Phytochem Rev (2025) 24:3649–3664
https://doi.org/10.1007/s11101-025-10096-8

# Artificial intelligence driven approaches in phytochemical research: trends and prospects

**Ressin Varghese · Harshita Shringi · Thomas Efferth · Siva Ramamoorthy**

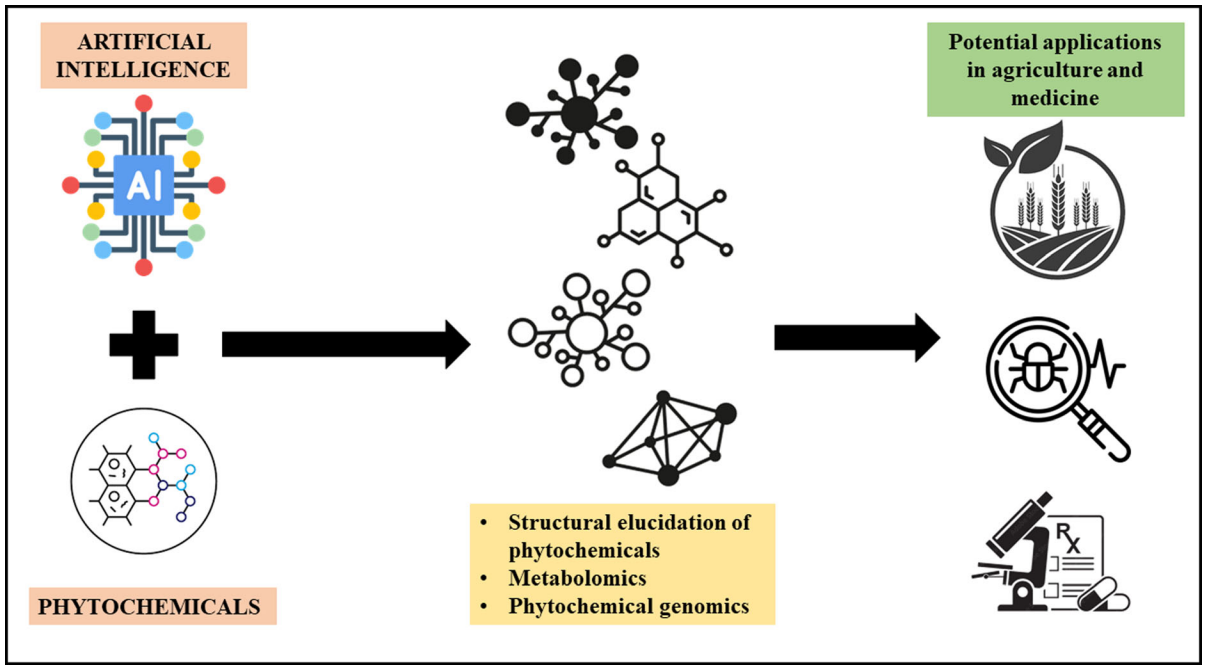
Received: 3 September 2024 / Accepted: 9 February 2025 / Published online: 27 February 2025


**Abstract** Tremendous scientific advancements have been witnessed in phytochemical research in pursuit of their therapeutic and nutritional value. Leveraging artificial intelligence (AI) is essential to handle the growing omics data and for the elucidation of novel potential phytochemicals. Interestingly, AI has transformed phytochemical research by enabling the efficient analysis of high-dimensional 'omics' data and facilitating the discovery of novel metabolites, structural elucidation, and metabolite profiling in plants. Taking together, this review highlights the implementation and significance of AI in various aspects of phytochemical research including analytical techniques, structural elucidation of phytochemicals, plant metabolomics, and genomics. The review also provides an outlook of prominent computational tools in phytochemical research including CASE followed by the present status and challenges of implementing AI in phytochemical research. We also propose the integration of more AI-driven analytical approaches in phytochemical research for the discovery of metabolites and to explore their applications in medicine and agriculture.

R. Varghese · H. Shringi · S. Ramamoorthy (✉)
School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
e-mail: siva.ramamoorthy@gmail.com

T. Efferth (✉)
Department of Pharmaceutical Biology, Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg University, 55122 Mainz, Germany
e-mail: efferth@uni-mainz.de

**Graphical abstract**




**Keywords** Phytochemicals · Artificial intelligence · Machine learning · AI driven NMR spectroscopy · Metabolomics · Genomics

## Introduction

Over the years, phytochemicals have emerged as pivotal natural products owing to their explicable therapeutic potential and commercial significance. In addition, these metabolites contribute to the defense system, communication, and environmental adaptation of the plant and maintain the homeostasis of plant growth and development which in turn influence the crop yield (Isah 2019). One million metabolites have been estimated to underline the unparalleled chemical diversity that should be elucidated to unleash their potential applications in agriculture, medicine, and industry (Saito 2013). Coupled with the advancement of high-throughput massive DNA sequencing technologies and metabolic profiling of a large number of plants, the volume of data in phytochemical research is mounting. In addition, these accumulating data should be explored to excavate more into the intricate relations of phytochemicals with functional genomics.

Interestingly, plant science research has been rationalized with the implementation of artificial intelligence to process the multiplying data. Computational phytochemistry, metabolomics, and phytochemical genomics are endorsing the implications of artificial intelligence. For instance, Machine learning (ML), one of the most discussed technical advancements of the century, has aided in estimating photosynthetic activity from plant cells to large terrestrial ecosystems by interpreting and analyzing these data in a fraction of the time (Varghese et al. 2023). Machine learning (ML), a subfield of artificial intelligence, utilizes algorithms that are known for performing the assigned task via the learning of the provided data set which may be labeled (supervised ML) or unlabeled (unsupervised ML). Deep learning is a highly speculated ML algorithm that trains the machine to learn via multiple processing layers. During learning, the algorithm automatically recognizes the patterns in the data. Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) are the best examples of Deep Learning (Arnold et al. 2011; Pouyanfar et al.

2018). Interestingly, ML has substantially alleviated the hurdles in the structural elucidation of unknown phytochemicals from known databases. Therefore, the current review accentuates the AI-driven approaches in various phytochemical research such as the present sophisticated analytical techniques and their part in the structural elucidation of novel phytochemicals. We also propose AI-assisted advanced analytical chemistry approaches to instigate the advancement of phytochemical research. Furthermore, the review presents the trends and prospects of AI in metabolomics and phytochemical genomics, a burgeoning domain.

## Computational phytochemistry: an outlook

It is a known fact that phytochemical research has been leveraged substantially by high-throughput computational approaches over the years. Computational phytochemistry is an umbrella term defining the implications of computational, mathematical, and statistical techniques to integrate the theories and experimental observations in phytochemical research. As aforementioned, the assertion of computational phytochemistry spans a multitude of levels beginning from the biosynthetic pathways to the high throughput screening of phytochemicals (Sarker and Nahar 2024). There are several computational approaches employed in the domain of phytochemical research even before the ascendancy of artificial intelligence in the twenty-first century. For instance, Density functional theory (DFT) illustrates a computational quantum mechanical model in phytochemical research exercised to predict the electronic structure of molecules and molecular properties including the vibrational frequencies, ionization, atomization energies, and electric and magnetic properties. DFT is concreted on the mathematical proof that the characteristics of quantum systems can be acquired from the density distribution of its fundamental constituents (Trappe and Chisholm 2023). Interestingly, Ullah et al. (2014) reported the similarity in the theoretical and computational modeling based on DFT defining the electronic and spectroscopic features of pistagremic acid from *Pistacia Integerrima*. In addition, electron affinities and ionization potential co-efficient of the highest and lowest unoccupied molecular orbital were estimated precisely. DFT coupled with molecular dynamics can be applied for deciphering the molecular structure of phytochemicals in organic solvents as compared to NMR data. $^{1}$H NMR spectrum of 5,4′-Dihydroxy-7,5′,3′-trimethoxyisoflavone from *Ouratea ferruginea* was theoretically profiled utilizing the hybrid B3LYP DFT functional (Hernandes et al. 2020).

Computer-assisted structural prediction, and determination was developed more than 50 years ago based on spectroscopic data. Remarkably, computer-aided structure elucidation (CASE) is primarily founded on the one-dimensional and two-dimensional NMR data which can minimize the errors in the structural elucidation of phytochemicals. CASE involves several axioms (statements) on spectrum-structure correlation such that if a molecule contains fragment AI, then structural features of the fragment can be observed in certain spectrum ranges. However, the general regularities-based hypothesis is considered more applicable than axioms when the structural elucidation is based on the 2D-NMR data. Here, the backbone of a phytochemical is derived from the main sources of structural information like COSY and HMBC correlations. The structure will be ultimately elucidated from the assembly of strict fragments (from 1D NMR, 2D COSY, and IR data) and fuzzy fragments (from 2DHMBC data) validating several axioms. The ACD structure elucidator is the oldest form of CASE program which requires only an empirical formula (from a high throughput spectrum analysis) and $^{13}$C chemical shifts edited from HSQC (heteronuclear single quantum correlation) and HMBC data. The program provides a list of probable structures correlating the spectral data and a molecular connectivity diagram in case of non-standard correlations. Bruker CMC-se program, The Mestrelb MNova structure elucidation program, Nuzillard's LSD program are some other programs available for the structural elucidation of phytochemicals (Burns et al. 2019). Interested readers can refer to Elyashberg et al. 2010 with an elaborate discussion outlining the methodology of structural elucidation from spectral data using CASE. Figure 1 provides an outline of steps involved in the structural elucidation of phytochemicals using CASE. CASE was coupled with NMR and HRESIMS data for the identification of phenylpropanoids from the *Ailanthus altissima.* The molecular formula of the unknown compounds was first elucidated from $^{13}$C, $^{1}$H NMR, and HRESIMS molecular-ion peak. Functional

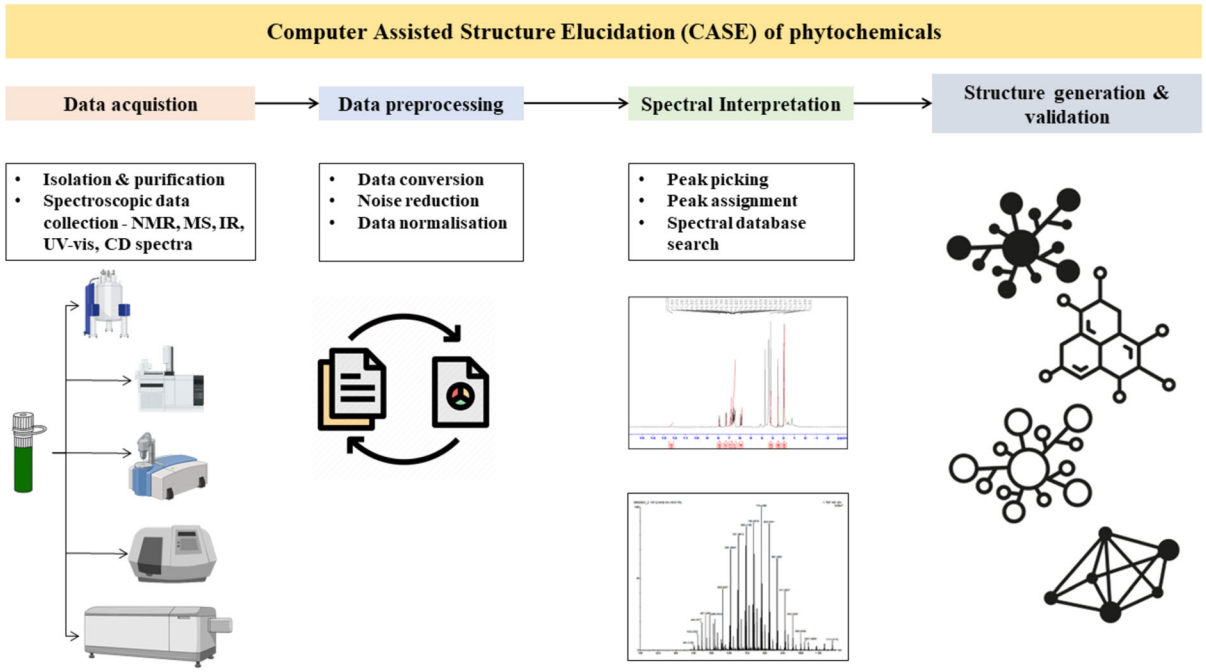



**Fig. 1** Outline of computer-assisted structure elucidation (CASE) of phytochemicals. The data of phytochemical collected from the spectroscopic instruments will be preprocessed. The preprocessed spectral data will be interpreted through peak picking, assignment and spectral database search which will results in the generation of structure

groups were assigned by HMBC correlations followed by the comparison with predicted $^{13}C$ with ACD/Spectrus processor (Du et al. 2019).

Another prominent computational implementation is chemometrics which provides required chemical information from experimental chemical data by mathematical, statistical, and other methods. Chemometrics analysis involves the following steps including the design of the experiment, data pre-processing, classification, calibration, knowledge provision, and understanding (Sarker and Nahar 2024). Remarkably, chemometrics involves the utilization of principal component analysis (PCA) and hierarchical cluster analysis (HCA) to offer descriptive and predictive explanations. PCA can be defined as an unsupervised pattern recognition approach for handling multivariate data. While HCA operates on the similarities of the processed data to receive the outcome (Patras et al. 2011). PCA, HCA, orthogonal projection to latent structure discriminant analysis (OPLS-DA), Pearson's correlation analysis, and batch learning self-organizing map analysis (BL-SOM) were utilized in evaluating the metabolite levels data of different Brassicaceae varieties received from GC–MS and HPLC. The statistical score plots aided in differentiating the metabolite profile of species. Thus, multivariate statistical assessment can be performed on metabolite data for discerning phenotypes (Lee et al. 2018). PCA and HCA were utilized recently in assessing the phenolics, flavonoid concentrations, and antioxidant potential in various parts of *Chaerophyllum bulbosum* from the HPL-DAD and spectrophotometric data (Tel-Çayan et al. 2022).

It is also obligatory to mention that the structural elucidation of phytochemicals has been accomplished from the spectral raw data by the fruitful utilization of various computational tools and databases over the years. The orbitrap-ion trap MS was coupled with CASE for the metabolic profiling of *Chimonanthus grammatus*. The UHPLC-HRMS-MS data of the plant extract was matched with ACD/MS workbook Suite with other databases including COCONUT. Besides, MS fragmentor was utilized for the prediction of the fragmentation pattern. ChromGenius was utilized for the retention time estimations which ultimately led to the identification of the major metabolites including the 2',4',5,7,8-pentahydroxyisoflavone (IS1), 3- (3,4-dihydroxyphenyl)-2,6,8-trihydroxy-4H-chromen-4-

one (IS2), melanoxetin. Table 1 represents the major computational tools and Table 2, the major databases utilized extensively in phytochemical research.

Most importantly, chemometrics, DFT, and all other computational approaches are more profusely employed in deciphering the therapeutic potential of the phytochemicals as compared to the metabolite profiling and structural prediction applications. It was tedious to handle the accumulating data on par with the rising number of databases and data generated by the computational tools. The advent of machine learning algorithms covered these lacunae and revolutionized phytochemical investigations which have been discussed in the following subsections.

## Machine-learning-assisted analytical techniques in phytochemical research

Analytical techniques are leeway to explore the hidden attributes of a sample including its molecular structure, molecular mass, isotropic signature, absorbance, and vibrational modes of molecules. The extracted data from analytical techniques are interpreted manually and computationally to establish the chemical fingerprint of a phytochemical. The flourishing of ML-directed analytical chemistry can be rationalized by the following thoughts. First, the necessity for critical computational strategies to obtain information on the phytochemicals from the large datasets generated by sophisticated instruments. Second, the accessibility of personal computers with high computational power by the plant chemists for the implementation of ML. Third, advanced information technologies permit rampant data mining for the development of computational tools, and databases and for the execution of ML (Debus et al. 2021). Machine learning algorithms regarded as the new element of artificial intelligence enable the machine to interpret large complex data sets. Robust computational algorithms in ML surpass human brains in predicting outcomes of input data in a very meticulous

**Table 1** Current tools that are used in phytochemical research

| Tool | Uses | Source | Reference |
|---|---|---|---|
| Response surface methodology (RSM) | Multifacetate statistic technique which is used to optimize various processes. RSM shows the effect of the independent variables on the dependent variables | https://develve.net/Response_surface_methodology.html | Yolmeh and Jafari (2017) |
| Self-organizing map (SOM) or (Kohonen Map) | An artificial neural network inspired by biological models is used for clustering and mapping techniques to map multidimensional data onto lower-dimensional data | http://livingforsom.com/ https://www.xlstat.com/en/solutions/features/self-organizing-maps | Kohonen (2013) |
| SPSS software | Used to recode data taken from measurement procedures in research projects. More precisely, it is used to do analyses, characterize features, and create new variables | https://www.ibm.com/products/spss-statistics | Yockey (2011) |
| Density Functional Theory (DFT) | It is used in the prediction of experimentally observable quantities, thus helping to understand the behavior of electrons in a material based on their density | https://gaussian.com/ | Geerlings et al. (2020) |
| Principal component analysis (PCA) | It removes noise and exposes hidden structures in the data set, helping in identifying the most meaningful basis to re-express a given data set | https://www.originlab.com/fileExchange/details.aspx?fid=328 | Kurita (2020) |
| Rstudio | An integrated development environment for the statistical computing and graphics programming language R | https://posit.co/download/rstudio-desktop/ | Kadiyala and Kumar (2017) |
| Statistica | It integrates the following features: multivariate analysis, design of experiments, graphical analysis, data mining, sigma six quality and process control statistics, random forests, CHAID, C&RT, MARsplines, GAM, random forests, classification, inter-action and boosted trees, MDS, and various machine-learning capabilities | https://statistica.software.informer.com/ | Hilbe (2012) |

**Table 2** Different types of plant databases and their applications

| Database | Source | Application |
|---|---|---|
| Ensembl Plants | https://plants.ensembl.org/index.html | Database presenting genome-scale information for many sequenced plant species. Data includes genome sequence, gene models, functional annotation, and polymorphic loci |
| National Center for Biotechnology (NCBI) Taxonomy | https://www.ncbi.nlm.nih.gov/taxonomy | Provides access to genomic information to advance science. The genome database currently comprises 578,406 species and can be sorted by kingdom, group, and subgroup |
| International Plant Names Index (IPNI) | https://www.ipni.org/ | Most complete tracheophyte database |
| Tropicos | https://www.tropicos.org/home | Database for bryophytes and New World tracheophytes |
| Plant Garden | https://plantgarden.jp/en/index | Database collect information from smaller databases and literature. Information on the organism, genome (chromosome number and genome size), markers, and genome-specific databases can be accessed |
| Phytozome | https://phytozome-next.jgi.doe.gov/ | Database portal wherein families of related genes representing the modern descendants of ancestral genes are constructed at key phylogenetic nodes |
| Plant DNA C-values database | https://cvalues.science.kew.org/ | Kew Royal Botanic Garden's Plant DNA C-values database contains C-values for 12,273 angiosperm, gymnosperm, pteridophyte, bryophyte, and algal species. Users can choose between analyzing C-value data across different groups of plants or searching just part of the database by selecting the specific plant group of interest |
| Plant GDB Genome Browser | https://goblinp.luddy.indiana.edu/ | The database displays current gene structure models and transcript evidence from spliced alignments of EST and cDNA sequences. If applicable, matched GSS contigs, community annotations, similar proteins, and microarray probes are displayed |

and well-defined manner. The classes of ML algorithms and their methodologies are beyond the scope of this review, and it is well expounded by Jordan and Mitchell (2015) and Naqa and Murphy (2015).

The ML-driven analysis of data retrieved from any techniques like magnetic resonance (NMR), mass spectroscopy (MS), vibrational spectroscopy, X-ray spectroscopy, atomic force microscope (AFM), electron microscope (EM), and two-dimensional (2D) chromatography can be performed through the stages of data enhancement i.e., inverse modeling, preprocessing followed by data modeling. The data preprocessing preliminarily involves the removal of artifacts in the spectral data which can be categorized as a forward and inverse problem (Houhou and Bocklitz 2021). A convolutional neural network (CNN) was trained utilizing simulated data for the preprocessing of Raman spectra by removing the cosmic ray, signal smoothing, and baseline subtraction (Wahl et al. 2020). Inverse modeling is another phase of data processing where any missing information is reconstructed from experimental measurements. For instance, the deep convolutional descattering autoencoder, a DL-based network has demonstrated the spectrum correction of the IR data with better robustness and speed as compared to the conventional multiplicative signal correction (ME-EMSC) algorithm (Magnussen et al. 2020). ML-driven data modeling can be exemplified in the context of different instruments spanning from vibrational spectroscopy to chromatographic techniques.

Vibrational spectroscopy deals with the vibrational frequency of each chemical bond which in turn aids in determining the chemical structure of a compound. FTIR vibrational spectrum can be exploited in the prediction of phytochemical content in conjunction with ML algorithms. A synchronous predictive model was developed for forecasting the polyphenol and epigallocatechin gallate in tea leaves. Interestingly, as aforementioned spectral data was preprocessed through Savitzky–Golay smoothing (SG), standard normal variate (SNV), vector normalization (VN),

multiplicative scatter correction (MSC), and first derivative (FD). The preprocessed spectral data was learned through Partial Least Squares Regression (PLSR) and Least squares support vector regression for developing the predictive models. Additionally, competitive adaptive reweighted sampling (CARS) and random forest (RF) were exercised to enhance the efficacy of the models. LS-SVR showed better performance, and the study underlined the potency of ML. ML-coupled FTIR approach in the prediction of phytochemicals for genotypic differentiation (Ye et al. 2023). Similarly, multivariate and ML-aided chemometrics were applied to the ATIR-FTIR vibrational profile to assess the pigment content in the lettuce. Preliminarily, the FTIR raw data was subjected to PCA, Linear Discriminant Analysis (LDA), (PLSR), and supervised machine learning (SVM) analysis and PCA was found to be the better predictive model. The classification of the lettuces as well as the prediction of the chemometric profiles like the chlorophyll content, pigment content, and flavonoid content was achieved with better accuracy and precision (Falcioni et al. 2022). The discovery of unknown phytochemicals has also been channeled by the integration of ML and NMR. Around 48 metabolites were identified by 1D and 2D NMR and were quantified by qNMR in *Capsicum annuum* cv. Jalapeño varieties. The analytical data were fed into the PCA, OPLS-DA for clustering and classification of varieties which unveiled the diversity and similarity between the races (Ramírez-Meraz et al. 2020).

Mass spectrometry ionizes the chemical species and arranges them for their mass-to-charge ratio (m/z). The complex interpretation of MS with several fragmented ions connotes the necessity of ML-coupled approaches. Most importantly, MS is a robust tool in structural prediction which will be discussed in the upcoming sections. Furthermore, phytochemical diversity analyzed in MS is an effective parameter in genotypic differentiation. Around 36 medicinal plant species have been identified by the m/z LC–MS data-trained ML classifiers like logistic regression (LR), SVM, RF (Nazarenko et al. 2016).

Chromatographic characterization of the large dataset of phytochemicals is also alleviated by the multilayered ML algorithms. HPLC is in conjunction with the ML models that are pertinent in anticipating the qualitative and quantitative profile of a phytochemical. The retention time of phytochemicals has been predicted in high-performance liquid chromatography through non-linear applied adaptive neuro-fuzzy inference systems (ANFIS). ANFIS is a complex model integrating the adaptive multilayer and feed-forward networks based on Takagi Sugeno models with higher predictive potential (Singh et al. 2023). The input variables were mobile phase composition, pH and the retention time was simulated as seen for the isoquercitrin in *Coriander sativum* L (Usman et al. 2021). The peak area was also later predicted to profile the qualitative and quantitative features of isoquercitrin through the artificial neural network (ANN), ANFIS, multilinear regression analysis (MLR), and SVM with input variables like the standard concentration, mobile phases, and pH. Another intriguing approach is that chromatograms can be predicted by imputation-augmented CNN with image-transformed temporal measurements as the input. Bacong and Juanico (2021) forecasted the chromatogram of *Blumea balsamifera* from sensorial data and correlated the environmental stress factors in terms of temperature, light, pH, and moisture with the phytochemical profile. Predicting the phytochemical composition of medicinal herbs on par with environmental fluctuations is a groundbreaking approach in the domain of phytomedicine.

## AI in structural elucidation of phytochemicals

Gratifyingly, the advancement of spectral technologies bourgeoned the variable data for interpreting molecular structural information. The identification of an unknown compound principally involves the correlation of descriptors like frequency range, multiple peaks, spacing between peaks, and maximum absorbance peak, from the experimental spectra with the available databases. Complex spectral data can be processed more promptly with the implementation of neural networks of AI as compared to conventional chemometrics. The enactment of AI in structural elucidation involves three approaches as explained by Xue et al. (2023). First is molecule to spectrum approach which involves the prediction of spectra followed by the discovery of candidate compounds in the database. Second, is the spectrum-to-structure way where the investigation or foreseeing of structures based on spectral data happens (Fig. 2), and the third approach is the de novo spectral generation of

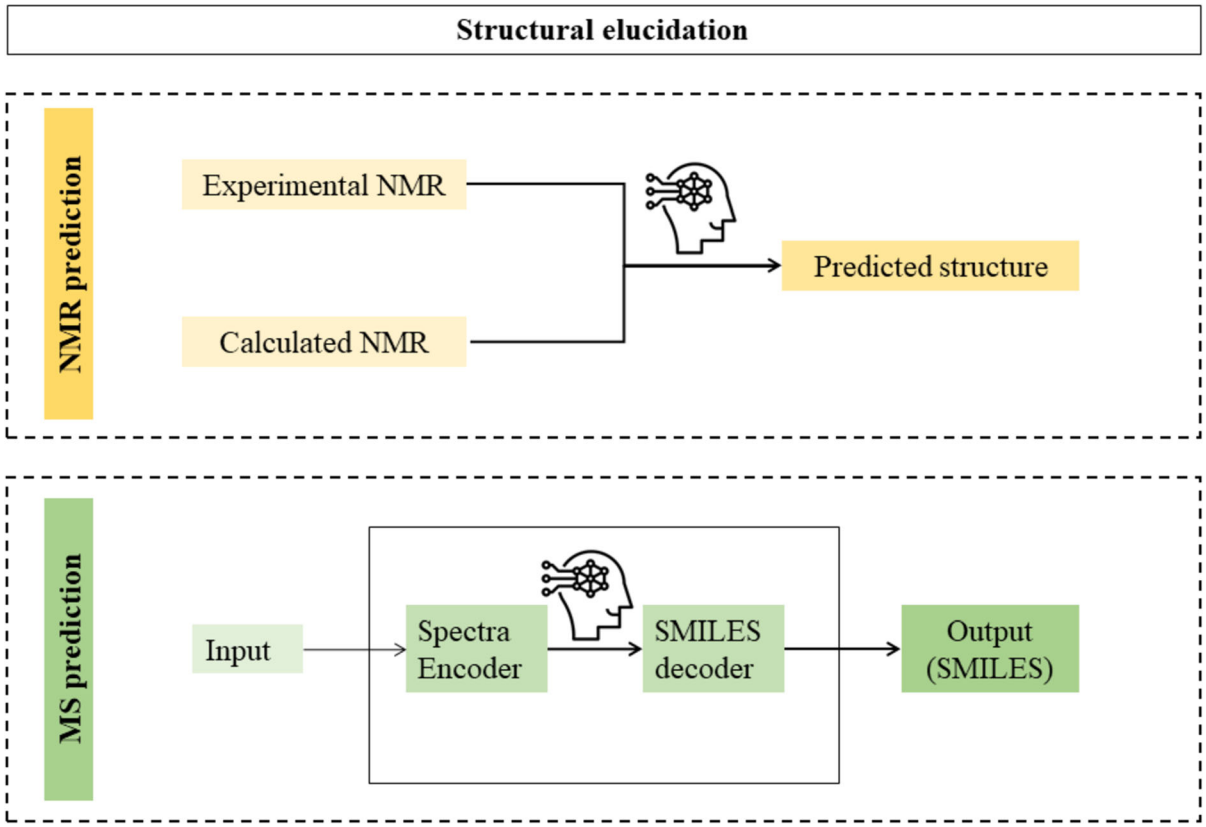



**Fig. 2** Block diagram structural elucidation by NMR and MS spectroscopy with the help of machine learning

molecules. Platforms like METLIN database can search, match, and score the experimental spectra with the library cataloged MS/MS spectra for the identification of the target compound which is a tiresome strategy. Unfortunately, AI-driven mass spectra-based structural elucidation reports were limited in plants and therefore we present some of the tools that can be prospected for phytochemical research. MetFID is one such application that employs MS/MS spectra for known compounds and their estimated fingerprints to train ANNs. MetFID follows a multicategory strategy for foreseeing fingerprints from the available databases for annotating a molecule (Fan et al. 2020). DeepEI is a DL-built program for the recognition of unknown compounds from Electron ionization–mass spectrometry data. The model is trained with database spectra and the trained model compares the predicted fingerprints and estimated fingerprints to rank the probable compounds (Ji et al. 2020).

In parallel, paramount efforts have been observed in the advancement of ML-aided approaches for the structural elucidation from NMR calculated data of natural products. Remarkably, the synergetic coupling of ML with NMR has lessened the structural misassignment and aided elucidation. Besides, the currently available empirical approaches often overlook the stereochemical and conformational effects on the prediction and impart computational cost burdens which can be assuaged by ML. The implementation of ML in NMR data interpretation can be categorized as prediction and correlation. Prediction involves the retrieval of quantum-quality NMR shifts and ML also accelerates the correlation of calculated NMR with the most probable structures (Cortés et al. 2023). Interestingly, Paruzzo et al. (2018) introduced a predictive ML tool termed shiftML which was trained using GIPAW DFT-calculated chemical shifts of structures from the Cambridge Structural Database (CSD) which was chosen using the farthest point sampling

algorithm. The trained shiftML was found to be highly accurate in the prediction of chemical shifts of solid molecules. Furthermore, pattern recognition analysis (PRA) was merged with ANN for the correlation and validation of spectral data with probable structures in the databases. The stipulation of merging ANN in PAR is to detect the regio- or stereochemical misassignments and to validate the diastereomers. The analysis was accomplished with mono-dimensional $^{1}H$ and $^{13}C$ NMR data with 2D HSQC correlations of 200 natural products as test and two-layer feed-forward ANNs. Besides, the validation was performed on originally misassigned 32 NPs with more than 400 trained ANNs which classified the right and wrong structural configurations for structural elucidation (Zanardi and Sarotti 2015). The discovery of premnafulvol from aerial regions of *Premna fulva* is an instant for ANN-PRA program. A complex 6/5/7/3- fused tetracyclic carbon skeleton was postulated through HR-ESI–MS and 1D and 2D-NMR data. $^{1}H$ and $^{13}C$ NMR chemical shift calculations for two possible diastereosiomers were estimated by the polarizable continuum model and NMR chemical shifts were assessed in DP4 + which predicted 1a was the preferred model. Furthermore, ANN-PRA approach corroborated 1a was the right model by employing tetramethylsilane and MSTD as references (Pu et al. 2018).

As discussed previously, it should be noted that DFT is the connoting link in the NMR spectra and structural prediction whereas ML is principally integrated to validate the credibility of the predicted data. Latterly, the comparison-based structural elucidation is performed by the computational approaches of allotting two sets of experimental data to two possible candidates (CP3) (Smith and Goodman 2009) or one set of experimental data to two or more reasonable structures (DP4) (Smith and Goodman 2010). Interestingly, Howarth et al. introduced DP4-AI, an AI-driven approach for the automatic interpretation of NMR spectra which could be effectively utilized in phytochemical research. The DP4-AI is cemented on the assignment algorithm (AA) that is responsible for designating the atoms in each candidate structure, as per the reported experimental peaks. DP4 also predicts chemical shift values utilizing the DFT GIAO methods. Interestingly, AA estimates the assignment probability matrix M to find the most likely identity of each experimental peak for the structural assignment and provides the probable structures (Howarth et al. 2020). Another sophisticated NMR structure validation software that can be utilized in phytochemical research is the DP5 by Howarth and Goodman. DP-5, a collective AI package predicts the uncertainty in structure from the $^{13}C$ NMR data and it handles the data processing, and the DFT calculations as well. DP5 estimates NMR shift using a previously established DP4-AI tool and raw data will be processed by NMR-AI. The pipeline involves the prediction error distribution for each atom that was found empirically by a Kernel Density Estimation, using a test-set of 5,140 molecules obtained from NMRShiftDB (Howarth and Goodman 2022). There are low apparent figures of investigations in phytochemical cosmos speculating above mentioned AI-driven NMR structure elucidation. Consequently, profuse attention is entailed to implement some of the AI-guided NMR strategies for the mining of phytochemicals from spectral data. Small Molecule Accurate Recognition Technology (SMART) is a deep CNN coupled with Non-Uniform Sampling (NUS) 2D NMR techniques for the uncovering of NPs through dereplication. Dereplication can be defined as the process of understanding the integrity of a novel molecule through its relation to the known ones. The SMART pipeline begins with the input of NUS HSQC spectra from our molecule of interest (here, phytochemical) followed by mapping of the spectra into a ten-dimensional space and then to 2D space by PCA for visualization through the CNN of 8 layers (4 convolutional layers and 4 fully connected layers). The training dataset includes the phytochemicals from different plant species which were clustered and the structure was interpreted (Zhang et al. 2017). Howbeit, the utility SMART in drug discovery, was also applied for the discovery of structurally diverse sesquiterpenes from *Litsea lancilimba* Merr. (Zhang et al. 2022) and from *Elephantopus tomentosus* L. (Bai et al. 2023).

## AI in plant metabolomics: a combinatorial cutting-edge technology

Plant metabolite diversity constitutes more than 1 million metabolites with impeccable bioactive properties as well as functions in plant growth and development. This rich metabolites consortium was estimated to be USD 23.7 billion worth in 2019 and is speculated to reach USD 59.4 billion by 2025

(Markets and Markets 2019). In this scenario, metabolomics developed as a comprehensive, nonbiased, high-throughput analysis of complex metabolites from plants (Hall et al. 2002). The integration of AI with metabolomics is requisite for the annotation of untargeted metabolomes and thus the isolation of novel metabolites (Tsugawa et al. 2021). It is obligatory to mention the metabolite annotation levels before dwelling into the AI assisted approaches. According to Metabolomics Standards Initiative (MSI), metabolite annotations start with level 1, i.e., the identification of unknown metabolites with validated standards, and level 2 is the annotation of metabolites using spectral libraries. While level 3 is the putative characterization of metabolite because of diagnostic ion and/or partial spectral similarities to chemical classes and level 4 is the unknown metabolite (Summer et al. 2007). These levels have been further rationalized and tailored by Schymanski et al. (2014). The level 2 and 3 annotations aid in the dereplication and unraveling of novel metabolites. The AI-based metabolite annotation has been discussed in detail in the previous subsection of AI in structural elucidation. Here, we focus on the relevance of AI in utilizing the targeted and untargeted metabolomics data and its application in plant growth and development with some case studies. While untargeted metabolomics is a top-down strategy featured by the subsequent measurement of the high number of known and unknown metabolites, targeted metabolomics directs the analysis of specific metabolites (Han et al. 2023). It should be noted that the metabolite profiling of the medicinal plants for the targeted drug receptors is rapidly escalating these days, which has been reviewed extensively in Cerny et al. (2020).

The objectives of AI-assisted metabolomics can be cataloged as first quality assessment of crops, second assessing the variations in interested metabolite levels, and third, early disease detection as depicted in Fig. 3. Machine learning was coupled with untargeted metabolomics to examine quality attributes of mung beans based on climate and geographical locations. Around 43 mung beans were subjected to UPLC-QTOF-MS resulting in the profiling of more than 11,000 metabolites. RF and SVM models classified mung beans based on the correlation of metabolites with locations and climate zones (He et al. 2023). Untargeted metabolomics has been used to profile the phenolic content of three species of *Bryophyllum* medicinal plants via UHPLC-QTOF/MS which engendered in the annotation of 485 metabolites. Furthermore, the authors implemented ML algorithms to decipher the factors affecting the biosynthesis of phenols. Neurofuzzy logic model evaluated the interactions of variables like organs, nutrient factors, and genotypes. Adaptive Spline Modeling of Data (ASMOD) was chosen for parameter minimization, and modeling includes fitness criteria like Bayesian Information Criterion, cross-validation, Structural Risk Minimization, Minimal Description Length, and Leave One Out Cross Validation. The persistent validation of models was performed by ANOVA and the model then generated IF–THEN rules to assess the factors determining phenolic biosynthesis. As aforementioned, non-targeted metabolomics has been coupled with ML classifiers including the Regularized logistic regression (LR-L2) and gradient-boosted decision tree (GBDT) for the early detection of Huanglongbing in citrus species (Wang et al. 2022a). The AI-integrated targeted metabolomics principally also has analogous applications like early disease detection, phenotypic differentiation, and quality assessment of crops. For instance, quality attributes were evaluated in different phenotypes of soybeans in correlation with growing periods, and various environmental factors were profiled through targeted metabolomics of 36 polyphenols. The neural model was identified as the best predictive model to determine the factor imparting the polyphenol content. It is an unequivocal fact that machine learning in conjunction with metabolomics is an assuring approach in promoting crop development amidst food crises and extreme climatic circumstances.

## Phytochemical genomics and machine learning

The escalating integrated omics approaches are opening avenues for ML algorithms to develop metabolite models for desirable plant species. There are wide possibilities to predict the integrated associations of phytochemicals with their key biosynthetic genes via datasets trained with omics information (Rai et al. 2019). Phytochemical genomics is the systematic integration of omics techniques (genomics, transcriptomics, proteomics, metabolomics) to evaluate the evolution, biosynthetic mechanism, and function of plant metabolites or phytochemicals (Saito 2013).

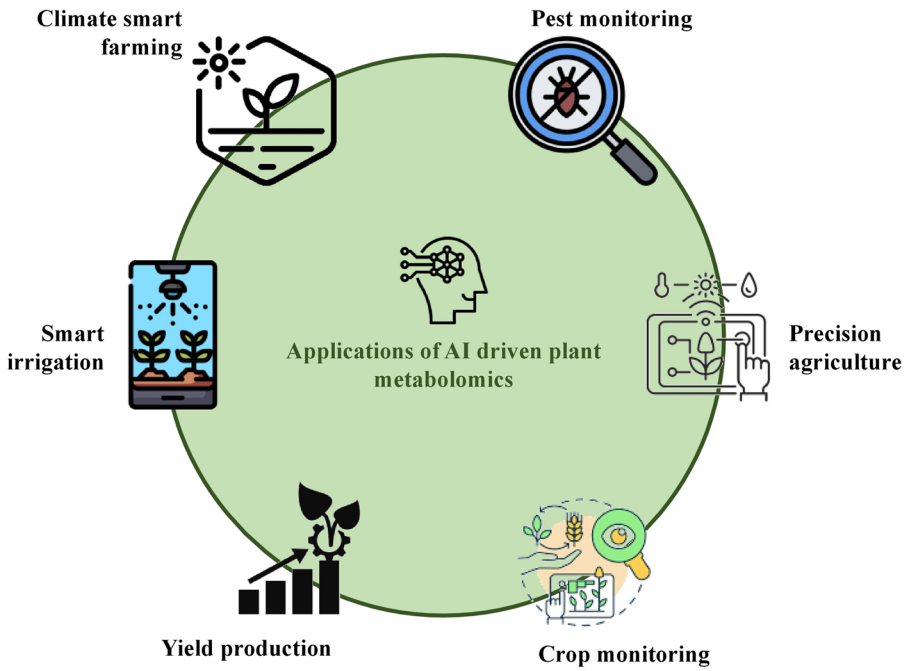



**Fig. 3** Application of AI-driven plant metabolomics. AI driven metabolomics also has its advantages in the field of agriculture where it can be used to monitor pests, and analyze the growth parameters for crop cultivation, thus enhancing crop yield. It also advance the precision agriculture, smart irrigation and paves way for climate smart farming

Designing phytochemical models with DL algorithms to foresee the metabolic variations in response to environmental stress factors is one such approach to establish. Genome-Scale Metabolic Models (GEMs) are network-based tools that assemble the whole metabolic data of a biological system such as the information on genes, enzymes, reactions, associated gene-protein-reaction (GPR) rules as well as metabolites. The GEMs have already been established in plants like *Arabidopsis thaliana* (Poolman et al. 2009). Integrating DL or CNN in model preparation will accelerate the design and use of GEMS. We propose that DL-driven phytochemical genomics can play a pivotal role in phytochemical research and crop improvement. DL can foresee the molecular phenotypes including the transcription factor binding, epigenetic marks, and gene expression levels using the genomic sequences as inputs. Besides, DL can be used to predict QTL in different species. Interestingly, another possibility is the utilization of DL for the prediction of allelic variation upon environmental stress factors (Wang et al. 2020).

Interestingly, there are several tools for predicting miRNAs like miTAR which integrates the CNN and RNN for learning intrinsic spatial and sequential features of miRNA and target (Gu et al. 2021). Another major leap is the substantial number of AI tools correlating plant genomics and stress resistance. Table 3 discusses some of the other AI-based tools utilized in omics. Unfortunately, most of the available AI-based tools are designed for research in human genetics and the implementation of these tools in plants is the need of the hour. Nonetheless, some of the promising innovations in AI-driven phytochemical genomics are considered here. Plant-specialized metabolism (SM) enzymes can synthesize lineage-specific metabolites with pivotal ecological, functional, evolutionary, and biotechnological inferences. Moore et al. (2019) developed an ML-based predictive model that predicts the SM genes which led to the discovery of around 1220 SM genes in *A. thaliana.* Fürtauer et al. (2018) introduced an ML-integrated predictive model to examine the metabolic stress response in *A. thaliana.* Abiotic stress can induce metabolic reprogramming in plants which involves an intricate association of signaling cascades connecting transcriptional, translational, and metabolic regulatory pathways. The SVM model identified a module of 23

**Table 3** Different types of AI-based tools that are used in metabolomics studies

| Tool | Application | Source | Reference |
|---|---|---|---|
| MS2LDA | Ms2lda.org allows users to decompose and annotate MS/MS data with MS2LDA | https://ms2lda.org/ | Wandy et al. (2018) |
| SIRIUS 4 | SIRIUS is an open java-based software that identifies metabolites using single and tandem mass spectrometry | https://bio.informatik.uni-jena.de/software/sirius-4-0-1/ | Dührkop et al. (2019) |
| CFM-ID | The CFM-ID 4.0 web server is an online tool for predicting, annotating and interpreting tandem mass (MS/MS) spectra of small molecules | https://cfmid.wishartlab.com/ | Wang et al. (2022b) |
| MetFrag | MetFrag is a freely available software used in the identification of the structure of molecules by annotation of high-precision tandem mass spectra of metabolites | https://msbi.ipb-halle.de/MetFrag/ | Ruttkies et al. (2016) |
| MAGMa | MAGMa is an online application for the automatic chemical annotation of mass spectrometry data | https://research-software-directory.org/software/magma | Ridder et al. (2014) |
| MS-FINDER | MS-FINDER is a universal compound annotation program that supports EI-MS (GC/MS) and MS/MS spectral mining | https://systemsomicslab.github.io/compms/msfinder/main.html | Lai et al. (2018) |
| RIKEN PRIMe | RIKEN PRIMe site is used for standardized measurements of metabolites obtained by using NMR, GC/MS, LC/MS, and CE/MS, and metabolomics tools that support related analyses | https://www.g6g-softwaredirectory.com/bio/metabolomics/metab-profiling/20659-RIKEN-PRIMe.php | Sawada et al. (2012) |
| MS-DIAL | MS-DIAL is an open-source software pipeline for the identification and quantification of small molecules by mass spectral deconvolution | https://systemsomicslab.github.io/compms/msdial/main.html | Tsugawa et al. (2015) |
| AllCCS | AllCCS is used for the annotation of small molecules, CCS (collision cross-section) prediction based on ML, and provides access to a unified CCS database. It is commonly used for ion mobility-mass spectrometry (IM-MS) data analysis | http://allccs.zhulab.cn/ | Zhang et al. (2023) |
| GNPS | GNPS is a web-based mass spectrometry, an open-access knowledge base used for sharing raw, processed, or annotated fragmentation mass spectrometry data (MS/MS) | https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp | Wang et al. (2016) |
| MetaboAnalyst | MetaboAnalyst is a comprehensive web application for metabolomic data analysis and interpretation of NMR and MS data | https://www.metaboanalyst.ca/ | Xia and Wishart (2016) |
| MZmine 2 | Used in data processing for GC–MS and LC–MS, in addition to data visualization and basic statistics analyses | https://mzio.io/ | Pluskal et al. (2010) |
| OpenMS | OpenMS offers an open-source C + + library is used for LC/MS data management, analysis and visualization | https://openms.de/ | Röst et al. (2016) |
| XCMS | XCMS software is used to process untargeted metabolomic data, it is commonly used to process LC/MS-based metabolomic data | https://xcmsonline.scripps.edu/landing_page.php?pgcontent=mainPage | Mahieu et al. (2016) |

MS2LDA, Mass Spectrometry to Latent Dirichlet Allocation; CFM-ID, Competitive Fragment Modeling Identification; GNPS, Global Natural Products Social Molecular Networking; XCMS, eXtensible Computational Mass Spectrometry

proteins related to temperature and extreme light stress factors. 18 proteins demonstrated a putative protein–protein interaction concerning transcriptional regulation and the modulation of primary and secondary metabolite synthesis. As exemplified, genomics investigations are often confined to the model plants. Optimistically, the advent of ML could aid collection of whole genomics data of industrially relevant plants, which will revolutionize phytochemical genomics within a short period.

## Concluding remarks and prospects

Phytochemical research is intricate, addressing several attributes including the biosynthetic pathways, novel metabolites, association of metabolite profiling with external stresses, and leeway to augment metabolite production. Artificial intelligence has drastically shaped the phytochemical cosmos in all these aspects as illustrated in the current review. In addition, we observed that AI-assisted approaches in phytochemical research majorly concentrate on the discovery of potential drugs for the ailment of cancers and viral infections which is not discussed here. Copious reviews are available, explicating the methodology, advantages, and limitations of the implementation of AI in drug discovery (Saldívar-González et al. 2022; Mullowney et al. 2023; Manochkumar and Ramamoorthy 2024).

ML approaches were proven to be effective in integrating analytical technologies with huge datasets, operating complex analyses, and supplying automated tools for prediction. ML algorithms like CNN, ANN, SVM, and RF were found to be thriving in aiding the spectroscopic and chromatographic techniques to process the spectral data and for structural elucidation. Despite the success in developing predictive models, there are concerns about interpreting the neural networks in correlation to physicochemical phenomena which require the need for preprocessing of data with tools like PCA. Besides, mass spectrometry is currently being used in the analysis of spatially resolved omics through mass spectrometry imaging (MSI) techniques which allow molecule detection from plant tissue. There are voids in the structural annotations which could be resolved through the integration of AI. Furthermore, AI could also be implemented in the MS in single cells (Ma 2022). In the future, we could expect sophisticated ML methods that can predict structural data from NMR/MS spectra /chromatographic data with high accuracy in real time. Optimistically, more accurate metabolite class, substructure, molecular backbone, and even the stereochemistry of structures will be predicted with more precision in the near future.

Despite these advancements, there are very few investigations on the integration of AI in the structural elucidation of phytochemicals. One of the bottlenecks which needed attention in AI implementation is the lack of user-friendly interfaces and software packages which might be hindering its widespread application in phytochemicals. Besides, most advancements in AI have occurred in the past four years and it will not take considerable time to transfer the techniques to phytochemical research. Further, artificial intelligence-aided metabolomics and genomics research will unquestionably contribute to hastening phytochemical research. Unfortunately, as aforementioned there should be unprecedented progress in integrating AI couple metabolomics to augment industrially valuable phytochemicals production.

**Acknowledgements** RV, HS, and SR thank Vellore Institute of Technology management for their constant support in the completion of the manuscript.

**Author contributions** RV contributed to data collection and manuscript preparation. HS prepared the figures and tables. TE revised and edited the manuscript. SR conceptualized, supervised and edited the manuscript.

**Funding** No funding was received to assist with the preparation of this manuscript.

**Declarations**

**Conflict of interest** The authors declare they have no conflicts of interest to declare.

**Research involving human participants and/or animals** Not applicable.

**Informed consent** Not applicable.

## References

Arnold L, Rebecchi S, Chevallier S & Paugam-Moisy H (2011) An introduction to deep learning. In: European symposium on artificial neural networks (ESANN)

Bacong JRC, Juanico DEO (2021) Predictive chromatography of leaf extracts through encoded environmental forcing on phytochemical synthesis. Front Plant Sci. https://doi.org/10.3389/fpls.2021.613507

Bai M, Xu W, Zhang X, Li Q, Du NN, Liu DF, Huang XX (2023) HSQC-based small molecule accurate recognition technology discovery of diverse cytotoxic sesquiterpenoids from *Elephantopus tomentosus* L. and structural revision of molephantins A and B. Phytochemistry. https://doi.org/10.1016/j.phytochem.2022.113562

Burns DC, Mazzola EP, Reynolds WF (2019) The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. Nat Prod Rep. https://doi.org/10.1039/C9NP00007K

Cerny MA, Kalgutkar AS, Obach RS, Sharma R, Spracklin DK, Walker GS (2020) Effective application of metabolite profiling in drug design and discovery. J Med Chem. https://doi.org/10.1021/acs.jmedchem.9b01840

Cortés I, Cuadrado C, Hernández Daranas A, Sarotti AM (2023) Machine learning in computational NMR-aided structural elucidation. Front Nat Prod. https://doi.org/10.3389/fntpr.2023.1122426

Debus B, Parastar H, Harrington P, Kirsanov D (2021) Deep learning in analytical chemistry. TrAC. https://doi.org/10.1016/j.trac.2021.116459

Du YQ, Yan ZY, Chen JJ, Wang XB, Huang XX, Song SJ (2021) The identification of phenylpropanoids isolated from the root bark of *Ailanthus altissima* (Mill.) Swingle. Nat Prod Res. https://doi.org/10.1080/14786419.2019.1643861

Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Böcker S (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. https://doi.org/10.1038/s41592-019-0344-8

El Naqa I, Murphy MJ (2015). In: El Naqa I, Murphy MJ (eds) What is machine learning? Springer, Berlin

Elyashberg M, Williams AJ, Blinov K (2010) Structural revisions of natural products by computer-assisted structure elucidation (CASE) systems. Nat Prod Rep. https://doi.org/10.1039/C002332A

Falcioni R, Moriwaki T, Gibin MS, Vollmann A, Pattaro MC, Giacomelli ME, Antunes WC (2022) Classification and prediction by pigment content in lettuce (*Lactuca sativa* L.) varieties using machine learning and ATR-FTIR spectroscopy. Plants. https://doi.org/10.3390/plants11243413

Fan Z, Alley A, Ghaffari K, Ressom HW (2020) MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. Metabolomics. https://doi.org/10.1007/s11306-020-01726-7

Fürtauer L, Pschenitschnigg A, Scharkosi H, Weckwerth W, Nägele T (2018) Combined multivariate analysis and machine learning reveals a predictive module of metabolic stress response in *Arabidopsis thaliana*. Mol Omics. https://doi.org/10.1039/C8MO00095F

Geerlings P, Chamorro E, Chattaraj PK, De Proft F, Gázquez JL, Liu S, Ayers P (2020) Conceptual density functional theory: status, prospects, issues. Theor Chem Acc. https://doi.org/10.1007/s00214-020-2546-7

Gu T, Zhao X, Barbazuk WB, Lee JH (2021) miTAR: a hybrid deep learning-based approach for predicting miRNA targets. BMC Bioinform. https://doi.org/10.1186/s12859-021-04026-6

Hall R, Beale M, Fiehn O, Hardy N, Sumner L, Bino R (2002) Plant metabolomics: the missing link in functional genomics strategies. Plant Cell. https://doi.org/10.1105/tpc.140720

Han W, Ward JL, Kong Y, Li X (2023) Targeted and untargeted metabolomics for the evaluation of plant metabolites in response to the environment. Front Plant Sci. https://doi.org/10.3389/fpls.2023.1167513

He L, Hu Q, Yu Y, Yu Y, Yu N, Chen Y (2023) Discrimination of mung beans according to climate and growing region by untargeted metabolomics coupled with machine learning methods. Food Control. https://doi.org/10.1016/j.foodcont.2023.109927

Hernandes IS, Da Silva HC, Dos Santos HF, De Almeida WB (2020) Conformational analysis of 5,4′-dihydroxy-7,5′,3′-trimethoxyisoflavone in solution using 1H NMR: a density functional theory approach. J Phys Chem A. https://doi.org/10.1021/acs.jpca.0c02996

Hilbe JM (2012) STATISTICA 7: an overview. Am Stat. https://doi.org/10.1198/000313007X172998

Houhou R, Bocklitz T (2021) Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. ASA. https://doi.org/10.1002/ansa.202000162

Howarth A, Ermanis K, Goodman JM (2020) DP4-AI automated NMR data analysis: straight from spectrometer to structure. Chem Sci. https://doi.org/10.1039/d0sc00442a

Howarth A, Goodman JM (2022) The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. Chem Sci. https://doi.org/10.1039/D1SC04406K

Isah T (2019) Stress and defense responses in plant secondary metabolites production. Bio Res. https://doi.org/10.1186/s40659-019-0246-3

Ji H, Deng H, Lu H, Zhang Z (2020) Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. Anal Chem. https://doi.org/10.1021/acs.analchem.0c01450

Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science. https://doi.org/10.1126/science.aaa8415

Kadiyala A, Kumar A (2017) Applications of R to evaluate environmental data science problems. Environ Prog Sustain Energy. https://doi.org/10.1002/ep.12676

Kohonen T (2013) Essentials of the self-organizing map. Neural Netw. https://doi.org/10.1016/j.neunet.2012.09.018

Kurita T (2020) Principal component analysis (PCA). In: Computer vision. Springer, Cham. https://doi.org/10.1007/978-3-030-03243-2_649-1

Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Fiehn O (2018) Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods. https://doi.org/10.1038/nmeth.4512

Lee KB, Kim YJ, Kim HJ, Choi J, Kim JK (2018) Phytochemical profiles of Brassicaceae vegtables and their multivariate characterization using chemometrics. Appl Biol Chem. https://doi.org/10.1007/s13765-017-0340-6

Ma X (2022) Recent advances in mass spectrometry-based structural elucidation techniques. Molecules. https://doi.org/10.3390/molecules27196466

Magnussen EA, Solheim JH, Blazhko U, Tafintseva V, Tøndel K, Liland KH, Kohler A (2020) Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. J Biophotonics. https://doi.org/10.1002/jbio.202000204

Mahieu NG, Genenbacher JL, Patti GJ (2016) A roadmap for the XCMS family of software solutions in metabolomics. Curr Opin Chem Biol. https://doi.org/10.1016/j.cbpa.2015.11.009

Manochkumar J, Ramamoorthy S (2024) Artificial intelligence in the 21st century: the treasure hunt for systematic mining of natural products. Curr Sci. https://doi.org/10.18520/cs/v126/i1/19-35

Market reports by markets and markets, plant extracts market—forecast to 2026. https://www.marketsandmarkets.com/Market-Reports/plant-extracts-market-942.html

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Shiu SH (2019) Robust predictions of specialized metabolism genes through machine learning. PNAS. https://doi.org/10.1073/pnas.1817074116

Mullowney MW, Duncan KR, Elsayed SS, Garg N, van der Hooft JJ, Martin NI, Medema MH (2023) Artificial intelligence for natural product drug discovery. Nat Rev Drug Discov. https://doi.org/10.1038/s41573-023-00774-7

Nazarenko DV, Kharyuk PV, Oseledets IV, Rodin IA, Shpigun OA (2016) Machine learning for LC–MS medicinal plants identification. Chemom Intell Lab Syst. https://doi.org/10.1016/j.chemolab.2016.06.003

Paruzzo FM, Hofstetter A, Musil F, De S, Ceriotti M, Emsley L (2018) Chemical shifts in molecular solids by machine learning. Nat Commun. https://doi.org/10.1038/s41467-018-06972-x

Patras A, Brunton NP, Downey G, Rawson A, Warriner K, Gernigon G (2011) Application of principal component and hierarchical cluster analysis to classify fruits and vegetables commonly consumed in Ireland based on in vitro antioxidant activity. J Food Compos Anal. https://doi.org/10.1016/j.jfca.2010.09.012

Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinform. https://doi.org/10.1186/1471-2105-11-395

Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol. https://doi.org/10.1104/pp.109.141267

Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, Shyu ML, Chen SC, Iyengar SS (2018) A survey on deep learning: algorithms, techniques, and applications. ACM Comput Surv (CSUR). https://doi.org/10.1145/3234150

Pu DB, Du BW, Chen W, Gao JB, Hu K, Shi N, Xiao WL (2018) Premnafulvol A: a diterpenoid with a 6/5/7/3-fused tetracyclic core and its biosynthetically related analogues from *Premna fulva*. Org Lett. https://doi.org/10.1021/acs.orglett.8b02845

Rai A, Yamazaki M, Saito K (2019) A new era in plant functional genomics. Curr Opin Syst Biol. https://doi.org/10.1016/j.coisb.2019.03.005

Ramírez-Meraz M, Méndez-Aguilar R, Hidalgo-Martínez D, Villa-Ruano N, Zepeda-Vallejo LG, Vallejo-Contreras F, Becerra-Martínez E (2020) Experimental races of *Capsicum annuum* cv. jalapeno: chemical characterization and classification by 1H NMR/machine learning. Food Res Int. https://doi.org/10.1016/j.foodres.2020.109763

Ridder L, van der Hooft JJ, Verhoeven S (2014) Automatic compound annotation from mass spectrometry data using MAGMa. Mass Spectrom. https://doi.org/10.5702/massspectrometry.s0033

Röst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, Kohlbacher O (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. https://doi.org/10.1038/nmeth.3959

Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J Cheminform. https://doi.org/10.1186/s13321-016-0115-9

Saito K (2013) Phytochemical genomics—a new trend. Curr Opin Plant Biol. https://doi.org/10.1016/j.pbi.2013.04.001

Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F (2022) Natural product drug discovery in the artificial intelligence era. Chem Sci. https://doi.org/10.1039/D1SC04471K

Sarker SD, Nahar L (2024). In: Sarker SD, Nahar L (eds) Computational phytochemistry: an overview in computational phytochemistry. Elsevier, Amsterdam

Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, Saito K (2012) RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. Phytochemistry. https://doi.org/10.1016/j.phytochem.2012.07.007

Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. https://doi.org/10.1021/es5002105

Singh YR, Shah DB, Kulkarni M, Patel SR, Maheshwari DG, Shah JS, Shah S (2023) Current trends in chromatographic prediction using artificial intelligence and machine learning. Anal Methods. https://doi.org/10.1039/D3AY00362K

Smith SG, Goodman JM (2009) Assigning the stereochemistry of pairs of diastereoisomers using GIAO NMR shift calculation. J Org Chem. https://doi.org/10.1021/jo900408d

Smith SG, Goodman JM (2010) Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. J Am Chem Soc. https://doi.org/10.1021/ja105035r

Summer LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Hardy N (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics. https://doi.org/10.1007/s11306-007-0082-2

Tel-Çayan G, Deveci E, Çayan F, Molo Z, Duru ME, Yeşil Y (2022) Chemometrics evaluation of phytochemicals and antioxidant activities of the extracts of *Chaerophyllum bulbosum* roots and aerial parts. Anal Lett. https://doi.org/10.1080/00032719.2021.1930026

Trappe MI, Chisholm RA (2023) A density functional theory for ecology across scales. Nat Commun. https://doi.org/10.1038/s41467-023-36628-4

Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, Arita M (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. https://doi.org/10.1038/nmeth.3393

Tsugawa H, Rai A, Saito K, Nakabayashi R (2021) Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. Nat Prod Rep. https://doi.org/10.1039/d1np00014d

Ullah H, Rauf A, Ullah Z, Anwar M, Uddin G, Ayub K (2014) Density functional theory and phytochemical study of Pistagremic acid. Spectrochim Acta A Mol Biomol Spectrosc. https://doi.org/10.1016/j.saa.2013.08.099

Usman AG, Işik S, Abba SI, Meriçli F (2021) Chemometrics-based models hyphenated with ensemble machine learning for retention time simulation of isoquercitrin in *Coriander sativum* L. using high-performance liquid chromatography. J Sep Sci. https://doi.org/10.1002/jssc.202000890

Varghese R, Cherukuri AK, Doddrell NH, Doss CGP, Simkin AJ, Ramamoorthy S (2023) Machine learning in photosynthesis: Prospects on sustainable crop development. Plant Sci. https://doi.org/10.1016/j.plantsci.2023.111795

Wahl J, Sjödahl M, Ramser K (2020) Single-step preprocessing of Raman spectra using convolutional neural networks. Appl Spectrosc. https://doi.org/10.1177/0003702819888949

Wandy J, Zhu Y, van der Hooft JJ, Daly R, Barrett MP, Rogers S (2018) Ms2lda. org: web-based topic modelling for substructure discovery in mass spectrometry. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx582

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Bandeira N (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat Biotechnol. https://doi.org/10.1038/nbt.3597

Wang H, Cimen E, Singh N, Buckler E (2020) Deep learning for plant genomics and crop improvement. Curr Opin Plant Biol. https://doi.org/10.1016/j.pbi.2019.12.010

Wang F, Allen D, Tian S, Oler E, Gautam V, Greiner R, Wishart DS (2022b) CFM-ID 4.0—a web server for accurate MS-based metabolite identification. Nucleic Acids Res. https://doi.org/10.1093/nar/gkac383

Wang Z, Niu Y, Vashisth T, Li J, Madden R, Livingston TS, Wang Y (2022a) Nontargeted metabolomics-based multiple machine learning modeling boosts early accurate detection for citrus Huanglongbing. Hortic Res. https://doi.org/10.1093/hr/uhac145

Xia J, Wishart DS (2016) Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. Curr Protoc Bioinf. https://doi.org/10.1002/cpbi.11

Xue X, Sun H, Yang M, Liu X, Hu HY, Deng Y, Wang X (2023) Advances in the application of artificial intelligence-based spectral data interpretation: a perspective. Anal Chem. https://doi.org/10.1021/acs.analchem.3c02540

Ye S, Weng H, Xiang L, Jia L, Xu J (2023) Synchronously predicting tea polyphenol and epigallocatechin gallate in tea leaves using Fourier transform–near-infrared spectroscopy and machine learning. Molecules. https://doi.org/10.3390/molecules28145379

Yockey RD (2011) SPSS demystified. A step-by-step guide to successful data analysis. Pearson Education Inc

Yolmeh M, Jafari SM (2017) Applications of response surface methodology in the food industry processes. Food Bioprocess Technol. https://doi.org/10.1007/s11947-016-1855-2

Zanardi MM, Sarotti AM (2015) GIAO C-H COSY simulations merged with artificial neural networks pattern recognition analysis. Pushing the structural validation a step forward. J Org Chem. https://doi.org/10.1021/acs.joc.5b01663

Zhang C, Idelbayev Y, Roberts N, Tao Y, Nannapaneni Y, Duggan BM, Gerwick WH (2017) Small molecule accurate recognition technology (SMART) to enhance natural products research. Sci Rep. https://doi.org/10.1038/s41598-017-13923-x

Zhang YJ, Bai M, Li KX, Yan QL, Huang XX, Zheng J, Song SJ (2022) Small molecule accurate recognition technology accelerated isolation of structurally diverse sesquiterpenes from *Litsea lancilimba* Merr. Fitoterapia. https://doi.org/10.1016/j.fitote.2022.105168

Zhang H, Luo M, Wang H, Ren F, Yin Y, Zhu ZJ (2023) AllCCS2: curation of ion mobility collision cross-section atlas for small molecules using comprehensive molecular representations. Anal Chem. https://doi.org/10.1021/acs.analchem.3c02267