Pharmedico Publishers
International Journal of Pharmaceutical Science and Medicine
Volume 4, Issue 1 (2026): 13-22
DOI: 10.70199/IJPSM.4.1.13-22

REVIEW ARTICLE

ISSN: 2584-1610 (Online)

AI-ASSISTED PHYTOCHEMICAL DEREPLICATION: INTEGRATION OF LC–MS/MS, MOLECULAR NETWORKING AND MACHINE LEARNING FOR NATURAL PRODUCT DISCOVERY

Khushboo Saxena[*1], Hema Choudhary[2], Amisha Vyas[3]

Assistant Professor[1], Dean[2], School of Medical and Allied Sciences, K. R. Mangalam University, Gurugram - 122103, Haryana, India.
Assistant Professor[3], School of Pharmaceutical Sciences, Apeejay Stya University, Sohna, Gurugram - 122103, Haryana, India.

Article History
Received: 07 January 2026
Revised: 28 January 2026
Accepted: 23 February 2026
Published: 25 March 2026

Correspondence should be addressed to
Dr. Khushboo Saxena,
Assistant Professor,
School of Medical and Allied Sciences, K. R. Mangalam University, Gurugram - 122103, Haryana, India.

Email- khushbusaxena1987@gmail.com

ABSTRACT: Natural products are still vital sources of structurally diverse and biologically active molecules; however, these products are often hard to discover due to the complexity of natural extracts and the isolation of known metabolites. Phytochemical dereplication is an effective tool for identification of known constituents prior to extensive purification and the selection of metabolites to study for possible chemical or biological novelty. The ability for rapid phytochemical investigation has greatly increased in the last few years with the development of liquid chromatography–tandem mass spectrometry (LC–MS/MS), molecular networking (MN), and artificial intelligence (AI). The accurate mass, chromatographic and fragmentation data obtained by LC–MS/MS is then used to correlate similar metabolites into molecular families through molecular networking, and to provide chemical context for unannotated features. Additional support can be provided by machine learning (ML) and AI for spectral classification, chemical-class prediction, candidate-structure generation, similarity assessment and bioactivity based prioritization. The integration then enables metabolite detection, computational metabolite annotation, molecular family analysis, metabolite candidate ranking, targeted isolation and experimental validation to be coupled together in a workflow. However, some important limitations include incomplete spectral libraries, structural isomerism, instrumental variability, training-data bias, limited model interpretability and false annotation. This means that the predictions made from a computer model should be interpreted as a structural hypothesis, unless backed up by suitable experimental evidence. This review explores how LC–MS/MS, molecular networking and AI/ML complement each other in the process of phytochemical dereplication, how they are used in the discovery of natural products, and the challenges and opportunities of explainable AI, multimodal data integration, better spectral databases, and increasingly automated dereplication workflows.

Keywords: Phytochemical dereplication; LC–MS/MS; Molecular networking; Artificial intelligence; Machine learning; Natural product discovery

I. INTRODUCTION

Structurally diverse and physiologically active molecules are still being discovered from natural resources across the various research fields of pharmaceuticals, nutraceuticals and bioactive compounds. Plants produce a host of chemical compounds called specialised metabolites such as terpenoids, alkaloids, flavonoids, phenolics, glycosides, and steroids [1]. The chemical complexity of plant extracts is a major challenge for natural product discovery; a simple extract can contain hundreds of different metabolites with varying amounts, polarities and structural complexity. Given this, phytochemical dereplication is an important technique for rapid identification or annotation of known compounds, before intensive isolation and structure elucidation. This approach enables scientists to focus their research on chemically distinct and biologically relevant components, without having to repeatedly isolate the known metabolites [2]. In the conventional dereplication, chromatography, UV-Vis spectroscopy, precise mass measurements, literature searches, and NMR data are used. By giving precise mass, retention time, isotope patterns, and fragmentation information, high-resolution LC-MS and tandem MS have greatly enhanced phytochemical profiling. This analytic data is supplemented by molecular networking which groups metabolites together based on shared MS/MS properties and provides chemical information about unannotated metabolites [3]. AI and ML technologies have recently enabled novel approaches for spectrum classification, prediction of chemical class, structure candidate development, similarity assessment, and metabolite prioritization [4]. Therefore, LC-MS/MS, molecular networking, and AI/ML could help to create a more efficient and hypothesis-driven approach for phytochemical dereplication [5]. Issues of restricted model interpretability, training-data bias, etc. remain to be solved due to challenges like structural isomerism and the lack of spectrum libraries. Therefore, computational forecasts should be considered hypotheses and experimental validation is desirable if needed [6].

2. Phytochemical Dereplication: Principles and Challenges

For more complex samples of natural products, phytochemical dereplication can rapidly identify or annotate known metabolites before intensive isolation and structure elucidation efforts. Many chemicals can be found in natural extracts; however, only a select few of these molecules would be suitable for further purification and studies, increasing its value [7]. Conventional dereplication is built upon chromatographic retention, ultraviolet spectra, precise mass, published data, nuclear magnetic resonance (NMR) and reference standards. These molecular formulas can be related to different structural isomers, so the quality and availability of reference compounds and spectra remain crucial when using a database for identification [8].

MS/MS can sometimes help provide additional structural information in the form of neutral losses and unique product ions that can be used to help identify structural subunits and classes of compounds [9]. Structurally related metabolites can have very similar fragmentation patterns, even though the various parameters of the spectrum (collision energy, ionisation, instrument platforms, acquisition parameters) can be affected. Other variables that could cause uncertainty are co-elution, ion suppression, matrix effects and similar effects, aside from low abundance and in-source fragmentation [10].

Nowadays, dereplication is more effective when it uses a combination of evidence kinds instead of depending on just one analytical measurement. Classification, candidate creation, and prioritisation may be assisted by AI/ML, whereas molecular networking offers links among chemically linked characteristics. Prioritising metabolites with higher chemical or biological value and decreasing the number of candidates needing extensive experimental examination is the main goal [11].

3. LC–MS/MS for Phytochemical Profiling and Dereplication

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) is an indispensable analytical tool for phytochemical profiling and dereplication, which is a comprehensive characterization of a mass spectrum [12]. It is able to detect a wide variety of metabolites in complex plant extracts, making comprehensive isolation and structure elucidation unnecessary. The complete data generated by the chromatographic data, accurate masses, and fragmentation information can all benefit the molecular networking, spectrum annotation, and AI/ML-based candidate prioritization [13].

3.1 Analytical Basis

LC-MS/MS, a combination of chromatography and mass spectrometry, is a specialty used to analyze complex phytochemical mixtures. A chromatographic separation might separate metabolites, which have different physicochemical properties, easier, as well as it simplifies the sample before the detection of ions. Reversed-phase LC is frequently used to separate non-polar and moderately polar metabolites; metabolites that are highly polar may be better separated using complementary separations. The stationary phase, mobile phase composition, gradient parameters and chromatographic resolution chosen may affect metabolite coverage and detection qualities [14]. Electrospray ionisation (ESI) is a two-pronged technique, making it an excellent tool for detecting all kinds of metabolites in plants. High resolution mass spectrometers give very accurate mass, isotope patterns and intensities of features allowing the characterisation of specific metabolites by using the precursor m/z and retention time in combination with related MS/MS data [15].

These analytical parameters generate this basic chemical data used for computing, spectral matching, molecular networking, and metabolite annotation. Factors, like extraction efficiency, chromatographic separation, ionisation response, and instrumental conditions, can also influence metabolite coverage, as well as quality of phytochemical profiling of the samples, and therefore sample preparation and analytical conditions are also critical. Get trustworthy datasets for downstream phytochemical dereplication by optimising your analyses thoroughly [16].

3.2 MS/MS Fragmentation and Annotation

By fragmenting precursor ions, tandem MS may provide more structural information about functional groups or structural subunits via the use of distinctive product ions and neutral losses. Losses associated with sugars, water or carbon dioxide among others could be helpful in understanding metabolites that have been glycosylated or oxygenated. Many classes of phytochemicals can be better detected using fragmentation patterns, such as flavonoids, phenolic compounds, alkaloids and terpenoids [17]. Structural similarity of metabolites may result in structural conclusions that were not conclusively determined by MS/MS evidence because of stereochemical or molecular connectivity differences. Thus, the mass, retention behaviour, chemical knowledge, and when needed, orthogonal analytical methods should be taken into account when interpreting fragmentation results [18].

3.3 DDA, DIA and Computational Processing

Data-dependent acquisition (DDA) selects the precursor ion(s) to be fragmented based on parameters such as the signal strength. It can yield good quality MS/MS spectra, but may under represent the low abundance components because of its focus on abundant metabolites [19]. However, Data Independent acquisition (DIA) can also reduce the window of precursors for larger windows and can potentially increase the MS/MS coverage overall. However, to interpret the results with accuracy sophisticated computational deconvolution is required as there is the potential for obtaining mixed spectra containing signals from more than one co-eluting chemical (Table 1) [20].

After being acquired, the computer processing steps consist of spectrum matching, isotope grouping, adduct annotation, noise reduction, and peak identification. There are large scale feature processing tools (*e.g.,* MZmine, MS-DIAL) and spectrum comparison/annotation tools (*e.g.*, GNPS, structure oriented). This is because LC-MS/MS is not only for identification of compounds but also it offers a holistic view of the chemical structure which could be helpful in developing prioritisation, molecular interactions and candidate structures [21].

Table 1: LC–MS/MS approaches and computational tools used in phytochemical dereplication

Approach/Tool	*Main function*	*Major strength*	*Key limitation*
LC–HRMS	Accurate-mass profiling	High mass accuracy and broad coverage	Limited structural evidence without MS/MS [22]
LC–MS/MS	Fragmentation-based annotation	Provides structural information	Isomers may remain unresolved [23]
DDA	Selective precursor fragmentation	High-quality spectra	May under-sample low-abundance features [24]
DIA	Broad MS/MS acquisition	Greater feature coverage	Mixed spectra require deconvolution [25]
MZmine / MS-DIAL	Feature detection and processing	Flexible large-scale workflows	Parameter and database dependence [26]
GNPS	Spectral matching and networking	Community-based comparison	Incomplete spectral-library coverage [27]
SIRIUS / MS-FINDER	Formula and candidate-structure support	Integrates mass and fragmentation evidence	Candidate ambiguity may persist [28]

4. Molecular Networking for Natural Product Dereplication

Complex metabolite mixtures may be seen as linked molecular characteristics by molecular networking (MN), which organises tandem mass spectra based on their similarity. The molecular properties are represented by nodes in a molecular network, and the similarity between MS/MS spectra is indicated by the connections between the nodes [29]. Molecular families are a set of nodes that are interconnected in a sample and are displayed graphically. Many metabolites of similar structure are synthesized in nature and will produce similar fragmentation patterns as a result, which makes this structure very useful in the field of natural product research. Understanding neighbouring unknown properties and researching possibly related metabolites may be greatly aided by using a known chemical within a molecular family as a reference point [30].

The GNPS platform has been invaluable to natural-product researchers for their access to molecular networking, enabling them to use spectral-library matching, development of the network, and community-based data sharing [31]. By incorporating chromatographic feature data with MS/MS interactions, feature-based molecular networking (FBMN) expands upon this method even further. This results in nodes that can store information like precursor m/z and retention time and abundance that can help provide more analytical context on metabolites [32].

Molecular families can identify groups of glycated metabolites, phenolics, alkaloids, terpenoids, and flavonoids. This may be because an unknown node adjacent to a known metabolite is a structurally similar or modified metabolite or because it may be part of the same biosynthetic series. This is very useful when direct spectral-library matching doesn't work (Table 2) [33].

Biological data can be used in conjunction with molecular networking to help identify natural products that are active. Chemical families that could be related to a specific phenotype by mapping features from the active fractions/exracts to the molecular network could be identified. Molecular networking improves the potential success of future chemical investigation by considering molecular families rather than aspects of a molecule separately (Fig. 1) [34].

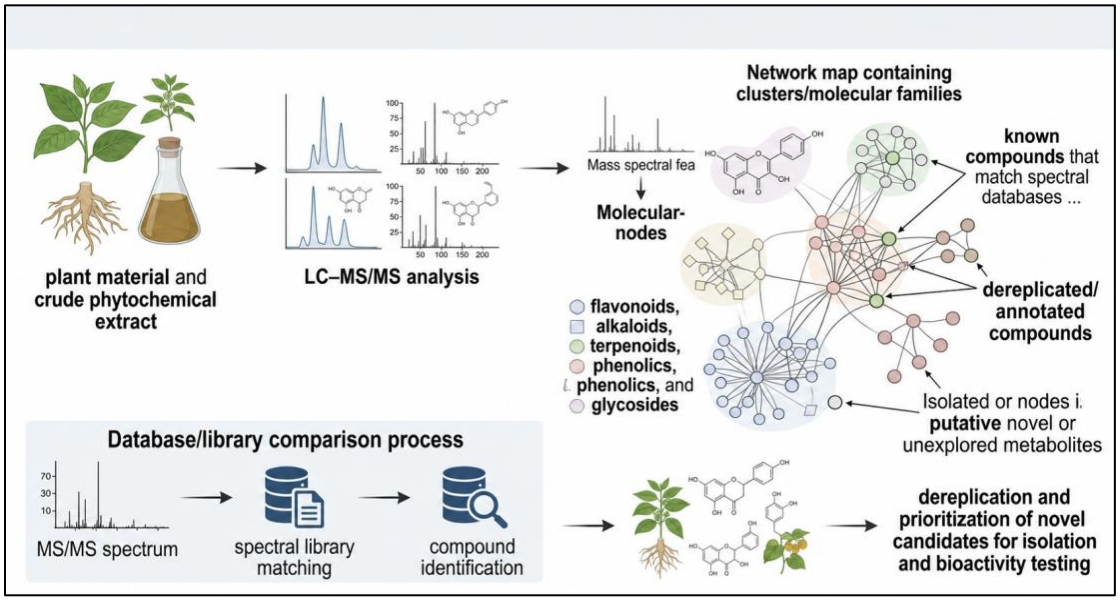

Figure 1: Conceptual representation of molecular networking for phytochemical dereplication

Table 2: Major molecular-networking approaches used in natural-product research

Approach	*Principle / application*	*Main strength*	*Limitation*
Classical MN	MS/MS spectral similarity and molecular-family organization [35]	Visualizes related metabolites	Limited chromatographic context
FBMN	MS/MS similarity combined with LC features [36]	Retains m/z, retention time and abundance	Requires reliable feature processing
IIMN	Connects ion/adduct forms of metabolites [37]	Reduces redundant features	Requires accurate ion-identity assignment
MS2LDA	Identification of recurring fragmentation motifs [38]	Provides substructure-oriented information	Motif interpretation may require expertise
Bioactivity MN	Links molecular features with biological activity [39]	Connects chemistry and phenotype	Requires reliable biological data
Network-assisted dereplication	Combines network context with library matching [40]	Supports interpretation of known compounds and analogues	Does not prove exact structure

5. Machine Learning and Artificial Intelligence in Phytochemical Dereplication

AI and ML can be used to help perform pattern recognition in high-dimensional, biological and structural datasets. In combination with database matching and molecular networking for phytochemical dereplication, they are most useful. Empirical descriptors that are associated with drugs, such as spectral or chemical descriptors, can be used to train supervised algorithms such as support vector machines and random forests [41]. These models can then be used to classify unknown data features into chemical families or metabolite classes. Some unsupervised learning methods, such as dimensionality reduction and clustering, can be useful in searching for patterns in the data without relying upon compound identities. These can be combined with molecular networking to improve the identification of groups of molecules with similar chemical structures [42]. The prediction of chemical structures and formulas with the help of AI is another significant application. For those features that do not have an exact match in the spectral-library, it is possible to generate or even rank candidate structures based on the accurate mass, isotope data, and MS/MS fragmentation [43]. Further, deep learning and representation-learning techniques can be used to learn molecular structure-fragmentation behaviour from a massive data set. It is possible to numerically depict spectra and use molecular graphs or fingerprints to represent molecules. The fact that graph neural networks can capture molecular connectivity the relationship between atoms and chemical bonds makes them very appealing [44].

Another use of AI and ML in candidate prioritisation, a crucial part of dereplication, is the identification of key words. Candidates can be ranked by biological data, molecular descriptors, expected chemical class, abundance, spectral uniqueness and/or network position. A compound with an unknown characteristic, not found in the library, but related to a family of molecular structures that have biological activity might then be prioritized over a common biogenic compound [45]. Bioactivity prediction is an additional layer of analysis, which assesses the likelihood of activity like antibacterial, antioxidant, anti-inflammatory, cytotoxic or enzyme inhibitory activity, but should be used to aid prioritising, not to replace biological assay (Table 3) [46].

Despite these advantages, the power of AI/ML depends heavily on the quality of the training data, and the chemical space that it covers. For some classes of chemicals, natural products databases may be incomplete, but natural metabolites that have been studied extensively are plentiful. This could result in models being biased [47]. Different instruments and acquisition settings may pose another challenge on transferability, while complicated deep-learning systems may be challenging for interpretability. Explainable Artificial Intelligence (XAI) approaches can help narrow down the spectral, chemical or network features that influence predictions to improve expert assessment and reduce overinterpretation [48].

Table 3: AI/ML approaches applicable to phytochemical annotation and prioritization

AI/ML approach	*Typical input*	*Potential use*	*Key limitation*
Random Forest	Spectral/molecular descriptors	Chemical-class prediction	Depends on feature quality [49]
Support Vector Machine	Spectral or molecular features	Metabolite classification	Requires optimized parameters [50]
Neural Networks	MS/MS spectral features	Spectral classification and prediction	Requires suitable training data [51]
Deep Learning	Spectra and molecular representations	Structure/spectral prediction	Data-intensive and less interpretable [52]
Graph Neural Networks	Molecular graphs	Structure and property prediction	Requires diverse datasets [53]
Similarity-learning models	Spectral/chemical representations	Unknown-candidate ranking	Similarity does not prove identity [54]
AI-assisted ranking	Spectral, network and bioactivity data	Novel-compound prioritization	Sensitive to training-data bias [55]

6. Integrated LC–MS/MS, Molecular Networking and AI/ML Workflow

A complete framework for current phytochemical dereplication is provided by the combination of LC-MS/MS, molecular networking, and AI/ML. Metabolites are organised by their spectrum similarity in molecular networking, whereas LC-MS/MS produces precise precursor weights and fragmentation patterns, among other specific chemical information [56]. AI/ML can significantly enhance this

process in various ways, such as metabolite categorization, structural prediction, and candidate prioritization. In doing so, scientists can investigate metabolites known or unknown in detail beyond the capabilities of simple database matches. It also simplifies the identification of chemical families of molecules having similar properties and helps focus biological or chemical research on molecules of greater importance [57].

Typically, the process begins with an extraction of plant materials, followed by LC-HRMS/MS analysis. The choice of extraction conditions, chromatographic separation, ionization mode and MS/MS parameters is determined by the anticipated chemical diversity of the sample. Positive- and negative-ionization modalities can be used to obtain complementary data. Raw data processing involves the subsequent steps of peak identification, peak alignment, isotope grouping, adduct annotation and removal of unwanted features and background signals [58].

The next step in the data processing is to form molecular networks based on the processed MS/MS data. Molecular anchors like known library matches can be utilized, while unannotated nodes can be evaluated using spectral or chromatographic associations. Molecular descriptors and network topology with these properties can be used in conjunction with the AI/ML models to further categorize and rank potential metabolites [59].

Rather than relying solely on the level of signal intensity, priority ranking in candidates could be based on various factors such as the uniqueness of their spectra, abundance, projected chemical class, structural similarity, biological significance, and the location of the candidate in the network. Another approach to association of certain molecular families or nodes with biological activity can be to compare active and inactive fractions [60].

The final step is to identify the most significant metabolites and then experimental validation. Structural confirmation can be obtained by the use of authentic standards, retention-time comparison, MS/MS matching, chromatographic separation, and NMR spectroscopy. The process is iterative as the chemicals identified through the experimental verification can be used to improve the spectrum libraries, annotations, and AI/ML training sets used in subsequent dereplication studies (Fig. 2) [61].

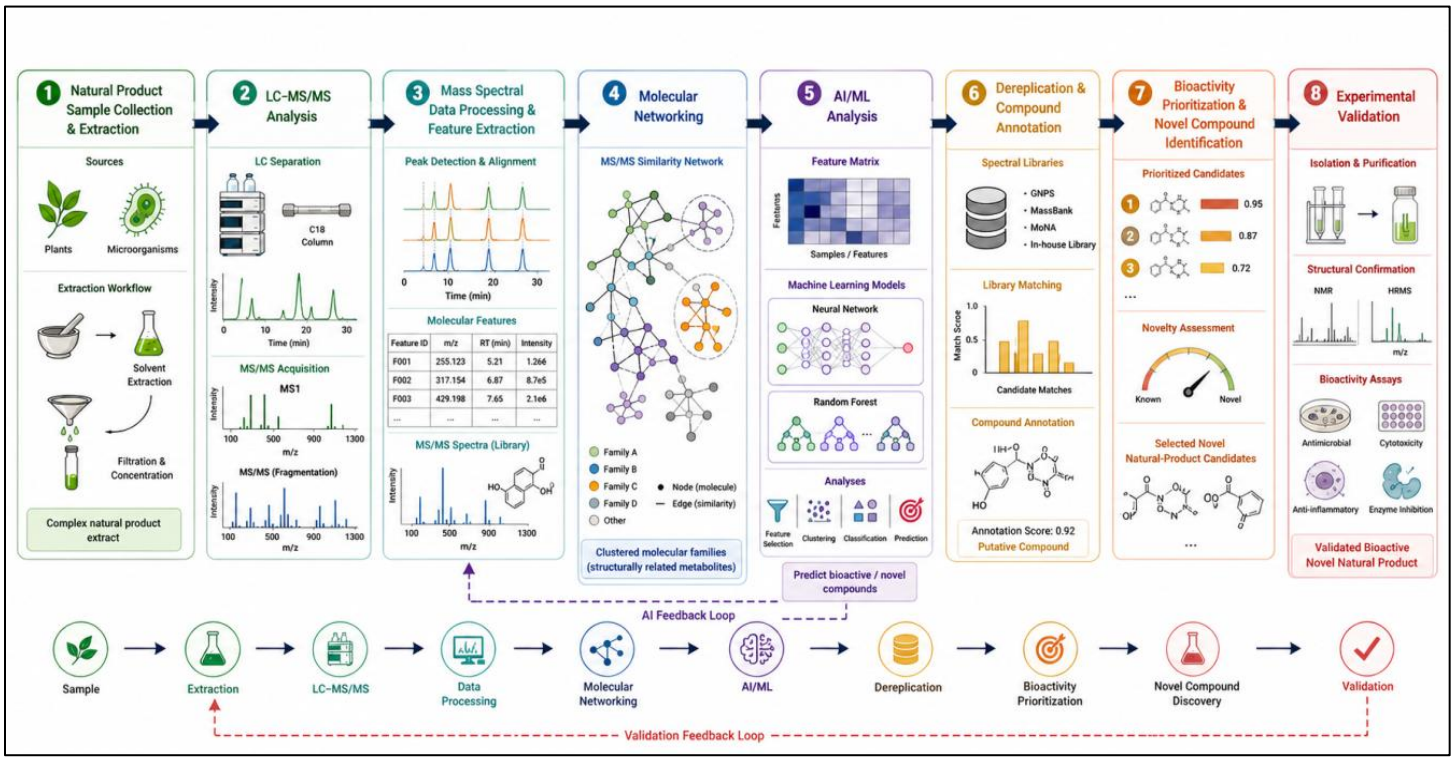

Figure 2: Integrated LC–MS/MS, molecular networking and AI/ML workflow for natural-product discovery

7. Applications in Natural Product Discovery

7.1 Rapid Metabolite Annotation

LC-MS/MS allows for the quick profiling of several phytochemical classes straight from crude extracts or fractions that have only been partly processed. Molecular networking can also be used for additional chemical context by clustering features based on their spectral similarity. These methods can be used together to provide a chemical picture of complex materials prior to the individual metabolite, and thus reduce the need for sequential isolation [62].

7.2 Structural Analogue Discovery

In natural products biosynthetic relationships, small analogues, altered metabolites etc. are often identified by the possibility

of matching them with other known chemicals in the same molecular family, which is detected by molecular networking [63]. Classifying these families and prioritising potential structural linkages may be further assisted by AI/ML methods. A thorough structural assignment is not necessary to prioritise an unannotated characteristic that indicates chemical similarity to a known metabolite; this is especially helpful in cases when direct spectral-library matching fails [64].

7.3 Novel-Compound Prioritization

One of the main objectives of dereplication is to identify chemicals that are of interest for isolation. To narrow down the pool of potential candidates for experimental testing, AI-assisted ranking may take into account spectral uniqueness, network location, chemical-class prediction, database representation, and other factors [65]. Most importantly, a characteristic shouldn't be ignored simply because it is uncommon. A little-known, but abundant, metabolite may be more interesting than a well-known, but low-abundance, metabolite, which may be chemically interesting [66].

7.4 Bioactivity-Guided Discovery

To find metabolites linked to certain phenotypes, chemical profiling may be combined with biological tests. Molecular networking may be used to identify molecular families that are present in large quantities in physiologically active materials, and LC-MS/MS properties can be compared between active and inactive fractions. In addition to selecting targets based on abundance, researchers could predict and prioritize metabolites that may be bioactive using AI/ML and focus on those that are chemically relevant [67].

7.5 Comparative Metabolomics and Chemotaxonomy

Integrated dereplication is a powerful tool for comparing plant populations, genotypes, tissues, developmental stages, habitats and more. Molecular networks can reveal new and universal classes of chemicals, while ML and statistical approaches can focus on the features responsible for the variation among samples [68]. This data could be useful for chemotaxonomic studies, finding unique chemical markers, and choosing plants for future natural product research [69].

7.6 Improving Isolation Efficiency

Dereplication is most helpful when used downstream. The selective purification can be used to remove chemically different or physiologically relevant properties, while favored metabolites can be de-emphasized. So experiments can be performed with less effort, fewer steps in the purification process, and using limited resources [70].

8. Challenges, Validation and Limitations

Although technology is rapidly developing, artificial intelligence (AI) phytochemical dereplication has a few drawbacks [71].

8.1 Incomplete databases

The area of natural products is chemically diverse and large, while the area of all possible structures in the public spectrum databases is limited. There is a dearth of experimentally confirmed MS/MS spectra for several specialised plant metabolites. Hence, a chemical with an unusual spectrum shouldn't be taken for granted as being completely new [72]. Enhancing dereplication and AI-model accuracy and performance relies on increasing the number of empirically verified spectra, especially those of natural-product classes that are under represented [73].

8.2 Isomerism and annotation confidence

While trying to differentiate between positional isomers, stereoisomers, and closely related analogues, precise mass and MS/MS data may not be enough. Molecular networking can demonstrate spectral similarities but not stereochemistry and/or exact connectivity [74]. So, until more proof is gathered, structures formed by AI should be referred to as forecasts or hypotheses. Confidence in the annotation needs to be explicitly expressed. There is a significant difference in the levels of evidence that involve feature detection, presumptive annotation, likely identification, and conclusive identification [75].

8.3 Data quality and model bias

AI/ML models are greatly influenced by the diversity and quality of the training data sets. Common metabolites may be overrepresented in the biased models, which do not represent rare natural products. The transferability of the model could be further affected by the type of instrument, collision energy, ionisation channels, chromatographic conditions, and pre processing [76]. Thus, benchmarking and independent validation data sets are needed. It would be nice to evaluate models over multiple labs, devices and chemical classes, rather than using only the datasets that were randomly split from one collection [77].

8.4 Reproducibility

The final molecular network may vary based on various selection parameters, such as filtering thresholds, peak-picking criteria, and options for network construction. Similarly, AI forecasts may be impacted by changes in model design and training data. Transparency of analytical and computational parameters and deposition of raw and processed MS/MS data can help to improve repeatability and aid independent verification [78].

8.5 Experimental validation

Computational dereplication is not to replace experimental natural-product chemistry, but is meant to complement it. Depending on the required level of certainty, orthogonal chromatography, high-resolution MS/MS, NMR spectroscopy, retention-time comparison, genuine standards and other

analytical techniques can be employed. Experimental structural elucidation is still crucial for new chemical entities. Finally, adequate analytical information will be used to determine identification, but calculations will help to rule out some candidates [79].

9. Future Perspectives

Improved integration of analytical chemistry, computational science and experimental natural product research is essential for future progress in the field of phytochemical dereplication. Bigger and more chemically diverse spectrum databases will improve library matching and generalisation of AI models [80]. These will be more accurately represented in these databases as they will contain more empirically confirmed spectra and unusual natural-product scaffolds. These improvements are essential to enhance the reliability and applicability of computational dereplication methods to be used in a broader spectrum of applications [81].

The future of AI is also promising, with multimodal AI being an exciting area of research. The chemistry of natural products could be better represented, with the integration of chemical structures, retention information, nuclear magnetic resonance (NMR), genomic or transcriptomic data and biological activity into future systems [82]. Models to accurately detect the fragments of the spectrum, chemical aspects, or network interactions that explain a prediction will also become more prominent in the development of explainable AI. Transparency of AI-assisted annotation would be improved, and scholars would be provided with the possibility to critically evaluate computational ideas. In the future, flows including multiple analytical steps like molecular networking, experimental prioritisation, database annotation, AI-assisted candidate rating, and raw LC-MS/MS processing, could also be automated [83].

Therefore, replacing experimental chemistry with AI isn't likely to be the most effective way to go in the future. Experimental chemistry should be able to prove it, but AI can be a smart decision support tool that can find patterns, propose structural hypotheses, and prioritize candidates [84].

CONCLUSION

AI-based phytochemical dereplication is an encouraging approach for improving the efficiency of the discovery of natural products. Molecular networking clusters metabolites into groups, LC MS/MS delivers accurate mass and fragmentation information, and AI/ML facilitates chemical classification, hypothesis generation and candidate prioritization. They can be combined to allow for less time spent isolating the same chemicals and more time spent studying metabolites that are chemically unique or are biologically significant. AI/ML can combine all the previous with chemical, structural, network, and biological information to create a more systematic and effective prioritisation of candidate metabolites, while molecular networking offers important chemical context for unannotated characteristics. It should not be confused, however, with an actual structural identification. Incomplete databases, structural isomerism, instrumental variability, bias in databases, and limited interpretation of models remain an issue. Appropriate confidence levels, computer-aided methods of dereplication with clear documentation, and experimental validation are essential for reliable dereplication. These developments should be continued to increase the speed, reliability and reproducibility of natural-product discovery, and are expected to include further advances in the development of spectral libraries and databases that have been validated, explainable AI, multimodal data integration and automated workflows. Moreover, a powerful tool for next generation phytochemical dereplication is possible by using the combination of LC-MS/MS, molecular networking and Artificial Intelligence/ML.

ACKNOWLEDGEMENT: Nil

CONFLICT OF INTEREST: Nil

REFERENCES:

1. Ali S, Khalil AA, Akhtar MS, Amin A, Zaman W. Comprehensive insights into natural bioactive compounds: From chemical diversity and mechanisms to biotechnological innovations and applications. ChemistryOpen. 2026 Apr;15(4):e202500469.
2. Radulovic NS, Blagojevic PD, Stojanovic-Radic ZZ, Stojanovic NM. Antimicrobial plant metabolites: structural diversity and mechanism of action. Current medicinal chemistry. 2013 Mar 1;20(7):932-52.
3. Pérez-Victoria I. Natural products dereplication: databases and analytical methods. Progress in the Chemistry of Organic Natural Products 124. 2024 Aug 6:1-56.
4. Gangwal A, Lavecchia A. Artificial intelligence in natural product drug discovery: current applications and future perspectives. Journal of medicinal chemistry. 2025 Feb 7;68(4):3948.
5. Lavecchia A. Artificial Intelligence-Driven Natural Product Drug Discovery: From Computational Genome Mining to Clinical Translation. Medicinal Research Reviews. 2026 Aug 7.
6. Feng S, Huang M, Li Y, Cai A, Yue X, Wang S, Chen L, Jiang J, Luo Y. Intelligent understanding of spectra: from structural elucidation to property design. Chemical Society Reviews. 2025 Sep 15;54(18):8243-86.
7. Gaudêncio SP, Bayram E, Lukić Bilela L, Cueto M, Díaz-Marrero AR, Haznedaroglu BZ, Jimenez C, Mandalakis M, Pereira F, Reyes F, Tasdemir D. Advanced methods for natural products discovery: bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. Marine drugs. 2023 May 19;21(5):308.
8. Hubert J, Nuzillard JM, Renault JH. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept?. Phytochemistry Reviews. 2017 Feb;16(1):55-95.
9. Qiao X, Lin XH, Ji S, Zhang ZX, Bo T, Guo DA, Ye M. Global profiling and novel structure discovery using multiple neutral loss/precursor ion scanning combined with substructure recognition and statistical analysis (MNPSS): characterization of terpene-conjugated curcuminoids in

Curcuma longa as a case study. Analytical chemistry. 2016 Jan 5;88(1):703-10.
10. Wolfender JL, Nuzillard JM, Van Der Hooft JJ, Renault JH, Bertrand S. Accelerating metabolite identification in natural product research: toward an ideal combination of liquid chromatography–high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. Analytical Chemistry. 2018 Nov 19;91(1):704-42.
11. Caesar LK, Cech NB. Synergy and antagonism in natural product extracts: when 1+ 1 does not equal 2. Natural product reports. 2019 Jun 1;36(6):869-88.
12. David PA, Norazhar AI, Che Zain MS. Integrating LC-MS/MS and Molecular Networking for Advance Analysis and Comprehensive Flavonoids Annotation. Journal of Separation Science. 2025 Jul;48(7):e70230.
13. Salem MA, Perez de Souza L, Serag A, Fernie AR, Farag MA, Ezzat SM, Alseekh S. Metabolomics in the context of plant natural products research: From sample preparation to metabolite analysis. Metabolites. 2020 Jan 15;10(1):37.
14. Mutha RE, Kalaskar M, Khan ZG. Modern analytical techniques for quality control and chemical identification of phytochemicals. Pharmacognosy and phytochemistry: principles, techniques, and clinical applications. 2025 Mar 13:167-88.
15. Misra BB, Assmann SM, Chen S. Plant single-cell and single-cell-type metabolomics. Trends in plant science. 2014 Oct 1;19(10):637-46.
16. Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, Petras D, Gauglitz JM, Sikora N, Vargas F, van Der Hooft JJ. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nature protocols. 2020 Jun 1;15(6):1954-91.
17. De Hoffmann E. Tandem mass spectrometry: a primer. Journal of mass spectrometry. 1996 Feb;31(2):129-37.
18. Dong C, Wang H, Xu K, Gu G, Zhang P. LC-MS/MS-guided discovery of herbicidal himeic acid derivatives from the Nicotiana tabacum-derived fungus Aspergillus japonicus 334. Journal of Agricultural and Food Chemistry. 2025 Aug 18;73(34):21463-72.
19. Defossez E, Bourquin J, Von Reuss S, Rasmann S, Glauser G. Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. Mass Spectrometry Reviews. 2023 Jan;42(1):131-43.
20. Bilbao A, Varesio E, Luban J, Strambio-De-Castillia C, Hopfgartner G, Müller M, Lisacek F. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. Proteomics. 2015 Mar;15(5-6):964-80.
21. Tsugawa H, Nakabayashi R, Mori T, Yamada Y, Takahashi M, Rai A, Sugiyama R, Yamamoto H, Nakaya T, Yamazaki M, Kooke R. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. Nature methods. 2019 Apr;16(4):295-8.
22. Evans AM, Bridgewater BR, Liu Q, Mitchell MW, Robinson RJ, Dai H, Stewart SJ, DeHaven CD, Miller LA. High resolution mass spectrometry improves data quantity and quality as compared to unit mass resolution mass spectrometry in high-throughput profiling metabolomics. Metabolomics. 2014 Apr 1;4(2):1.
23. Ernst M, Kang KB, Caraballo-Rodríguez AM, Nothias LF, Wandy J, Chen C, Wang M, Rogers S, Medema MH, Dorrestein PC, Van Der Hooft JJ. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. Metabolites. 2019 Jul 16;9(7):144.
24. Ochoa JL, Crittenden CM. In-Source Fragmentation of Pyrrolidine-Containing KRAS Scaffolds Leads to Enhanced Structure Elucidation and Tandem Mass Spectral Quality. Journal of the American Society for Mass Spectrometry. 2026 Jan 8;37(2):363-6.
25. Cho K, Schwaiger-Haber M, Naser FJ, Stancliffe E, Sindelar M, Patti GJ. Targeting unique biological signals on the fly to improve MS/MS coverage and identification efficiency in metabolomics. Analytica chimica acta. 2021 Mar 8;1149:338210.
26. Samak T, Gunter D, Goode M, Deelman E, Juve G, Mehta G, Silva F, Vahi K. Online fault and anomaly detection for large-scale scientific workflows. In2011 IEEE International Conference on High Performance Computing and Communications 2011 Sep 2 (pp. 373-381). IEEE.
27. Quinn RA, Nothias LF, Vining O, Meehan M, Esquenazi E, Dorrestein PC. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. Trends in pharmacological sciences. 2017 Feb 1;38(2):143-54.
28. Gupta P, Gunning M, Ahmed S, Tagkopoulos I. From metabolomics dark matter to bioactivity: a computational framework for prioritizing food-derived bioactive compounds. npj Science of Food. 2026 Aug 25.
29. Ramabulana AT, Petras D, Madala NE, Tugizimana F. Metabolomics and molecular networking to characterize the chemical space of four Momordica plant species. Metabolites. 2021 Nov 8;11(11):763.
30. Ekins S, Bugrim A, Brovold L, Kirillov E, Nikolsky Y, Rakhmatulin E, Sorokina S, Ryabov A, Serebryiskaya T, Melnikov A, Metz J. Algorithms for network analysis in systems-ADME/Tox using the MetaCore and MetaDrug platforms. Xenobiotica. 2006 Jan 1;36(10-11):877-901.
31. Muthuraj R, Chandrasekaran J. Nature meets machine: the AI renaissance in natural product drug discovery. Natural Products and Bioprospecting. 2026 Dec;16(1):37.
32. Qu B, Liu Y, Shen A, Guo Z, Yu L, Liu D, Huang F, Peng T, Liang X. Combining multidimensional chromatography-mass spectrometry and feature-based molecular networking methods for the systematic characterization of compounds in the supercritical fluid extract of Tripterygium wilfordii Hook F. Analyst. 2022 Dec 20;148(1):61-73.
33. Campos-Vega R, Oomah BD. Chemistry and classification of phytochemicals. Handbook of plant food phytochemicals. 2013 Jan 2:5-48.
34. Fox Ramos AE, Evanno L, Poupon E, Champy P, Beniddir MA. Natural products targeting strategies involving molecular networking: different manners, one goal. Natural product reports. 2019 Jul 1;36(7):960-80.
35. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J. MS/MS networking guided analysis of molecule and gene cluster families. Proceedings of the National Academy of Sciences. 2013 Jul 9;110(28):E2611-20.
36. Pang Z, Xu L, Viau C, Lu Y, Salavati R, Basu N, Xia J. MetaboAnalystR 4.0: a unified LC-MS workflow for global metabolomics. Nature communications. 2024 May 1;15(1):3675.
37. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: a computational solution for streamlining metabolomics analysis. Analytical chemistry. 2017 Nov 3;90(1):480.
38. Yang M, Wu Y, Zhang Z, Xu Y, Lei T, Wang X, Jin Z, Hou K, Cai Y, Sun S. StrucGAP: a modular, streamlined and traceable data mining platform for structural and site-

specific glycoproteomics. Nature Communications. 2026 Mar 19;17(1):2579.
39. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. Nature reviews Molecular cell biology. 2019 Jun;20(6):353-67.
40. Lo YC, Senese S, Damoiseaux R, Torres JZ. 3D chemical similarity networks for structure-based target prediction and scaffold hopping. ACS Chemical Biology. 2016 Jun 20;11(8):2244.
41. Tariq U, Ahmed I, Khan MA, Bashir AK. Bridging biosciences and deep learning for revolutionary discoveries: A comprehensive review. IAES International Journal of Artificial Intelligence (IJ-AI). 2025 Apr;14(2):867-83.
42. Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, Petras D, Gerwick WH, Rousu J, Dorrestein PC, Böcker S. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nature biotechnology. 2021 Apr;39(4):462-71.
43. Alghadeer M, Aisyah ND, Hezam M, Alqahtani SM, Baloch AA, Alharbi FH. Machine learning prediction of materials properties from chemical composition: Status and prospects. Chemical Physics Reviews. 2024 Dec 1;5(4).
44. Liu H, Zhu B, Nie S, Li H, Lin Y, Ma T, Shao X, Chen Q, Shen M, Zheng Y, Fan X. Advancing ADMET prediction through multiscale fragment-aware pretraining with MSformer-ADMET. Briefings in Bioinformatics. 2025 Sep;26(5):bbaf506.
45. Hussein MA, Munirathinam G. Artificial intelligence-driven natural product discovery for cancer metastasis and chemoresistance: from computational prediction to preclinical validation. Cancers. 2026 Feb 24;18(5):719.
46. Lagunin A, Filimonov D, Poroikov V. Multi-targeted natural products evaluation based on biological activity prediction with PASS. Current Pharmaceutical Design. 2010 May 1;16(15):1703-17.
47. Elkhatat A, Al-Muhtaseb SA. Harnessing Artificial Intelligence (AI) for a greener future: a review of AI advancements in green chemistry, chemical processes and sustainable materials. Applied Intelligence. 2026 Jun;56(8):306.
48. Tran AT, Zeevi T, Payabvash S. Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging. BioMedInformatics. 2025 Apr 14;5(2):20.
49. Goldman S, Wohlwend J, Stražar M, Haroush G, Xavier RJ, Coley CW. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. Nature Machine Intelligence. 2023 Sep;5(9):965-79.
50. Raghav M, Dubey A, Singh J. Hyperspectral imaging for detection and classification of plant primary and secondary metabolites: a review. Phytochemical Analysis. 2026 Jan;37(1):4-27.
51. Djoumbou-Feunang Y, Pon A, Karu N, Zheng J, Li C, Arndt D, Gautam M, Allen F, Wishart DS. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. Metabolites. 2019 Apr 13;9(4):72.
52. Sorrentino S, Gussoni A, Calcagno F, Pasotti G, Avagliano D, Rivalta I, Garavelli M, Polli D. Mol2Raman: a graph neural network model for predicting Raman spectra from SMILES representations. Digital Discovery. 2026 Jan 1;5(1):161-76.
53. Zang X, Zhao X, Tang B. Hierarchical molecular graph self-supervised learning for property prediction. Communications Chemistry. 2023 Feb 17;6(1):34.
54. Ma G Kurita KL, Glassey E, Linington RG. Integration of high-content screening and untargeted metabolomics for comprehensive functional annotation of natural product libraries. Proceedings of the National Academy of Sciences. 2015 Sep 29;112(39):11999-2004.
55. Sena S, Joseph GG, Kumar V. Artificial Intelligence revolutionizing plant biotechnology research: current breakthroughs and future outlooks. Plant Cell, Tissue and Organ Culture (PCTOC). 2026 Sep;166(3):66.
56. Chigozie VU, Ugochukwu CG, Igboji KO, Okoye FB. Application of artificial intelligence in bioprospecting for natural products for biopharmaceutical purposes. BMC Artificial Intelligence. 2025 Jul 15;1(1):4.
57. Ponticelli M, Esposito G, Labanca F, Scognamiglio P, Sinisgalli C, Milella L, Faraone I, Costantino V. Aglianico grape pomace as a source of antioxidant and anti-proliferative biomolecules: Eco-friendly extraction and HRMS/MS-based molecular networking. Food Chemistry. 2025 Mar 30;469:142573.
58. Olivon F, Grelier G, Roussi F, Litaudon M, Touboul D. MZmine 2 data-preprocessing to enhance molecular networking reliability. Analytical chemistry. 2017 Aug 1;89(15):7836-40.
59. Qeli E, Omasits U, Goetze S, Stekhoven DJ, Frey JE, Basler K, Wollscheid B, Brunner E, Ahrens CH. Improved prediction of peptide detectability for targeted proteomics using a rank-based algorithm and organism-specific data. Journal of proteomics. 2014 Aug 28;108:269-83.
60. Trunzer M, Faller B, Zimmerlin A. Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. Journal of medicinal chemistry. 2009 Jan 22;52(2):329-35.
61. Raju KS, Kadian N, Taneja I, Wahajuddin MJ. Phytochemical analysis of isoflavonoids using liquid chromatography coupled with tandem mass spectrometry. Phytochemistry Reviews. 2015 Jun;14(3):469-98.
62. Shugrue CR, Miller SJ. Applications of nonenzymatic catalysts to the alteration of natural products. Chemical reviews. 2017 Sep 27;117(18):11894-951.
63. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ. The Pfam protein families database: embracing AI/ML. Nucleic acids research. 2025 Jan 6;53(D1):D523-34.
64. Lang G, Mayhudin NA, Mitova MI, Sun L, van der Sar S, Blunt JW, Cole AL, Ellis G, Laatsch H, Munro MH. Evolving trends in the dereplication of natural product extracts: new methodology for rapid, small-scale investigation of natural product extracts. Journal of Natural Products. 2008 Sep 26;71(9):1595-9.
65. Ritchhart R. Intellectual character: What it is, why it matters, and how to get it. John Wiley & Sons; 2004 25.
66. Andrew Clayton T, Lindon JC, Cloarec O, Antti H, Charuel C, Hanton G, Provost JP, Le Net JL, Baker D, Walley RJ, Everett JR. Pharmaco-metabonomic phenotyping and personalized drug treatment. Nature. 2006 Apr 20;440(7087):1073-7.
67. James ME, Brodribb T, Wright IJ, Rieseberg LH, Ortiz-Barrientos D. Replicated evolution in plants. Annual review of plant biology. 2023 May 22;74(1):697-725.

68. Domingo-Fernández D, Gadiya Y, Mubeen S, Healey D, Norman BH, Colluru V. Exploring the known chemical space of the plant kingdom: insights into taxonomic patterns, knowledge gaps, and bioactive regions. Journal of Cheminformatics. 2023 Nov 10;15(1):107.
69. Strejcek M, Smrhova T, Junkova P, Uhlik O. Whole-cell MALDI-TOF MS versus 16S rRNA gene analysis for identification and dereplication of recurrent bacterial isolates. Frontiers in microbiology. 2018 Jun 19;9:1294.
70. Paerhati Y, Aikebaier A, Dilimulati D, Baishan A, Yusufujiang N, Qiu X, Wusiman Y, Zhou W. Rethinking nature's pharmacy: AI era and natural product drug discovery. Pharmaceuticals. 2026 Feb 11;19(2):301.
71. Van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, Neto FC, Castano-Espriu L, Chang C, Clark TN, Cleary Little JL. The natural products atlas: an open access knowledge base for microbial natural products discovery. ACS central science. 2019 Nov 27;5(11):1824-33.
72. Mirza A, Patiny L, Jablonka KM. End-to-end multimodal structure elucidation from raw spectra combining contrastive learning and evolutionary algorithms. Nature Communications. 2026 Jun 5;17(1):5013.
73. Furuhashi T, Weckwerth W. Isomer analysis by mass spectrometry in clinical science. TrAC Trends in Analytical Chemistry. 2023 Feb 1;159:116907.
74. Josephson JR, Chandrasekaran B, Smith JW, Tanner MC. A mechanism for forming composite explanatory hypotheses. IEEE Transactions on Systems, Man, and Cybernetics. 1987 May 31;17(3):445-54.
75. Kumar A, Dhanka S, Sharma A, Sharma A, Nain M, Kumar P, Gupta AR, Bansal J, Saxena NK, Pant R. A comprehensive review of bias in AI, ML, and DL models: methods, impacts, and future directions. Archives of Computational Methods in Engineering. 2026 May;33(4):5743-73.
76. Phoon KK, Shuku T, Ching J, Yoshida I. Benchmark examples for data-driven site characterisation. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards. 2022 Oct 2;16(4):599-621.
77. Egan JM, van Santen JA, Liu DY, Linington RG. Development of an NMR-based platform for the direct structural annotation of complex natural products mixtures. Journal of natural products. 2021 Mar 22;84(4):1044.
78. Deng C, Dong T, Zhang M, Zi Y, Yang W, Hao F, Zhu L, Zhang S, Li C, Ding Z, Cai L. Computer-Assisted Analytical Workflows for Natural Product Dereplication, Structure Elucidation, and Bioactivity-Oriented Prioritization. Analytical Chemistry. 2026 Jul 13.
79. Sarker SD, Nahar L, editors. Computational phytochemistry. Elsevier; 2018 May 2.
80. Avellaneda-Tamayo JF, Agudo-Muñoz NA, Sánchez-Galán JE, López-Pérez JL, Medina-Franco JL. Chemoinformatic characterization of NAPROC-13: A database for natural product 13C NMR dereplication. Journal of Natural Products. 2024 Sep 13;87(9):2216.
81. Sun YZ, Sun HL, Ma JC, Zhang P, Huang XY. Multimodal agent ai: A survey of recent advances and future directions. Journal of Computer Science and Technology. 2025 Jul;40(4):1046-63.
82. Gallegos M, Vassilev-Galindo V, Poltavsky I, Martín Pendás Á, Tkatchenko A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. Nature Communications. 2024 May 21;15(1):4345.
83. Sadeghi S, Canty RB, Mukhin N, Xu J, Delgado-Licona F, Abolhasani M. Engineering a sustainable future: harnessing automation, robotics, and artificial intelligence with self-driving laboratories. ACS Sustainable Chemistry & Engineering. 2024 Aug 6;12(34):12695-707.

How to cite this article:
Saxena K, Choudhary H, Vyas A. AI-Assisted Phytochemical Dereplication: Integration of LC–MS/MS, Molecular Networking and Machine Learning for Natural Product Discovery. International Journal of Pharmaceutical Science and Medicine 2026; 4(1): 13-22.