Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

Machine Learning For Specialized-Metabolite Prediction In Medicinal Plants: A Validation-Focused Review Of Five Ethnomedicinal Species

Najma Farid[1], Muntaha Kanwal[2], Saffi Ur Rehman[3]*, Aqeela Kanwal[4]

[1]College of Materials Science and Engineering, Hohai University, Nanjing 210024, China
[2]Department of Botany, Faculty of Sciences, University of Agriculture, Faisalabad, 38000, Pakistan
[3]College of Agricultural Science and Engineering, Hohai University, Nanjing 210024, China
[4]Department of Chemistry, University of Lahore, Sargodha Campus, Sargodha 40100, Pakistan

[1]najmafarid11@gmail.com, [2]montahakanwal@gmail.com *[3] safi.rehman.mkw@gmail.com
[4]aqeelakanwal6@gmail.com

DOI: https://doi.org/

Keywords
Machine Learning; Specialized Metabolites; Chemometrics; Model Validation; Metabolomics; Medicinal Plants

Article History
Received on 20 Jan 2026
Accepted on 20 Mar 2026
Published on 30 Mar 2026

Copyright @Author

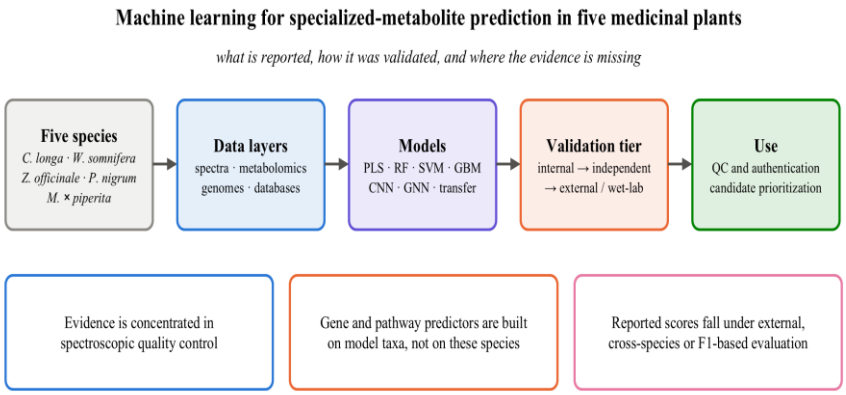
Corresponding Author: *

Saffi Ur Rehman
College of Agricultural Science and Engineering, Hohai University, Nanjing 210024, China Email: safi.rehman.mkw@gmail.com

Abstract

Machine learning (ML) is widely used to quantify, annotate and prioritize plant specialized metabolites, but reported performance is rarely comparable across studies because validation designs differ. We critically reviewed ML applications published between 2021 and 2026, using five ethnomedicinal species with contrasting chemistry as case studies: *Curcuma longa*, *Withania somnifera*, *Zingiber officinale*, *Piper nigrum* and *Mentha* × *piperita*. Each study was coded by task, data type, algorithm, validation design and reported metric, and graded on a five-tier validation scale. Species-level evidence was concentrated in spectroscopic quantification, authentication and extraction optimization, almost always with internal validation only; no ML-based biosynthetic-gene or pathway prediction was found for any of the five species. The most transferable methods, namely multi-omics gene predictors, pathway classifiers and mass-spectrometry class annotators, were developed on model or reference taxa. Their scores declined under stricter evaluation: gene-prediction AUC-ROC fell from 0.924 within species to 0.584 in an external taxon. We propose minimum validation and reporting practices and research priorities for transferring these methods to medicinal species.

Graphical Abstract



INTRODUCTION

(SVR), decision-tree, random-forest and XGBoost models for piperine content (Rahaman et al. 2025), a convolutional neural network (CNN) combined with cyclic voltammetry for piperine prediction (Banerjee et al. 2024, preprint), hybrid deep learning on hyperspectral images to detect papaya-seed adulteration (Balakrishnan et al. 2026), and SVR for optimizing tannin extraction (Khajeh et al. 2025). Because piperine enhances the bioavailability of co-administered drugs, black pepper is also relevant to ML tools for assessing herb–drug interactions (Ding et al. 2026; Spanakis et al. 2025).

Mentha × *piperita* L. (peppermint). Peppermint oil consists mainly of menthol and menthone, while peppermint leaves contain flavonoid glycosides like eriocitrin (Hudz et al. 2023; Khameneh et al. 2025). ML classifiers applied to metabolic fingerprints have been used to authenticate *Mentha* oils (Ratnasekhar et al. 2025). The most relevant content-prediction model in the genus was developed for *M. spicata* (Van Haute et al. 2023) and has not been tested on peppermint. Elicitor-induced changes in menthol-pathway gene expression have been characterized (Khaldari et al. 2025) but not yet modelled with ML.

Data resources

Model performance is bounded by the training data. Table 1 compares the main resources by contribution: general chemistry and bioactivity databases, natural-product structure–organism databases, a regional medicinal-plant collection and tandem mass-spectrometry libraries. Genomic platforms now link metabolite records to functional annotation in medicinal plants (MedMeta; Meng et al. 2025) and detect physically clustered biosynthetic genes (plantiSMASH; Kautsar et al. 2017). Common limitations include inconsistent identifiers, sparse quantitative data, missing tissue and growth metadata, and taxonomic bias (Rutz et al. 2022; Street 2025). Public data deposition remains inconsistent (Gouveia et al. 2024); reporting to Metabolomics Standards Initiative (MSI) requirements (Sumner et al. 2007) and benchmarking guidelines for annotation tools (de Jonge et al. 2022) help address this.

Table 1 Data resources for ML-based plant specialized-metabolite research

Resource	Content	Scope (release cited)	Main limitation	Reference
PubChem	General chemical structures and bioassays	>119 million compounds	Mostly synthetic; organism of origin often absent	Kim et al. 2025
ChEMBL	Curated bioactivity of drug-like molecules	Millions of measurements	Little pathway or source-organism metadata	Zdrazil et al. 2024
COCONUT 2.0	Open natural-product structures with organism links	695,133 structures	Quantitative activity sparse	Chandrasekhar et al. 2025
LOTUS	Referenced structure-organism pairs	>750,000 pairs	Occurrence only; no concentrations	Rutz et al. 2022

Resource	Content	Scope (release cited)	Main limitation	Reference
KNApSAcK	Species-metabolite relationships	Tens of thousands of pairs	Biased toward well-studied taxa	Shinbo et al. 2006
NPASS	Natural-product activity with species source	Regularly updated	Incomplete for non-model species	Zhao et al. 2023
IMPPAT 2.0	Indian medicinal plants and phytochemicals	4,010 plants; 17,967 compounds	No quantitative or chemotype data	Vivek-Ananth et al. 2023
GNPS	Public MS/MS data and spectral libraries	Community repository	Low coverage of plant chemistry	Wang et al. 2016

ML approaches and their reported performance

Table 2 summarizes the algorithm families used in this field, and Fig. 2 places them in a pipeline with evaluation and experimental confirmation. Table 3 reports quantitative results with their validation designs, and Fig. 3 shows the highest level of validation accomplished for each species and task.

Table 2 Algorithm families used for plant specialized-metabolite prediction

Family	Typical task	Strengths	Limitations
PLS regression / PLS-DA	Content prediction from NIR, Raman and hyperspectral data	Robust with collinear spectra and small samples; interpretable	Linear; sensitive to preprocessing and instrument drift
Random forest, gradient boosting	Gene and compound classification; QSAR; spectral regression	Handle many correlated features; feature importance	Overfit redundant data under random splits
Support vector machine / regression	Nonlinear tasks on small datasets	Effective with moderate sample size	Sensitive to hyperparameters; opaque
Convolutional neural network	Images, spectra, sensor signals	Automatic feature extraction	Data- and compute-intensive
Graph neural network / transformer	Pathway and property prediction from molecular graphs	Encodes atom- and bond-level structure	High data demand; split-sensitive
Spectrum-to-class deep models	Class annotation of unknowns from MS/MS	Work without reference spectra	Depend on fragmentation coverage

IEER
Institute for Excellence in Education & Research

Family	Typical task	Strengths	Limitations
Transfer and multi-species learning	Extending gene predictors to data-poor species	Reduce labelled-data requirements	Degrade with phylogenetic distance

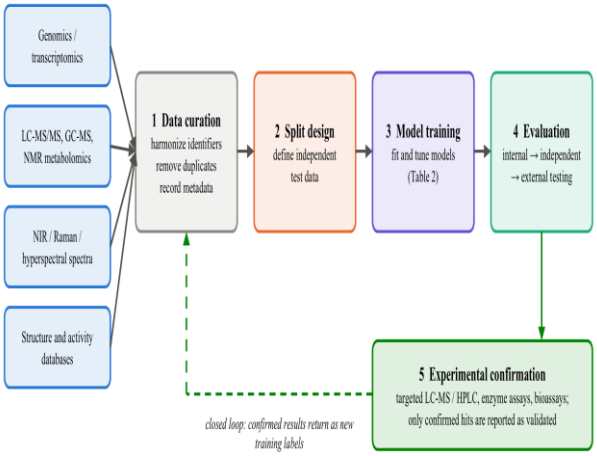


Fig. 2 Pipeline for ML-based specialized-metabolite prediction. Confirmed experimental results feed back into data curation as new training labels

Chemometric and classical models

PLS regression remains the standard for spectroscopic quantification and the benchmark against which other models (SVR, random forests, gradient boosting and small multilayer perceptrons) are typically compared (Samrat et al. 2022; Van Haute et al. 2023; Rahaman et al. 2025). Random forests (Breiman 2001) and support vector machines (Cortes and Vapnik 1995) suit small tabular datasets. On multi-omics gene features, tree-based ensembles outperformed neural networks, and a model with five interpretable features performed as well as the full 234-feature model (Bai et al. 2024). With only tens of samples, the size and independence of the test set matter more than the choice of algorithm (e.g. Khajeh et al. 2025).

Deep learning and graph-based models

Convolutional neural networks (CNNs) extract features from images and spectra, transformers learn long-range dependencies, and graph neural networks (GNNs) work directly with molecular graphs (Choi and Lee 2023; Zhang et al. 2025). A graph transformer combined with a SMILES-based CNN classified plant compounds into 18 pathway classes, outperforming random-forest and graph-convolution baselines (Bao et al. 2023), and image-based CNNs have predicted metabolite-content classes of neem fruit (Krishnan et al. 2024). These models require more labelled data than are available for most medicinal species (Alzubaidi et al. 2023).

Mass-spectrometry-based annotation

ML has had its most practical impact on plant chemistry through the annotation of untargeted LC-MS/MS data. MS2DeepScore enhances analogue search (Huber et al. 2021), CANOPUS assigns compound classes to spectra of compounds not found in the libraries (Dührkop et al. 2021), and NPClassifier assigns structures to pathways, superclasses and classes (Kim et al. 2021). These tools are taxon-agnostic, but none has been specifically evaluated on the five species.

Biosynthetic-gene prediction and transfer learning

For *Arabidopsis thaliana*, models based on duplication, conservation, expression and network features captured 87% of the known specialized-metabolism genes (Moore et al. 2019). The critical step is transfer to data-poor species: using an *Arabidopsis* model to filter mis-annotated tomato genes substantially improved tomato predictions (Moore et al. 2020), and models trained on three species performed as well as or better than within-species models, whereas a model trained on tomato alone performed markedly worse on *Arabidopsis* (Bai et al. 2024). Few-shot and transfer methods are therefore promising for

medicinal plants (Street 2025), provided they are validated on phylogenetically distant targets.

Cross-study comparison

Two patterns emerge. First, species-level evidence is concentrated in spectroscopic quantification, authentication and extraction optimization, almost exclusively at tier 2 (Fig. 3). No species-level study reaches tier 4; the only tier-3 study is the turmeric work evaluated on an independent prediction set with benchtop and portable instruments (Khongkaew et al. 2022). No ML method has been applied to the biosynthesis of curcuminoids, withanolides, gingerols or piperine in these species.



Fig. 3 Highest validation tier reached by ML studies of the five case-study species and of model or reference taxa, by task. Tier 0 indicates that no study was located in the reviewed corpus. *Study on *Mentha spicata*, the closest congener.

Second, reported performance depends strongly on evaluation design (Table 3). The same gene-prediction method yielded an AUC-ROC of 0.924 within *Arabidopsis*, 0.785 when trained on tomato alone and 0.584 in the external species *Papaver somniferum* (Bai et al. 2024). In multi-label pathway classification, an accuracy of 97.8% corresponded to an F1 of only 0.837 because accuracy was dominated by correctly predicted negative labels (Bao et al. 2023), and the mean average precision of NPClassifier fell from 0.992 at the pathway level to 0.895 at the class level (Kim et al. 2021). For chemometric models, RPD values below about 3 (1.9–2.8 in the studies reviewed here) support screening but not quality control or pharmacopoeial quantification (Williams 2014), and the robustness of single-harvest calibrations across seasons and suppliers is unknown. Headline accuracies above 95% therefore say little about performance on an under-studied species.

Table 3 Reported performance of representative ML studies and their validation design

Study	Taxon	Task	Data (n)	Model	Validation design	Reported performance
Moore et al. 2019	*Arabidopsis thaliana*	Specialized vs general metabolism genes	Multi-omic gene features	ML classifiers	Held-out genes	TPR 87%, TNR 71%

Study	Taxon	Task	Data (n)	Model	Validation design	Reported performance
Moore et al. 2020	*Solanum lycopersicum*	Specialized-metabolism genes	Tomato genes; Arabidopsis source model	Transfer learning	Held-out test set	F 0.74 → 0.92 after cross-species filtering
Bai et al. 2024	Arabidopsis, tomato, maize; poppy external	Biosynthetic genes	594 gold-standard genes; 234 features	AutoGluon ensemble	80/20 split; cross-species; external taxon	AUC 0.924 within; 0.785 tomato-trained; 0.584 poppy
Bao et al. 2023	Plant compounds (KEGG)	Pathway class (18 labels)	2,028 compounds; test 577	Graph transformer + CNN	5-fold CV; independent test	Test accuracy 97.8 %, F1 0.837
Kim et al. 2021	Natural products	Pathway / superclass / class	73,607 structures	Deep neural network	Test split; external DNP set	mAP 0.992 / 0.952 / 0.895
Dührkop et al. 2021	Natural products and metabolites	Compound class from MS/MS	2,497 classes	Deep neural network	Cross-validation	Mean accuracy 99.7 %
Krishnan et al. 2024	*Azadirachta indica*	Metabolite-content class from images	1,045 samples	CNNs	Validation and test sets	F1 0.93 (validation), 0.88 (test)
Khongkaew et al. 2022	*Curcuma longa*	Curcuminoid content	Turmeric powders	PLS (NIR, Raman)	Prediction set; two instrument types	RMSEP 0.41 % (NIR), 0.44 % (Raman) w/w
Samrat et al. 2022	*Zingiber officinale*	6-Gingerol: 6-shogaol ratio	89 samples	PLSR, LASSO	80/20 random split	$R^2 \geq 0.73$; RPD ≥ 1.92

Study	Taxon	Task	Data (n)	Model	Validation design	Reported performance
Van Haute et al. 2023	*Mentha spicata*	Essential-oil content	NIR hyperspectral images	PLS + MLP	Repeated prediction sets	R^2p 0.863; RPD 2.84
Khajeh et al. 2025	*Piper nigrum*	Tannin extraction yield	30 runs	SVR-RSM	Train/test split	RMSE 0.035; NSE 0.91
Ronen et al. 2025	*Withania somnifera*	Root localization of constituents	>600 training points	Random forest	Internal validation	Accuracy 0.84; AUC 0.98

TPR/TNR, true-positive/true-negative rate; F, F-measure; mAP, mean average precision; DNP, Dictionary of Natural Products; RMSEP, root-mean-square error of prediction; RPD, ratio of performance to deviation; NSE, Nash–Sutcliffe efficiency; AUC, area under the receiver operating characteristic curve; CV, cross-validation; PLSR, partial least squares regression; LASSO, least absolute shrinkage and selection operator; MLP, multilayer perceptron; SVR-RSM, support vector regression with response surface methodology; R^2p, coefficient of determination for the prediction set.

Validation and reporting practice

Table 4 lists pitfalls common in this literature; the recommended practices draw on the DOME recommendations for supervised ML in biology (Walsh et al. 2021), scaffold-aware splitting (Wu et al. 2018), and the OECD principles for QSAR validation (OECD 2007). At a minimum, studies should report sample origin and independence, the split procedure, a class-balanced metric or RMSEP and RPD, results on an independent or external set with the validation tier stated, and deposited data and code. Docking and network-pharmacology results should be reported separately from ML results.

Table 4 Common validation pitfalls and recommended practice

Pitfall	Consequence	Recommended practice	Basis
Random split of structurally redundant compounds	Near-duplicates leak into the test set	Scaffold or cluster split	Wu et al. 2018
Accuracy on imbalanced or multi-label outputs	Accuracy near ceiling while F1 is lower	Report per-class F1, MCC or PR-AUC	Walsh et al. 2021
Small samples with one random split	Unstable, optimistic error estimates	Independent set from another harvest, site or instrument; report RMSEP and RPD	Walsh et al. 2021
Within-species evaluation only	Overstated transfer to medicinal species	Leave-species-out and external-taxon tests	Moore et al. 2020
Noisy training labels	Performance ceiling and spurious errors	Curated gold standards; state annotation source	Rutz et al. 2022

Pitfall	Consequence	Recommended practice	Basis
No applicability domain	Silent extrapolation	Define and report the applicability domain	OECD 2007
No data or code release	Results cannot be reproduced	Deposit data, code and models	Walsh et al. 2021

MCC, Matthews correlation coefficient; PR-AUC, area under the precision–recall curve; RMSEP, root-mean-square error of prediction; RPD, ratio of performance to deviation.

Research priorities and conclusions

Overall, the main bottleneck is the scarcity of species-level data and external evaluation rather than algorithms (Zhang and Wang 2023; Murmu et al. 2024). Four priorities follow. First, paired multi-omics and targeted quantification datasets with MSI-compliant metadata are needed for under-studied species, particularly *W. somnifera*. Second, existing gene predictors and spectral annotators should be applied to the five species and evaluated at tier 4. Third, spectroscopic models should be calibrated and validated across sites, seasons and instruments. Fourth, ML prioritization should be coupled with targeted LC-MS, enzyme assays and bioassays, and confirmed predictions should be fed back into training (Msobo et al. 2026). At similar accuracy, inherently interpretable models are preferable because their outputs can be traced to testable gene–enzyme–metabolite hypotheses (Rudin 2019).

In conclusion, ML has already proved useful for quantifying and authenticating the specialized metabolites of turmeric, ginger, black pepper and peppermint, and for annotating metabolomes and ranking biosynthetic genes in model species. However, evidence for the five species is largely limited to internally validated quality control, ML-based biosynthetic-gene prediction is absent, and *W. somnifera* has very few predictive studies. Because scores decline markedly under external and class-balanced testing, ML predictions should be confirmed by analytical chemistry and bioassays.

References

Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, Duan Y, Abdullah A, Farhan L, Lu Y, Gupta A, Albu F, Abbosh A, Gu Y (2023) A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data 10(1):46. https://doi.org/10.1186/s40537-023-00727-2

Ashokkumar K, Murugan M, Dhanya MK, Pandian A, Warkentin TD (2021) Phytochemistry and therapeutic potential of black pepper [*Piper nigrum* (L.)] essential oil and piperine: a review. Clin Phytosci 7(1):52. https://doi.org/10.1186/s40816-021-00292-2

Atanasov AG, Zotchev SB, Dirsch VM, Orhan IE, Banach M, Rollinger JM, Barreca D, Weckwerth W, Bauer R, Bayer EA, Majeed M, Bishayee A, Bochkov V, Bonn GK, Braidy N, Bucar F, Cifuentes A, D'Onofrio G, Bodkin M, Diederich M, Dinkova-Kostova AT, Efferth T, El Bairi K, Arkells N, Fan TP, Fiebich BL, Freissmuth M, Georgiev MI, Gibbons S, Godfrey KM, Gruber CW, Heer J, Huber LA, Ibanez E, Kijjoa A, Kiss AK, Lu A, Macias FA, Miller MJS, Mocan A, Müller R, Nicoletti F, Perry G, Pittalà V, Rastrelli L, Ristow M, Russo GL, Silva AS, Schuster D, Sheridan H, Skalicka-Woźniak K, Skaltsounis L, Sobarzo-Sánchez E, Bredt DS, Stuppner H, Sureda A,

Tzvetkov NT, Vacca RA, Aggarwal BB, Battino M, Giampieri F, Wink M, Wolfender JL, Xiao J, Yeung AWK, Lizard G, Popp MA, Heinrich M, Berindan-Neagoe I, Stadler M, Daglia M, Verpoorte R, Supuran CT (2021) Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 20(3):200–216. https://doi.org/10.1038/s41573-020-00114-z

Bai W, Li C, Li W, Wang H, Han X, Wang P, Wang L (2024) Machine learning assists prediction of genes responsible for plant specialized metabolite biosynthesis by integrating multi-omics data. BMC Genomics 25(1):418. https://doi.org/10.1186/s12864-024-10258-6

Balakrishnan SB, Padmanaban P, Malvannan L (2026) PiperNet: a hybrid deep learning approach for monitoring papaya seed adulteration in black pepper using hyperspectral imaging. Food Addit Contam Part A Chem Anal Control Expo Risk Assess 43(1):15–31. https://doi.org/10.1080/19440049.2025.2598389

Banerjee S, Ghorai S, Dhara M, Naskar H, Ali SB, Das N, Saha P, Tudu B, Chatterjee A, Bandyopadhyay R, Tudu B (2024) Accurate prediction of piperine content in black pepper using combined CNN and regression modelling with Padmam@G electrode and cyclic voltammetry. SSRN preprint. https://doi.org/10.2139/ssrn.5046816

Bao H, Zhao J, Zhao X, Zhao C, Lu X, Xu G (2023) Prediction of plant secondary metabolic pathways using deep transfer learning. BMC Bioinformatics 24:348 https://doi.org/10.1186/s12859-023-05485-9

Bhatla SC, Lal MA (2023) Secondary metabolites. In: Plant physiology, development and metabolism. Springer, Singapore, pp 765–808. https://doi.org/10.1007/978-981-99-5736-1_33

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Chandrasekhar V, Rajan K, Kanakam SRS, Sharma N, Weißenborn V, Schaub J, Steinbeck C (2025) COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. Nucleic Acids Res 53(D1):D634–D643. https://doi.org/10.1093/nar/gkae1063

Choi SR, Lee M (2023) Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology 12(7):1033. https://doi.org/10.3390/biology12071033

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/BF00994018

Das D, Shafi S (2023) Bioactivity-guided fractionation and identification of bioactive molecules: a basic method in drug discovery. In: Drugs and a methodological compendium: from bench to bedside. Springer, Singapore, pp 41–78. https://doi.org/10.1007/978-981-19-7952-1_3

de Jonge NF, Mildau K, Meijer D, Louwen JJR, Bueschl C, Huber F, van der Hooft JJJ (2022) Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. Metabolomics 18(12):103. https://doi.org/10.1007/s11306-022-01963-y

Dhaniaputri R, Suwono H, Amin M, Lukiati B (2022) Introduction to plant metabolism, secondary metabolites biosynthetic pathway, and in-silico

molecular docking for determination of plant medicinal compounds: an overview. In: Proceedings of the 7th International Conference on Biological Science (ICBS 2021). Adv Biol Sci Res 22:373–382. https://doi.org/10.2991/absr.k.220406.053

Ding L, Fei T, Xiao J, Wang L (2026) Unveiling the multifaceted potential of *Piper nigrum*: a comprehensive review of its chemical components, health benefits, safety assessment and industrial application. J Agric Food Chem 74(4):3353–3376. https://doi.org/10.1021/acs.jafc.5c15160

Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, Petras D, Gerwick WH, Rousu J, Dorrestein PC, Böcker S (2021) Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol 39(4):462–471. https://doi.org/10.1038/s41587-020-0740-8

Elshafie HS, Camele I, Mohamed AA (2023) A comprehensive review on the biological, agricultural and pharmaceutical properties of secondary metabolites based-plant origin. Int J Mol Sci 24(4):3266. https://doi.org/10.3390/ijms24043266

Fuloria S, Mehta J, Chandel A, Sekar M, Rani NNIM, Begum MY, Subramaniyan V, Chidambaram K, Thangavelu L, Nordin R, Wu YS, Sathasivam KV, Lum PT, Meenakshi DU, Kumarasamy V, Azad AK, Fuloria NK (2022) A comprehensive review on the therapeutic potential of *Curcuma longa* Linn. in relation to its major active constituent curcumin. Front Pharmacol 13:820806. https://doi.org/10.3389/fphar.2022.820806

Gouveia GJ, Head T, Cheng LL, Clendinen CS, Cort JR, Du X, Edison AS, Fleischer CC, Hoch J, Mercaldo N, Pathmasiri W, Raftery D, Schock TB, Sumner LW, Takis PG, Copié V, Eghbalnia HR, Powers R (2024) Perspective: use and reuse of NMR-based metabolomics data: what works and what remains challenging. Metabolomics 20(2):41 https://doi.org/10.1007/s11306-024-02090-6

Hilal B, Khan MM, Fariduddin Q (2024) Recent advancements in deciphering the therapeutic properties of plant secondary metabolites: phenolics, terpenes, and alkaloids. Plant Physiol Biochem 211:108674. https://doi.org/10.1016/j.plaphy.2024.108674

Huber F, van der Burg S, van der Hooft JJJ, Ridder L (2021) MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. J Cheminform 13(1):84. https://doi.org/10.1186/s13321-021-00558-4

Hudz N, Kobylinska L, Pokajewicz K, Horčinová Sedláčková V, Fedin R, Voloshyn M, Myskiv I, Brindza J, Wieczorek PP, Lipok J (2023) *Mentha piperita*: essential oil and extracts, their biological activities, and perspectives on the development of new medicinal and cosmetic products. Molecules 28(21):7444. https://doi.org/10.3390/molecules28217444

Jan R, Asaf S, Numan M, Lubna, Kim KM (2021) Plant secondary metabolite biosynthesis and transcriptional regulation in response to biotic and abiotic stress conditions. Agronomy 11(5):968. https://doi.org/10.3390/agronomy11050968

Kaur K, Dolker D, Behera S, Pati PK (2022) Critical factors influencing in vitro propagation and modulation of important secondary metabolites in *Withania somnifera* (L.) Dunal. Plant Cell Tissue Organ Cult 149(1–2):41–60. https://doi.org/10.1007/s11240-021-02225-w

Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res 45(W1):W55–W63. https://doi.org/10.1093/nar/gkx305

Khajeh M, Ghaffari-Moghaddam M, Piri J, Barkhordar A, Şenol H, Saloglu D (2025) Using microwave-assisted extraction with advanced artificial intelligence models for predicting tannins in black pepper (*Piper nigrum* L.). J Appl Res Med Aromat Plants 44:100594. https://doi.org/10.1016/j.jarmap.2024.100594

Khaldari I, Afshoon E, Nik SH (2025) Phytohormonal elicitation triggers oxidative stress and enhances menthol biosynthesis through modulation of key pathway genes in *Mentha piperita* L. Sci Rep 15(1):30495. https://doi.org/10.1038/s41598-025-16667-1

Khalid MU, Sultan MT, Baig I, Abbas A, Noman AM, Zinedine A, Bartkiene E, Rocha JM (2025) A comprehensive review on the bioactivity and pharmacological attributes of *Withania somnifera*. Nat Prod Res. https://doi.org/10.1080/14786419.2025.2499070

Khameneh AN, Eghlima G, Ranjbar ME, Ghorbanpour M (2025) Exploring the role of abiotic elicitors in the biosynthesis of secondary metabolites in peppermint (*Mentha piperita* L.). Acta Sci Pol Hortorum Cultus 24(6):137–154. https://doi.org/10.24326/ASPHC.2025.5402

Khan A, Kanwal F, Ullah S, Fahad M, Tariq L, Altaf MT, Riaz A, Zhang G (2025) Plant secondary metabolites—central regulators against abiotic and biotic stresses. Metabolites 15(4):276. https://doi.org/10.3390/metabo15040276

Khongkaew P, Cruz J, Puig Bertotto J, Cárdenas V, Alcalà M, Nuchtavorn N, Phechkrajang C (2022) A comparative study of benchtop and portable NIR and Raman spectroscopic methods for the quantitative determination of curcuminoids in turmeric powder. Foods 11(15):2187. https://doi.org/10.3390/foods11152187

Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, van der Hooft JJJ, Dorrestein PC, Gerwick WH, Cottrell GW (2021) NPClassifier: a deep neural network-based structural classification tool for natural products. J Nat Prod 84(11):2795–2807. https://doi.org/10.1021/acs.jnatprod.1c00399

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2025) PubChem 2025 update. Nucleic Acids Res 53(D1):D1516–D1525. https://doi.org/10.1093/nar/gkae1059

Koirala M, Yan L, Mohamed Z, DiPaola M (2025) AI-integrated QSAR modeling for enhanced drug discovery: from classical approaches to deep learning and structural insight. Int J Mol Sci 26(19):9384. https://doi.org/10.3390/ijms26199384

Krishnan NM, Kumar S, Panda B (2024) Fruit-In-Sight: a deep learning-based framework for secondary metabolite class prediction using fruit and leaf

images. PLoS One 19(8):e0308708. https://doi.org/10.1371/journal.pone.0308708

Latif R, Nawaz T (2025) Medicinal plants and human health: a comprehensive review of bioactive compounds, therapeutic effects, and applications. Phytochem Rev 25(3):2299–2342. https://doi.org/10.1007/s11101-025-10194-7

Li H, Guo Z, Xu M, Zhao J, Xu D (2024) Molecular mechanism of miRNA mediated biosynthesis of secondary metabolites in medicinal plants. Plant Physiol Biochem 208:108524. https://doi.org/10.1016/j.plaphy.2024.108524

Mao QQ, Xu XY, Cao SY, Gan RY, Corke H, Beta T, Li HB (2019) Bioactive compounds and bioactivities of ginger (*Zingiber officinale* Roscoe). Foods 8(6):185. https://doi.org/10.3390/foods8060185

Meng F, Zhang G, Xiao W, Mao Y, Shu Y, Yang X, Xu G, Tang X, Zhang M, Liu Z, Zhang X, You S, Wang B, Yu Z, Chen S, Chen W (2025) MedMeta: an AI-enabled and genomics-based database for functional profiling of secondary metabolites in medicinal species. Engineering. https://doi.org/10.1016/j.eng.2025.09.007

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu SH (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci U S A 116(6):2344–2353. https://doi.org/10.1073/pnas.1817074116

Moore BM, Wang P, Fan P, Lee A, Leong B, Lou YR, Schenck CA, Sugimoto K, Last R, Lehti-Shiu MD, Barry CS, Shiu SH (2020) Within- and cross-species predictions of plant specialized metabolism genes using transfer learning. In Silico Plants 2(1):diaa005. https://doi.org/10.1093/insilicoplants/diaa005

Msobo A, Maphari PW, Koorsen G, Singab ANB, Mhlongo MI (2026) Modernising antiviral drug discovery: harnessing medicinal plants through machine learning and metabolomics to target the SARS-CoV-2 main protease. In Silico Pharmacol 14(2):120. https://doi.org/10.1007/s40203-026-00620-9

Murmu S, Sinha D, Chaurasia H, Sharma S, Das R, Jha GK, Archak S (2024) A review of artificial intelligence-assisted omics techniques in plant defense: current trends and future directions. Front Plant Sci 15:1292054. https://doi.org/10.3389/fpls.2024.1292054

OECD (2007) Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD Series on Testing and Assessment, No. 69. OECD Publishing, Paris

Paul A, Acharya K, Chakraborty N (2023) Involvement of phenylpropanoid pathway and shikimic acid pathway in environmental stress response. In: Biology and biotechnology of environmental stress tolerance in plants, vol 1. CRC Press, pp 27–66. https://doi.org/10.1201/9781003346173-3

Rahaman SH, Banerjee S, Ali SB, Tudu B, Das N, Naskar S, Ghorai S, Chatterjee A, Bandyopadhyay R, Tudu B (2025) Regression based prediction of piperine content in black pepper using near infrared spectroscopy. J Near Infrared Spectrosc 33(3):64–76. https://doi.org/10.1364/JNIRS.33.000064

Ratnasekhar CH, Khan S, Rai AK, Mishra H, Verma AK, Lal RK, Ananda Kumar TM, Elliott CT (2025) Rapid metabolic fingerprinting meets machine learning models to identify authenticity and detect adulteration of essential oils with vegetable oils: *Mentha* and *Ocimum* study. Food Chem 471:142709. https://doi.org/10.1016/j.foodchem.2024.142709

Reshi ZA, Ahmad W, Lukatkin AS, Javed SB (2023) From nature to lab: a review of secondary metabolite biosynthetic pathways, environmental influences, and in vitro approaches. Metabolites 13(8):895. https://doi.org/10.3390/metabo13080895

Ronen Y, Ebert C, Tamim-Yecheskel BC, Zev S, Kantor O, Arbel HBH (2025) Comprehensive safety evaluation of *Withania somnifera* (Ashwagandha): an AI-driven meta-analysis and quantitative structure–activity relationship based toxicity assessment. Front Nutr 12:1658265. https://doi.org/10.3389/fnut.2025.1658265

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, Graham JG, Stephan R, Page R, Vondrášek J, Steinbeck C, Pauli GF, Wolfender JL, Bisson J, Allard PM (2022) The LOTUS initiative for open knowledge management in natural products research. eLife 11:e70780. https://doi.org/10.7554/eLife.70780

Samrat NH, Johnson JB, White S, Naiker M, Brown P (2022) A rapid non-destructive hyperspectral imaging data model for the prediction of pungent constituents in dried ginger. Foods 11(5):649. https://doi.org/10.3390/foods11050649

Shinbo Y, Nakamura Y, Altaf-Ul-Amin M, Asahi H, Kurokawa K, Arita M, Saito K, Ohta D, Shibata D, Kanaya S (2006) KNApSAcK: a comprehensive species-metabolite relationship database. In: Plant metabolomics. Springer, Berlin, Heidelberg, pp 165–181. https://doi.org/10.1007/3-540-29782-0_13

Spanakis M, Tzamali E, Tzedakis G, Koumpouzi C, Pediaditis M, Tsatsakis A, Sakkalis V (2025) Artificial intelligence models and tools for the assessment of drug–herb interactions. Pharmaceuticals 18(3):282. https://doi.org/10.3390/ph18030282

Street NR (2025) Data-driven resources and computational tools in non-model plant species. Physiol Plant 177(1):e70133. https://doi.org/10.1111/ppl.70133

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3(3):211–221. https://doi.org/10.1007/s11306-007-0082-2

Tian WW, Liu L, Chen P, Yu DM, Li QM, Hua H, Zhao JN (2025) *Curcuma longa* (turmeric): from traditional applications to modern plant medicine research hotspots. Chin Med 20:76 https://doi.org/10.1186/s13020-025-01115-z

Van Haute S, Nikkhah A, Malavi D, Kiani S (2023) Prediction of essential oil content in spearmint (*Mentha spicata*) via near-infrared hyperspectral imaging and chemometrics. Sci Rep 13(1):4261.

https://doi.org/10.1038/s41598-023-31517-8

Varghese R, Shringi H, Efferth T, Ramamoorthy S (2025) Artificial intelligence driven approaches in phytochemical research: trends and prospects. Phytochem Rev 24(5):3649–3664. https://doi.org/10.1007/s11101-025-10096-8

Vivek-Ananth RP, Mohanraj K, Sahoo AK, Samal A (2023) IMPPAT 2.0: an enhanced and expanded phytochemical atlas of Indian medicinal plants. ACS Omega 8(9):8827–8845. https://doi.org/10.1021/acsomega.3c00156

Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, ELIXIR Machine Learning Focus Group, Harrow J, Psomopoulos FE, Tosatto SCE (2021) DOME: recommendations for supervised machine learning validation in biology. Nat Methods 18(10):1122–1127. https://doi.org/10.1038/s41592-021-01205-4

Wang L, Chen M, Lam PY, Dini-Andreote F, Dai L, Wei Z (2022) Multifaceted roles of flavonoids mediating plant-microbe interactions. Microbiome 10(1):233. https://doi.org/10.1186/s40168-022-01420-x

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, et al. (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34(8):828–837. https://doi.org/10.1038/nbt.3597

Williams P (2014) The RPD statistic: a tutorial note. NIR News 25(1):22–26. https://doi.org/10.1255/nirn.1419

Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530. https://doi.org/10.1039/C7SC02664A

Ye Y, Zhang X, Chen X, Xu Y, Liu J, Tan J, Li W, Tembrock LR, Wu Z, Zhu G (2022) The use of widely targeted metabolomics profiling to quantify differences in medicinally important compounds from five *Curcuma* (Zingiberaceae) species. Ind Crops Prod 175:114289. https://doi.org/10.1016/j.indcrop.2021.114289

Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR (2024) The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52(D1):D1180–D1192. https://doi.org/10.1093/nar/gkad1004

Zhang O, Lin H, Zhang X, Wang X, Wu Z, Ye Q, Zhao W, Wang J, Ying K, Kang Y, Hsieh CY, Hou T (2025) Graph neural networks in modern AI-aided drug discovery. Chem Rev 125(20):10001–10103. https://doi.org/10.1021/acs.chemrev.5c00461

Zhang Y, Wang Y (2023) Recent trends of machine learning applied to multi-source data of medicinal plants. J Pharm Anal 13(12):1388–1407. https://doi.org/10.1016/j.jpha.2023.07.012

Zhao H, Yang Y, Wang S, Yang X, Zhou K, Xu C, Zhang X, Fan J, Hou D, Li X, Lin H, Tan Y, Wang S, Chu XY, Zhuoma D, Zhang F, Ju D, Zeng X, Chen YZ (2023) NPASS database update 2023: quantitative natural product activity and species source database for biomedical research. Nucleic Acids Res

51(D1):D621–D628. https://doi.org/10.1093/nar/gkac1069

Zhaogao L, Yaxuan W, Mengwei X, Haiyu L, Lin L, Delin X (2023) Molecular mechanism overview of metabolite biosynthesis in medicinal plants. Plant Physiol Biochem 204:108125. https://doi.org/10.1016/j.plaphy.2023.108125

IEER
Institute for Excellence in Education & Research