Role of Artificial Intelligence in Lc-Ms/Ms Metabolomics

Saleem Javid[1*], Ayshathul Shahala S[2], Ashwini Hanmant Vatre[2], Fathimai Thaisena[2], Ayisha Rifa P[2], Mohammed Gulzar Ahmed[3].

[1]*Department of Pharmaceutical Chemistry, Yenepoya Pharmacy College & Research Centre, Yenepoya (Deemed to be University), Mangalore 570 018, Karnataka, India.*

[2]*Department of Pharmacology, Yenepoya Pharmacy College & Research Centre, Yenepoya (Deemed to be University), Mangalore 570 018, Karnataka, India.*

[3]*Department of Pharmaceutics, Yenepoya Pharmacy College & Research Centre, Yenepoya (Deemed to be University), Mangalore 570 018, Karnataka, India.*

Corresponding author Email: saleemjavid@yenepoya.edu.in
Doi: 10.5281/zenodo.18835044
Received: 12 February 2026
Accepted: 19February 2026

ABSTRACT

Artificial intelligence is transforming LC-MS/MS metabolomics by automating data handling and improving metabolite identification, leading to higher sensitivity, reproducibility, and clinical relevance. Machine learning and deep learning enhance baseline correction, noise reduction, peak detection, deconvolution, and retention-time alignment, reducing manual curation. Tools such as PeakBot, MS2DeepScore, MIST-CF, SingleFrag, and CFM-ID strengthen spectral matching, fragmentation prediction, and de novo annotation, expanding detectable metabolites. AI-driven feature selection and classification support robust biomarker discovery, disease diagnosis, and multi-omics integration for systems-level pathway insights and precision medicine. Remaining challenges include limited annotated data, inter-laboratory variability, model opacity, computational cost, and lack of standardized benchmarks. Future priorities are unified, explainable end-to-end AI workflows, shared LC-MS/MS repositories with community benchmarks, and accessible cloud platforms to enable routine AI-driven metabolomics in research and clinical practice.

Keywords: Artificial intelligence, LC-MS/MS, metabolomics, biomarker discovery, multi-omics integration

1. INTRODUCTION

Metabolomics aims at measuring and comprehending every small molecule (metabolites) that exist in biological systems such as cells, tissues or biofluids. Liquid chromatography combined with tandem mass spectrometry (LC–MS/MS) has become thebest analysis platform because of its outstanding sensitivity, selectivity and versatility across various metabolite classes [1,2]. LC MS/MS is a combination of the separating capacity of

chromatography with the molecular specificity of mass spectrometru. This lets you find and scale hundreds to thousands of metabolites at a time from complex biological samples. The general LC-MS/MS metabolomics workflow includes sample extraction followed by chromatographic separation, ionization, precursor and fragment ions detection through downstream data analysis and interpretation [3]. The mass to charge ratio (m/z), signal intensity, and retention time of each of the detected features are altogether included in the multidimensional datasets that are generated. The fragmentation spectra (MS/MS) consist of structural information that are useful in the identification of metabolite are provided by the tandem mass spectrometry. Despite such advantages, there are huge challenges for data analysis and biological interpretation for LC-MS/MS metabolomics [4]. The initial challenge is caused by the volume and complexity of LC-MS/MS data. The ability of large cohort studies to compute hundreds of gigabytes of data is enabled by the ability of the contemporary fine-resolution tools to trace of millions of data points at once. The non-trivial process of extracting trustworthy peak features, aligning them across samples, and then connecting them to metabolite identities need to be carried out by strong computational pipelines. Noise, baseline drift, matrix effects and co-eluting compounds make the procedure even more difficult [5]. The conventional approaches rely on human curation and rule-based algorithms, that are tedious and often random. These traditional workflow have trouble scaling which is useful when the datasets are large, which may result in large false discovery rates and poor laboratory reproducibility [6]. So in order to enhance automation, accuracy and interpretability throughout the LC-MS/MS data pipeline, the computational metabolomics has growingly sought artificial intelligence (AI) and its subfields, machine learning (ML) and deep learning (DL). AI is used to refer to computer programs that can perform such operations as pattern recognition, learning, and decision-making which normally require human intelligence. Although DL (as an artificial neural networks variant) will automatically discover complex features in high-dimensional data spaces, machine learning (ML), which is a sub-topic of artificial intelligence, applies algorithms that learn relationships and rules directly from the data [8]. Through direct learning of raw LC-MS/MS signals and spectra, these methods are transforming metabolomics by making it easy by means of automated features extraction, peak detection, spectral matching, and identification of metabolite with minimal human support [9]. The importance of AI use in LC–MS/MS workflows has many important advantages. To begin with, both ML and DL models are less sensitive to batch effects and variations in instruments since they are able to adapt and adjust to heterogeneous and noisy data. Second, they are able to reveal patterns which traditional statistical procedures may be ignorant of, and thus lead to the finding of greater sensitivity of biomarkers. Finally AI accelerates clinical, environmental research and nutritional metabolomics through data-driven hypothesis generation and biological interpretation [10,11].

2. OVERVIEW OF LC-MS/MS DATA AND ANALYTICAL CHALLENGES

2.1. Complexity and volume of LC-MS/MS data

Each LC -MS /MS analysis produces a large, multifaceted dataset that represents data on ion intensity, retention time and m/z. Whereas the mass spectrometric dimension provides molecular specificity, and the chromatographic dimension provides temporal separation. High resolution systems have the capacity to identify over 10 4 -10 5 features per sample, many of which are not distinct metabolites instead it is isotopes, adducts, in-source fragments or noise [12].

Extensive studies are done on hundreds or thousands of samples in different conditions of metabolomics investigations. Consequently, the datasets are found to be highly dynamic, varied and complex. The instrument variability, chromatographic retention time drift and ionization efficiency variation introduces systematic errors, thus the volume of data is not the only problem [13]. It is possible that these variables can confound the biological variation and give rise to misinterpretations in the case that they are not adjusted accordingly. With the help of AI, this massive data is being visualized, compressed and organized now with techniques, such as unsupervised learning algorithms, such as clustering, and principal component analysis (PCA) and autoencoders. Deep learning techniques can even work directly on unprocessed LC-MS maps as multidimensional "images," automatically removing noise and extracting significant signal patterns.

2.2 Noise, peak overlap, and matrix effects

Solvents, mobile phases, and sample contaminants can produce noise and background signals in metabolomics. Accurate quantification may be hampered by overlapping peaks caused by co-eluting analytes. Furthermore, signals can be non-uniformly suppressed or enhanced by matrix effects, which are variations in ionization efficiency brought on by sample composition [15]. Traditional peak detection algorithms frequently use predefined shapes or simple thresholds (such as Gaussian models), which can be ineffective when there is fluctuating noise or an irregular peak shape. On the other hand, using training data, machine learning-based peak detection techniques can identify the distinctive forms and temporal profiles of "real" chromatographic peaks. More precisely than rule-based methods, neural network architectures, such as convolutional neural networks (CNNs), have been used to separate overlapping signals and locate peaks in noisy chromatograms [16]. By identifying ions that belong to the same compound based on their correlation structures and co-elution patterns, AI can also help with deconvolution. AI lowers downstream errors in identification and quantification, which are frequently exacerbated in conventional workflows, by increasing the precision of peak picking and deconvolution. Studies have shown that deep learning– based algorithms can reduce false peak detection rates by up to 50% compared with conventional methods such as XCMS or MZmine [17].

2.3 Difficulties in metabolite identification and quantification

Metabolite identification, or giving each detected feature a chemical structure, is one of the main problems in LC-MS/MS metabolomics. Usually, identification entails comparing MS/MS fragmentation spectra to reference libraries like Mass Bank, GNPS, or METLIN. Annotation is challenging or impossible because many physiologically significant metabolites lack reference spectra [18. Furthermore, it is more difficult to identify by comparison because LC retention times are not consistent across instruments or laboratories. By combining experimental and theoretical data, AI-based models have been created to forecast retention durations and fragmentation patterns from molecular structures, enhancing annotation [19]. For instance, a structured machine learning model called LC–MS2Struct greatly outperforms spectral similarity alone in the ranking of candidate structures by combining retention order and MS/MS spectral similarity [20]. Quantification tends to have other complications. Even after the successful identification, ion suppression, calibration drift, and sample preparation variability all have an impact on accurate quantification. It is possible to do data-driven normalization and by modeling such nonlinear sources of variation, the machine learning algorithms are capable of eliminating as much technical noise as possible with preserving biological signals [21].

2.4. Need for improved data processing and interpretation methods

Biological interpretation is the final goal of metabolomics and it involves relating the changes in metabolite profiles with physiological or pathological states which are not limited to identification and quantification. Nonetheless, information interpretation instead of information generation is the bottleneck at the moment. The so-called metabolomic dark matter, in which the most features are unknown, has the outcomes of the huge number of features present that we can annotate [22].

With the combination of contextual and multi-omic data, AI tools can help address this gap. ML models can be used to correlate unidentified spectral features, e.g. transcriptomic phenotypes, phenotypes of disease or biological pathways that are known. In clinical metabolomics, the disease predictive metabolite signatures have been determined by using supervised learning such as random forests, gradient boosting and neural networks, with a provision towards the production of diagnostic models and biomarker discovery [23, 24].

Nevertheless, harsh validation and interpretability should be used as a benchmark to incorporate AI. When they are trained on small or biased datasets, they are often susceptible to overfitting, data residual, and ineffective intercohort generalization. External testing, cross-validation and explainable AI (XAI) techniques, such as saliency maps and SHAP values are being frequently used in order to discuss this and ensure useful predictions [25].

3. ARTIFICIAL INTELLIGENCE TECHNIQUES RELEVANT TO LC-MS/MS METABOLOMICS

LC-MS/MS is a fundamental tool used in Metabolomics, which provides high sensitivity and wide spectrum detection of small molecule metabolites [1][3]. Nevertheless, it still has some issues with raw data, metabolite identification, and result interpretation, which limit its overall utility [26, 27]. To address these challenges, machine learning (ML), deep learning (DL), clustering strategies, and natural language processing (NLP) artificial intelligence (AI) solutions have been adopted more and more in metabolomics workflows [6][28][29]. In this paper, AI strategies that can facilitate progress in areas such as automated peak detection, metabolite classification, biomarker identification, and literature-based data mining are reviewed [16][30][31].

3.1 Machine learning

Machine learning (ML) has become an essential tool in Metabolomics, which assists a researcher in analysing complex LC-MS/MS metabolomics data [32]. There are two leading approaches under machine learning, which are supervised learning and unsupervised learning. Supervised learning methods, including random forests, support vector machines, and gradient boosting, are used in most of the studies to forecast the biological outcome as well as to classify the metabolites based on labelled datasets [32][33]. These methods can be extensively applied to biomarker discovery, in which the researchers can identify the difference in the metabolites between experimental groups [34]. In contrast, unsupervised techniques of learning include principal component analysis (PCA), hierarchical clustering, and k-means clustering, can be used to explore unlabelled datasets [35][36]. The tools may be employed to identify clusters of metabolites, accentuate outliers, and discover latent relationships [37].

Interestingly, there have been several recent reports indicating that supervised and unsupervised methods can be combined to enhance the process of feature selection and metabolite annotation, which subsequently leads to increased reliability of LC-MS/MS analyses [32][38].

3.2 Deep learning

The field of medicine learning has also been increasingly used for analysing complex LC-MS/MS

Metabolomics data [6][39]. Models such as Convolutional neural networks (CNNs) and artificial neural networks (ANNs) have the potential to automatically identify non-linear patterns in spectral data to enhance metabolite annotation and peak deconvolution without the need to preprocess it manually on a large-scale basis [6][39]. DL can improve the discovery of biomarkers and make the interpretation of data more dependable by emphasizing small metabolomic changes as well as integrating with traditional Machine learning [40]. Altogether, Deep learning is becoming an influential instrument of LC-MS/MS Metabolomics as it can study the complicated patterns in high-dimensional data that are difficult to study by traditional methods [40].

3.3 Pattern recognition and clustering algorithms.

This type of algorithm is crucial in the study of intricate LC-MS/MS Metabolomics datasets [30][35]. Such unsupervised techniques are useful in determining intrinsic data structures, exposing outliers, and classifying metabolites whose behaviours are similar [36][37].

The discovery of natural groups of metabolites or samples helps in the discovery of biomarkers and the addition of information on biological variability without any labels. Besides, the clustering may be used with the methods of pattern recognition, including PCA, which would enable researchers to visualise the differences in metabolic patterns of experimental groups [37]. In general, pattern recognition and clustering cannot be or will never be discarded in the organisation of data, the uncovering of relationships that are not evident, and assisting downstream analysis in systems biology and clinical studies [37].

3.4 Natural language processing (NLP)

NLP methods are being more actively used in metabolomics to import any valuable information from a large scientific literature and databases [29][40]. NLP can automatically extract published studies, identify metabolite-disease interactions, aggregate results of the experiments, and facilitate hypothesis testing [29]. A combination of NLP and computational metabolomics pipelines can empower a researcher to effectively connect NLP-LC-MS/MS data to existing knowledge to facilitate biomarker identification and functionalization faster [29][41].

Table 1. summarizes selected AI and machine learning tools applied in LC-MS/MS metabolomics

TASK	AI TOOL/METHOD	ALGORITHM/APPROACH
Peak detection	PeakBot	Convolutional Neural Network (CNN)
Spectral matching	MS2DeepScore	Neural network similarity scoring
Fragment prediction	MIST-CF	Deep learning
Fragment prediction	SingleFrag	Deep learning
Metabolite annotation	CFM-ID	Probabilistic fragmentation modeling
Biomarker discovery/ classification	Random forest, SVM	Supervised machine learning

Multi-omics integration	Graph neural networks, Variational Autoencoders	Deep learning

workflows. The tasks indicate the primary application of each tool, and the algorithm or approach specifies the underlying AI method.

4. APPLICATIONS OF AI IN LC-MS/MS METABOLOMICS

Noise, baseline drifts, and matrix effects often tends to influence the complicated procedures involved in LC-MS/MS Metabolomics (Data preprocessing, metabolite identification, and quantification). AI and machine-based learning methods have been used to overcome these difficulties, which allows the detection of peaks using automated peak detection, spectral prediction, robust quantification, and biomarker discovery. The next section points out the most important AI applications throughout the LC-MS/MS workflow.

a. Data preprocessing

Preprocessing of LC-MS/MS metabolomics data includes baseline correction, noise reduction, automated peak detection and deconvolution, and retention time alignment. The true metabolite signals can be obscured by baseline drifts and random noise [33].

AI-driven methods have addressed these limitations. For instance, one of the machine learning approaches is PeakBot developed for detection of chromatographic peaks in LC-HRMS profile-mode data [33]. It works by detecting local signal maxima in a chromatogram, which are then extracted as super-sampled standardized areas (retention-time versus m/z), followed by inspection using convolutional neural network [33]. In training and independent validation datasets, PeakBot demonstrated exceptional performance in distinguishing between chromatographic peaks and background signals, achieving an accuracy of 0.99 [33]. PeakBot is developed in Python (v3.8) with TensorFlow (v2.5.0) and is freely available for noncommercial use under a CC-BY-NC-SA license and has been successfully tested on Linux and Windows systems [33]

In LC -MS metabolomic, pre-processing steps, such as baseline correction and noise reduction are critical for accurate resources and reliable quantification [33]. Sun et al. (2024) emphasized that basal deviations resulting from instrument fluctuations; spine bleeding and background chemical noise can obscure genuine analyte signals, making correction an essential step [34]. More Recently, Wang and Chen introduced an Artificial Intelligence - Driven Search Algorithm That Dynamically Adapts to Complex Spectral Backgrounds, Outperforming Traditional Methods in Separating Metabolite peaks From Baseline Distortions [34]. Similarly, noise reduction increases the signal / noise ratio, minimizing the random variability introduced during ionization and detection. Established techniques such as Savitzky - Golay Smoothing, Fourier Transform Filtering and Wavelet Denishing are widely applied; However, AI-based emerging Denishing models now allow the suppression of random noise without distorting the true forms of chromatographic peak [33][34]. Together, these developments illustrate how AI is advancing in pre-processing data to improve sensitivity, reproducibility and accuracy in the LC-MS/MS metabolomics.

b. Metabolite identification

Traditionally, metabolite identification in LC-MS/MS anticipates on spectral library matching; however incomplete libraries and fragmentation variations limits the annotation confidence. Recent AI-based approaches have improved database searching in order to overcome this challenge [42].

Deep learning tools such as MIST-CF and SingleFrag enhance the accuracy of spectral matching by predicting the fragmentation spectra or intensity of fragment ion [36][37]. Similarly, MS2DeepScore utilizes neural networks for evaluating spectral similarity more robustly than conventional cosine score methods [38]. For molecules absent from libraries, AI models- including CFM-ID and graph-based neutral networks- capture molecular structure context to enable de novo annotation of unknown metabolites [39].

Such AI-based methods increase the proportion of metabolites that can be detected through LC-MS/MS, as well as improve the level of reproducibility and the confidence in the identification of the metabolites [37][38][39].

c. **Quantification and Calibration**

Co-elution of substances produces matrix effects that either inhibit or augment ionization that creates significant challenges to quantitative analysis. The AI models are capable of forecasting and rectifying such effects when compared to the current conventional strategies,which are based on the internal criteria or sample dilution [34]. Supervised algorithms - including random forests, gradient boosting, and studies on deep learning frameworks studies the correlation between composition of the sample and variability of ionization, and hence enhances the quality of calibration [34][40].

Convolutional and recurrent neural networks are deep learning architectures that reduce the inter-sample variation [41][42]. Integration of advanced quantification workflows and Artificial Intelligence-driven matrix effect forecasting increases sensitivity, reliability and reproducibility of complex biological samples [32]

d. Discovery and pattern recognition of biomarkers

In metabolomics, AI and machine learning are broadly used in the discovery of biomarkers. Random forests, support vector machines and gradient boosting are examples of supervised algorithms that rigorously differentiate between biological samples, hence identifying disease and healthy conditions [32][40]. The top performing variants of feature selection, such as LASSO regression, recursive feature elimination as well as random forest importance scoring are used to determine the most informative metabolites [32][40]. Likewise, dimensionally reduction methods like PCA, t-SNE, and UMAP are useful in visualizing the data of high dimensions and minimising the noise [32][40]. Such methods can be combined with the classification that facilitates biomarker discovery, enhances interpretability, predictive modelling, and candidate biomarker discovery to detect disease and personalized medicine [32].

e. Combination with various omics data

A combination of metabolomics, genomics, proteomics, and transcriptomics is useful in the study of system-level biological processes[41]. Machine learning and AI are able to handle heterogeneous multi-omics data and show the correlations and causality between the molecular layers [32][41].

Graph neural networks and variational autoencoders are deep multi-model learning techniques that enable the joint analysis of various omics datasets [42]. This allows the analysis of networks, reconstruction of pathway and prioritization of candidate biomarkers to reveal molecular interactions that single-omics studies may

overlook, thereby improving specificity, sensitivity and interpretability in system biology investigation [41][42].

5. CASE STUDIES AND EXAMPLES

Artificial intelligence (AI) contributed substantially to different procedures of data interpretation and processing within LC-MS/ MS metabolomics pipelines. One such notable development is the introduction of end-to-end deep learning models, e.g., DeepMSProfiler, that can directly process the raw LC-MS data, and also extract the features and classify them using the same pipeline. This method has been more powerful, accurate and interpretable in high throughput metabolomics investigations rather than the traditional techniques which need multiple preprocessing and peak curation procedures. It has reduced batch impacts, and improved repeatability [14].

The process of detecting and curating the peak used to take time and was prone to inconsistency but AI has helped advance the process. NeatMS is software that significantly reduces the manual evaluation of large cohort studies that is labor-intensive by utilizing convolutional neural networks (CNNs) to distinguish actual chromatographic peaks and noise [16]. On the same note, recent advances like MassCube provide more compact algorithms of peak deconvolution and feature alignment to increase sensitivity and reproducibility of untargeted metabolomics data sets [43]. These methods address a longstanding problem of LC-MS/MS processes high-quality and consistent feature extraction on heterogeneous data sets.

Aside from preprocessing, metabolite identification through MS/MS spectral interpretation has been facilitated by deep neural network models. Rule-based spectral matching is much less accurate than models that predict chemical fingerprints from experimental spectra or sort out candidate structures. The extent of chemical space in biological science has been grown, for example, with the advent of novel deep learning-based fingerprint predictors in improving the assignment of structures in untargeted metabolomics [44]. This has significant implications to translational research, as structural elucidation is still a major bottleneck.

AI-powered LC-MS/MS metabolomics are also increasingly being used for clinical applications. Metabomic profile-based machine learning classifiers have been found to effectively diagnose disease and predict prognosis. For instance, it was demonstrated that machine learning algorithms based on metabolites could identify minute solid biomarker panels with great diagnostic capability in metabolic disease and cancer [45]. In pharmacometabolomics, ML-enhanced metabolomics methods are assisting in informing personalized medicine approaches and target pathway-based therapeutic intervention by facilitating patient classification based on drug response patterns [46].

AI has also proved to be useful in metabolomics, especially in agricultural and environmental metabolomics. To assist in food authentication and environmental surveillance, machine learning and chemo metric methods have been applied together with LC-MS data to analyze plant metabolite fingerprints, geographic origin, and agricultural practices [47]. AI has also been utilised in microbial metabolomics research to differentiate between resistant and susceptible bacterial strains and has been found to be potentially useful as a rapid diagnostic tool for the treatment of infectious diseases [48].

All these illustrate that AI has a wide ranging effect throughout LC-MS/MS metabolomics, including raw data processing, metabolite annotation, clinical diagnostics, pharmacology, and environmental uses.

6. CHALLENGES AND LIMITATIONS

There are several obstacles that prevent the implementation of AI in LC-MS/MS metabolomics, despite the rapid development. The major limitation is the requirement of extensive and annotated training sets. The lack of valid spectral libraries that verify the identities of the metabolites, their retention times, and data makes the use of supervised learning models in other biological contexts challenging [49][50]. Inter-instrument variability and inter-laboratory variability contribute to this shortage of extensive training data, curtailing model performance when applied beyond the training range.

The interpretability of AI models is the other key issue. While deep neural networks have surpassed conventional methods in most benchmarking tasks, they are "black box" machines that are hard to understand biologically and for use in the clinic. While explainable architectures, used in DeepMSProfiler, are one such example of ongoing attempts to handle this by encouraging decision making transparency, the large-scale adoption of standardized interpretability techniques is still lacking [14].

Increased application of AI methods is also hindered by their need for computing resources. High-resolution LC-MS/MS datasets are computationally expensive to train and deploy complex deep models (GPUs, high-memory architectures) that could be unaffordable for small labs or clinics [49]. As domain shift reduces the transferability of AI models because of differences in chromatography, fragmentation approaches, or population groupings, transfer learning and domain adaptation techniques have been innovated [46].

Additionally, there are no accepted benchmarking datasets or measures of evaluation in the community that would facilitate comparison of the efficacy of AI software for peak picking, deconvolution, or metabolite identification. Therefore, without such benchmarks, reproducibility and validation across labs are restricted [16].

Lastly, since most AI solutions are research prototypes and not user-friendly, proven software, integration into current workflows is difficult. Severe validation, LIMS compatibility, and process documentation are required for practical application in controlled environments to satisfy clinical or regulatory requirements [46].

AI can revolutionize LC-MS/MS metabolomics, but translating it into typical biomedical and environmental applications will rest on the overcoming of these challenges: data availability, interpretability, accessibility of computation, standardisation, and workflow integration.

7. FUTURE PERSPECTIVES AND OPPORTUNITIES

The use of artificial intelligence (AI) in the future is likely to transform the way that metabolomics is utilized in research and healthcare significantly. AI can provide the entire picture of the work of the diseases in the body by integrating the information about various areas of the so-called omics, such as genomics, proteomics, and metabolomics. This can assist scientists in identifying all new metabolic pathways and identifying helpful biomarkers in diagnosis and treatment. In personalized medicine, AI can analyze the specific metabolic characteristics of the individual to what effect they will take certain drugs or how their illness will develop. It is then possible to develop more effective and less side-effect treatments. AI also has the ability to accelerate the

process of biomarker discovery by analyzing big and complex data sets faster than a man would analyze the data and is relevant in assisting in the early detection of diseases and monitoring treatment [51].

A secondary advantage is that of automation, AI-driven systems have the potential to gather and process metabolomics in real-time, which may be utilized in hospitals and labs to swiftly screen health and make medical-related decisions. In drug discovery, AI is used together with metabolomics to identify how a new drug will act in the body and its metabolism, safety and the side effects. This aids in saving time, cost and animal testing in the development. Even the anticipation of the diseases before they manifest themselves can be aided by AI, and healthcare can take a more preventative approach instead of a curative one. Standardized AI-based data will facilitate the sharing and comparison of metabolomic data across scientists in the future. Last but not the least, it will be necessary to create clear and understandable AI models, which will allow doctors and researchers to trust them and use such tools in clinical practice [51].

Artificial intelligence (AI) makes the bioprocess instruments dynamic and autonomous in control by interpreting real-time data and adaptive decision-making. AI models in biorefineries and bioprocessing include reinforcement learning and model predictive control, which can be used to perform continuous sensor feedback optimization to maximize parameters like temperature, pH and nutrient flow. This guarantees the stable functioning and increased yield at a low consumption of resources. With digital twins and soft-sensing technologies, the simulations can be used to make adjustments and predictive corrections without disrupting the current processes. Adaptive sampling is an AI-based sampling strategy that uses critical variations in processes to develop a more precise model and eliminate redundant measurements. A combination of these technologies can be used to provide better precision, efficiency, and sustainability of self-optimizing bioprocesses [52].

Standardizing AI pipelines to analyze metabolomics is one of the crucial steps towards the increased trustworthiness of data analysis, its transparency, and broader applicability. Today, various data formats, techniques of preprocessing and a variety of analytical tools are commonly used in metabolomics studies, thus giving conflicting outcomes in different laboratories. Establishing uniform AI pipelines, researchers are able to get the data of different sources processed and analyzed in a similar manner. These pipelines would integrate automated processes in data cleaning, data normalization, feature extraction, statistical modeling and machine learning-based interpretation [53].

The standardized AI workflow would facilitate easier comparison of studies, reproduction of findings and combination of metabolomics data with other omics disciplines like genomics and proteomics. It would also accelerate the discovery of biomarkers and eliminate human error through the automation of complicated analytical procedures. Besides, open-source and cloud-based AI platforms might enable the world of researchers to exchange data and models, which would stimulate collaboration and openness. Such standardized pipelines will be necessary in the future to make AI-driven metabolomics accessible to clinical and pharmaceutical application, and so forth valid, repeatable, and understandable outcomes that can be relied upon in making the decision [53].

The democratization of metabolomics by AI is possible as it turns the complicated datasets into standardized, machine-readable knowledge models that can be accessed by experts and non-experts alike. By introducing high levels of standardization, explicit data encoding, and the critical efforts of biocurators, raw metabolomics data may be translated into interoperable formats, which can be interpreted by people and processed by computer programs. It allows reuse of data on a larger scale, interdisciplinary integration and creation of AI-based tools that can hasten the discovery speed. Although routine information is easily standardized, new metabolomics technologies still need expert consensus and curation cycling. To maximize the potential of AI to reduce barriers, improve reproducibility, and democratise metabolomics research, it is important to invest in data infrastructure and biocuration over a long period [54].

8. CONCLUSION

The artificial intelligence has become a strong facilitator in LC-MS/MS based metabolomics to overcome most of the challenges that have been posed over the years regarding data complexity, feature extraction, metabolite identification and biological interpretation. The combination of machine learning, as well as deep learning technologies have played an important role in enhancing crucial steps of the metabolomics workflow, such as preprocessing of data, peak detection, deconvolution, spectral matching and quantification. PeakBot, MS2DeepScore, MIST-CF, SingleFrag, and CFM-ID are all examples of AI-driven tools that can be found to be more sensitive, accurate, and reproducible in comparison with traditional rule-based tools. Outside the field of data processing, AI has broadened the horizon of metabolomics through increasing the identification of biomarkers, disease classification and integration of multi-omics. Supervised/unsupervised learning paradigms can be used to identify disease-related metabolic signatures with high confidence; multifaceted deep learning models can be applied to provide the information on the systems level, i.e. to correlate metabolite signatures with genomic, proteomic, and transcriptomic data. These advancements point to the increased applicability of AI-assisted metabolomics in precision medicine, pharmacometabolomics, clinical diagnostics, and in the environmental and agricultural fields. Nevertheless, despite such advances, there are a number of limitations that impede the extensive use of AI in LC-MS/MS metabolomics. Inadequate access to large, well-annotated training set, inter-laboratory variance, minimal model explainability, high computational cost and standardized benchmarking infrastructure are all still major challenges. These dilemmas will be essential in solving the challenge of applying AI methodologies in research environments and integrating them into the daily clinical and industrial practice. The next steps will rely on the creation of standard, explainable, and end-to-end AI pipelines with the help of shared repositories of metabolomics and community-based benchmarks. Accessibility and reproducibility as well as trust will be further improved by cloud-based platforms and transparent AI models. In general, the further integration of artificial intelligence and LC-MS/MS metabolomics is a huge prospective in changing the metabolomic studies and facilitating the application of the research in healthcare and systems biology.

Conflicts of interest:

The authors declare that they have no conflicts of interest with the article's contents.

Acknowledgments:

We greatly appreciate the support for this review article provided by Yenepoya Pharmacy College & Research Centre, Mangalore, Karnataka, India.

Funding statement: The authors have not received any funding from any source.

Abbreviations:

AI: Artificial intelligence

ANN: Artificial Neural Networks

CNN: convolutional neural network

CFM-ID: Competitive fragmentation modelling for metabolite identification

DL: Deep learning

ML: Machine learning

MIST-CF: Metabolite interference with spectrum transformers for chemical formulae

PCA: Principal component analysis

LASSO: Least absolute shrinkage & selection operator

t-SNE: t- distributed stochastic neighbor embedding

LCMS/MS: Liquid Chromatography–Mass Spectrometry

NLP: Natural Language Processing

GNPS: Global Natural Products Social Molecular Networking

METLIN: METabolite LINk

SHAP: Shapley Additive exPlanations

LC-HRMS: Liquid chromatography- High resolution mass spectrometry

LIMS: Laboratory information management system

GPU: Graphics processing unit

UMAP: Uniform manifold approximation and projection

XCMS: Extensible Computational Mass Spectrometry

REFERENCES

1. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol.*2012;13(4):263-269. https://doi:10.1038/nrm3314.
2. Dunn, W.B., Erban, A., Weber, R.J.M. *et al.* Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **9** (Suppl 1), 44–66 (2013). https://doi.org/10.1007/s11306-012-0434-4.
3. Cajka T, Fiehn O. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. *Anal Chem.* 2016;88(1):524-545. https://doi:10.1021/acs.analchem.5b04491.
4. Misra BB, van der Hooft JJJ. Updates in metabolomics tools and resources: 2023-2024. *Metabolomics.*2024;20(3):46.
5. Zhou B, Xiao JF, Tuli L, Ressom HW. LC–MS-based metabolomics. *Mol Biosyst.* 2012;8(2):470-481.

6. Pang Z, Chong J, Li S, Xia J. MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. Metabolites. 2020 May 7;10(5):186.

7. Das S, Reddy PH. Artificial intelligence and machine learning in metabolomics research. *Biomed Pharmacother*.2021;142:112045. https://doi:10.1016/j.biopha.2021.112045.

8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. https://doi:10.1038/nature14539.

9. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol.* 2016;17(7):451-459.

10. Triba MN, Le Moyec L, Amathieu R, et al. Untargeted metabolomics and machine learning: a synergistic approach for biomarker discovery. *Comput Struct Biotechnol J.* 2015;13:55-64.https://doi:10.1016/j.csbj.2015.02.005.

11. Wehrens R, et al. Artificial intelligence in metabolomics: from peak picking to biomarker discovery. *Anal Chim Acta*. 2022;1200:339611.

12. Misra BB. Data complexity in mass spectrometry-based metabolomics: opportunities and challenges. *Anal Bioanal Chem.* 2021;413(9):23852403.

13. Broadhurst D, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics.* 2006;2(4):171196. https://doi:10.1007/s11306-006-0037-z.

14. Deng Y, Yao Y, Wang Y, et al. An end-to-end deep learning method for mass spectrometry data analysis to reveal disease-specific metabolic profiles. *Nat Commun.* 2024;15:7136. doi:10.1038/s41467-024-47136-1.

15. Taylor PJ. Matrix effects: the Achilles heel of quantitative high performance liquid chromatography-electrospray-tandem mass spectrometry. Clin Biochem 2005: 38 (4):328-334.https://doi:10.1016/j.clinbiochem.2004.10.022.

16. Gloaguen Y, Kirwan JA, Beule D. Deep learning-assisted peak curation for large-scale LC-MS metabolomics. Analytical chemistry. 2022 Mar 15;94(12):4930-7. https://doi:10.1021/acs.analchem.1c04811.

17. Pang Z, Chong J, Xia J. Recent advances in computational metabolomics: data integration and interpretation. *Metabolomics.* 2021;17(7):55. https://doi:10.1007/s11306-021-01805-2.

18. Schymanski EL, et al. Identifying small molecules via high-resolution mass spectrometry: communicating confidence. *Environ Sci Technol.* 2014;48(4):2097-2098. https://doi:10.1021/es5002105.

19. Creek DJ, et al. Retention time prediction using machine learning to improve metabolite annotation. *Anal Chem.* 2020;92 (15): 10194-10202. https://doi:10.1021/acs.analchem.0c02061.

20. Bach E, Schymanski EL, Rousu J. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. *Nat Mach Intell.*

2022;4:1224-1237. https://doi:10.1038/s42256-022-00568-8.

21. Engskog MK, et al. Machine learning–based normalization strategies in LC–MS metabolomics. *Anal Bioanal Chem.* 2023;415:1567–1581.
22. Misra BB, Lange BM. Metabolomic dark matter: addressing the unknowns in metabolite identification. *Trends Anal Chem.* 2021;143:116331. https://doi:10.1016/j.trac.2021.11633
23. Zhou J, et al. Machine learning approaches for metabolomics-based biomarker discovery in disease diagnosis. *Metabolomics.* 2020;16(5):89.
24. Liu X, et al. AI-driven metabolomics for clinical biomarker discovery: recent advances and applications. *Front Mol Biosci.* 2023;10:1123459.
25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst.*2017;30:4765–4774.
26. Broadhurst D, Goodacre R, Reinke SN, Kuligowski J, Wilson ID, Lewis MR, et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*. 2018;14(6):72. https://doi:10.1007/s11306-018-1367-3
27. Misra BB. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *Eur J Mass Spectrom*. 2021;27(3):165–92. https://doi: 10.1177/1469066720918446
28. Zampieri M, Sekar K, Zamboni N, Sauer U. Frontiers of high-throughput metabolomics. *Curr Opin Chem Biol*. 2018;36:15–23.
29. Chi J, Zhao Y, Hu C. Artificial intelligence in metabolomics: A current review. *Pharmacol Res*. 2024;197:107333.
30. Mahieu NG, Patti GJ. Systems-level annotation of a metabolomics data set reduces 25,000 features to fewer than 1,000 unique metabolites. *Anal Chem*. 2017;89(19):10397–406.
31. Zhang A, Sun H, Wang P, Han Y, Wang X. Recent and potential developments of biofluid analyses in metabolomics. *J Proteomics*. 2020;219:103758.
32. Pang Z, et al. Integrative machine learning approaches for metabolomics feature selection and classification. *Bioinformatics*. 2021;37(14):1957–65.
33. Bueschl C, Doppler M, Varga E, Seidl B, Flasch M, Warth B, et al. PeakBot: Machine-learning-based chromatographic peak picking. *Bioinformatics*. 2022;38(13):3422–8.
34. Sun J, Xia Y. Pretreating and normalizing metabolomics data for statistical analysis. *Genes Dis*. 2024;11(3):100979.
35. Wang X, Chen Y. Baseline correction based on a search algorithm from artificial intelligence. *Appl Spectrosc*. 2021;75(5):531–40.
36. Goldman S, Xin J, Provenzano J, Coley CW. MIST-CF: chemical formula inference from tandem mass spectra. *Nat Mach Intell*. 2023;5(8):861–71.
37. Pérez-Ribera M, Khan MF, Giné-González M, Olivella J, Roldan C, Vilarrasa J, et al. SingleFrag: a deep learning tool for MS/MS fragment and spectral prediction and metabolite annotation. *Brief*

Bioinform.
2025;26(4):bbaf333.

38. Huber F, Ridder L, Verhoeven S, Rogers S. MS2DeepScore: a novel deep learning similarity measure for mass fragmentation spectra. *J Cheminform*. 2021;13:84. https://doi:10.1186/s13321-021-00558-4
39. Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res*. 2014;42(W1):W94–9.
40. Chi Z, Li J, Liu Z, Zhang X, Chen L. Artificial intelligence-driven metabolomics for biomarker discovery and systems biology. *Metabolomics*. 2024;20(3):29.
41. Zhou Z, Luo J, Xu Y, Zhu ZJ. Machine learning and artificial intelligence in metabolomics: Applications and future perspectives. *TrAC Trends Anal Chem*. 2023;163:117012.
42. Roberts LD, Souza AL, Gerszten RE, Clish CB. Challenges and opportunities in quantitative metabolomics using liquid chromatography– mass spectrometry. *Nat Rev Chem*. 2022;6(8):564–81.
43. Yu H, Zhang L, Wang S, Ding J, Shen T, Liu M, et al. MassCube improves data processing for metabolomics from raw files to phenotype classifiers. Nat Commun. 2025;16(1):5487. https://doi:10.1038/s41467-025-60640-5.
44. Chau HYK, Ao H, Zhang X, Gao S, Varghese RS, Ressom HW. Deep learning–based molecular fingerprint prediction from MS/MS spectra. J Proteome Res. 2025;24(1):112-23.https://doi:10.1109/EMBC40787.2023.10341007.
45. Chen Y, Wang B, Zhao Y, Shao X, Wang M, Ma F, et al. Metabolomic machine learning predictor for diagnosis and prognosis of gastric cancer. Nat Commun. 2024;15(1):1657. https://doi:10.1038/s41467-024-46043-y.
46. Che Y, Liu Z, Zhao P, et al. Application of machine learning for mass spectrometry–based metabolomics. Front Mol Biosci. 2024;11:1483326. https://doi:10.3389/fmolb.2024.1483326.
47. García-Pérez P, Losada-Barreiro S, Bravo-Díaz C, et al. Metabolomics and chemometrics in agri-food and environmental sciences. Trends Food Sci Technol. 2024;147:104481. https://doi:10.1016/j.tifs.2024.104481.
48. Chávez-Pérez A, Ríos-Castro E, Jiménez-Mendoza A, et al. AI-driven metabolomics for microbial resistance diagnostics. Microbiol Spectr. 2025;13(1):e01025-24.
49. Misra BB. Challenges in AI applications in metabolomics: data and interpretability. Metabolomics. 2023;19(4):58.
50. Coler EA. Metabolomics in the era of artificial intelligence: promises and pitfalls. Trends Biotechnol. 2024;42(6):715-27.
51. Yipeng Xu, Xiaojuan Jiang, Zeping Hu, Synergizing metabolomics and artificial intelligence for advancing precision oncology, Trends in Molecular Medicine, Volume 31, Issue 8, 2025,

Pages 692-701

52. Butean, A.; Cutean, I.; Barbero, R.; Enriquez, J.; Matei, A. A Review of Artificial Intelligence Applications for Biorefineries and Bioprocessing: From Data-Driven Processes to Optimization Strategies and Real-Time Control. Processes 2025, 13, 2544.

53. Mitchell JM, Chi Y, Thapa M, Pang Z, Xia J, Li S (2024) Common data models to streamline metabolomics processing and annotation, and implementation in a Python pipeline. PLoS Comput Biol 20(6): e1011912.

54. Dessimoz, C., Thomas, P.D. AI and the democratization of knowledge. *Sci Data* 11, 268 (2024).