Review Articles

Acupuncture and Herbal Medicine

From fragments to function: a review of AI-driven mass spectrometry and fragment-based strategies

Tianyi Ma[1,2], Zhaoxin Xu[1,2], Yugang Lin[3], Jie Liao[1,2,*], Xiaohui Fan[1,2,*]

[1]State Key Laboratory of Chinese Medicine Modernization, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China; [2]Zhejiang Key Laboratory of Chinese Medicine Modernization, Innovation Center of Yangtze River Delta, Zhejiang University, Jiaxing, China; [3]Department of Pharmacy, Affiliated Jinhua Hospital, Zhejiang University School of Medicine, Jinhua, China

Abstract

Traditional Chinese medicine (TCM) possesses unique advantages in disease prevention and treatment, yet its inherent complexity and diversity pose tremendous challenges for structural elucidation, mechanism research, and bioactivity characterization. High-resolution mass spectrometry (HRMS) technology demonstrates immense potential in TCM analysis due to its high sensitivity, high resolution, high throughput, and high efficiency. However, its application in TCM research remains constrained by the lack of intelligent analytical methods and unified standardized databases. Therefore, this paper focuses on the integration of artificial intelligence (AI) and mass spectrometry, providing a systematic review of the applications and potential of AI-driven mass spectrometry analysis in structural elucidation, data resource integration, multi-omics mechanism studies, and chemical biology. Furthermore, this article emphasizes that by leveraging AI models to learn the complex mapping from chemical structures to biological functions, fragment-based characterization has emerged as the bridge connecting chemical structures with biological activities. Molecular fragments themselves serve as the core "knowledge units" that carry bioactive information. Future research will focus on establishing high-quality mass spectrometry databases for the complete chemical profiles of TCM and promoting the standardization and open sharing of mass spectrometry databases, thereby advancing the integration of AI and mass spectrometry in TCM analysis and providing new tools for TCM research.

Keywords: Artificial Intelligence, Fragment-based representation, Mass spectrometry, Structural elucidation, Traditional Chinese medicine

Graphical abstract: https://links.lww.com/AHM/A220

Introduction

Traditional Chinese medicine (TCM) is a precious treasure of the Chinese nation, having attracted widespread attention for its unique advantages in preventing and treating complex diseases[1]. As an important branch of natural products, TCM is also a primary source for the discovery and design of small-molecule drugs[2]. However, the research progress of TCM is constrained by various factors: the complexity and diversity of TCM components pose challenges for structural elucidation[3]; the multi-component and multi-target characteristics of TCM make the screening of active components and the exploration of their mechanisms of action complex[4]. Therefore, how to promote the modernization of TCM research has become a key issue in the field of TCM.

In the field of TCM research, mass spectrometry(MS) is often combined with liquid chromatography (LC) and widely applied in TCM component identification, qualitative and quantitative analysis, metabolomics, and TCM quality control[5–6]. However, manual analysis and annotation of complex MS data generated from TCM are highly dependent on specialized knowledge and are time-consuming and labor-intensive, which significantly hinders the progress of related research[3]. Emerging technologies such as single-cell mass spectrometry imaging (MSI) now offer unprecedented spatial resolution, enabling the mapping of TCM compound distribution and action heterogeneity at the single-cell level. Matrix-assisted laser desorption/ionization (MALDI)-MSI combined with single-nucleus RNA sequencing (snRNA-seq) revealed ginsenoside spatial heterogeneity in ginseng at single-cell resolution, guiding metabolic engineering[7].

Addressing the aforementioned challenges urgently requires technological innovations. The proliferation of digital technologies—particularly big data, artificial

Tianyi Ma and Zhaoxin Xu contributed equally to this work.

**Corresponding author. Jie Liao, E-mail: liaojie@zju.edu.cn; Xiaohui Fan, E-mail: fanxh@zju.edu.cn.*

How to cite this article: *Ma TY, Xu ZX, Lin YG, Liao J, Fan XH. From fragments to function: a review of AI-driven mass spectrometry and fragment-based strategies. Acupunct Herb Med 2026;6(1):28–41. doi: 10.1097/HM9.0000000000000190*

Received 25 November 2025 / Accepted 10 February 2026

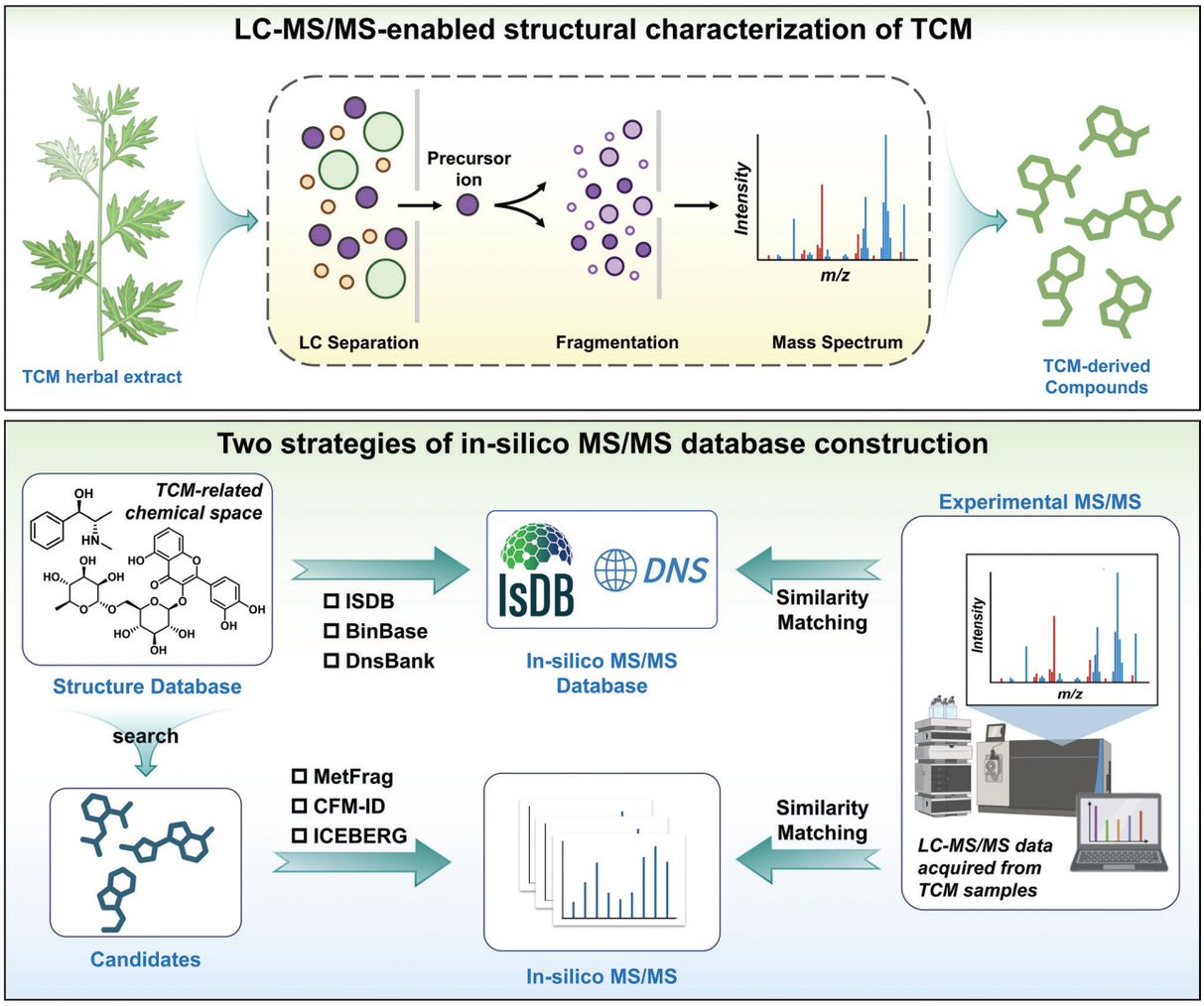

Figure 1. LC-MS/MS-based structural characterization and *in-silico* MS/MS annotation workflows for compounds in TCM. LC-MS/MS is used to acquire tandem mass spectra from complex TCM herbal extracts for structural characterization. Two *in-silico* MS/MS annotation strategies are illustrated: (i) static *in-silico* databases generated from large-scale structure repositories and matched directly to experimental spectra, and (ii) dynamic candidate-based approaches in which virtual spectra are predicted on demand for candidate compounds and ranked by spectral similarity. These strategies enable the annotation of TCM-derived compounds beyond existing experimental spectral libraries. LC: Liquid chromatography; MS: Mass spectrometry; TCM: Traditional Chinese medicine.

intelligence (AI), and, more recently, large language models (LLMs)—is propelling TCM into a new era of digital intelligence, fostering systematic innovation and modernization[8–10]. LLM4MS leverages LLMs to generate discriminative spectral embeddings, significantly improving MS-based compound identification accuracy and speed[11]. This AI-driven MS analysis paradigm has been widely applied in multi-omics research[12] and is gradually penetrating the field of TCM.

This article will focus on the integration of AI and MS analysis, exploring its potential and applications in TCM research: structural analysis, data resource integration, multi-omics mechanism research, and chemical biology applications. It mainly includes intelligent analysis methods for the structure of TCM components, functional prediction methods based on molecular fragments, and mechanism exploration methods combined with omics research. Some of the models in this article are currently mainly applied to the field of small-molecule drugs, but their principles are also of great reference value for the analysis of complex TCM systems. This article reviews the cross-scale application of "AI + mass spectrometry" in the "chemical structure-bioactivity" relationship, driving the evolution of fragments from "breakdown units" to "knowledge units", ultimately providing new insights for TCM research.

AI-assisted structural analysis technology

Nowadays, high-resolution MS is widely used for characterizing natural products with different structures (Figure 1)[3]. Tandem MS/MS generates fragment ions through collision-induced dissociation (CID), providing detailed structural information[13]. However, for complex samples containing a large number of unknown compounds, challenges such as manual annotation difficulties and insufficient reference MS databases remain[3]. Meanwhile, MSI enables the detection of the spatial distribution of compounds on the surface of tissues or samples, compensating for the lack of spatial resolution in MS/MS. However, MSI primarily focuses on obtaining the spatial distribution of parent ions and does not

include fragment information, which restricts the application of MSI-based structural annotation methods[14–15]. In recent years, AI has provided a new paradigm for structural analysis strategies. Therefore, this paper systematically reviews AI-assisted tandem MS/MS and MSI structural annotation methods to provide insights for the structural analysis of TCM components.

Annotation method based on tandem MS

Structural annotation based on dependency library matching

For MS/MS annotation of known compounds, the most widely used method is library matching[16]. Library matching–based methods typically generate a ranked list of candidate molecules based on similarity scores between query spectra and reference spectra[16–17]. Among these, cosine similarity is the most commonly used spectral similarity metric, as it is more sensitive to structural similarities between molecules[16]. Kim et al.[17] investigated the performance of 15 binary similarity metrics in compound identification. The Huber team[18] innovatively applied the Word2Vec algorithm from NLP to MS data, creating the Spec2Vec score. Through data validation, they demonstrated that Spec2Vec similarity better reflects structural similarity than traditional cosine scores. Meanwhile, similarity scores lay the foundation for constructing MS molecular networks (MN). Dorrestein's team[19] first reported MS/MS-based MN in 2012. The principle is that molecules with similar structures will show strong similarities in their MS/MS spectra. By comparing the secondary mass spectra of different compounds, their structural similarity can be assessed, and an MN diagram can be constructed. In an MS MN, each node represents a compound, and the lines connecting the nodes indicate the structural relationships between the compounds. This method can intuitively display the structural relationships between compounds in complex samples. By observing the spatial location characteristics of MN node clusters, it accelerates the deduplication of MS data and the discovery of new structures.

Based on similarity matching and MS MN, researchers have developed a series of AI-assisted structural annotation algorithms. Wang et al.[20] introduced an on-line MS search engine, MASST, which integrates and queries spectral data from multiple public MS databases to perform similarity matching on target spectra. MS2DeepScore[21] employs a convolutional ANN architecture to predict the similarity between two chemical structures using only MS/MS spectra. MS2Query[22] integrates Spec2Vec[18] and MS2DeepScore[21] to construct a composite similarity score using Random Forests, achieving efficient and reliable MS analog search with a 40% improvement in structural similarity score at 35% recall. To address the coverage bottleneck of standard spectral libraries, Shen et al.[23] proposed an innovative hypothesis: metabolites in reaction pairs (e.g., substrates and products) within a metabolic reaction network (MRN) have structurally similar MS/MS spectra, leading to high correlation between their MS/MS spectra. Based on this hypothesis, MetDNA was developed using a small-scale standard spectrum library as a "seed" constructing an MN *via* MRN, and recursively annotating neighboring metabolites in reaction pairs. Ultimately, nearly 2,000 metabolites were annotated using an extremely small standard spectrum library (accuracy >77%), significantly alleviating reliance on "comprehensive" standard spectrum libraries. Building on MetDNA, to address complex spectral shifts caused by biotransformation, DeepMASS[24] calculates cross-correlation using fast Fourier transform (FFT) and captures nonlinear shifts between spectra through frequency domain transformation, addressing the limitations of traditional linear similarity calculations. Similarly, the coverage of MRNs remains limited. SGMNS[25] constructs a globally connected molecular network (GCMN) based on molecular fingerprint similarity, enabling annotation of over 2,000 metabolites with only 10 seeds (accuracy >83%). This breaks through the constraints of biological reaction pathways and covers a broader chemical space.

However, compared to the chemical space comprising over 68 million known compounds, the scale of publicly available MS libraries remains small[26], and a large number of MS/MS spectra cannot be matched to corresponding compounds in existing databases[27]. An annotation strategy relying on virtual spectrum libraries has emerged. Specifically, researchers use ML algorithms to predict virtual spectra of molecular structures, thereby constructing a "virtual spectrum library". Subsequently, experimental MS/MS spectra to be identified are compared with the virtual database, and the closest structure is matched based on similarity[16]. We divide the virtual spectrum libraries into a precomputed *in-silico* MS/MS library and an on-the-fly candidate-based *in-silico* MS/MS framework (Figure 1). The former are typically constructed using virtual spectral libraries based on massive compound databases such as PubChem[26] and KEGG[28], such as ISDB[29], BinBase[30], and DnsBank[31], and then employ the same structure analysis strategy as when matching real spectral libraries. The latter first identifies candidate compounds based on molecular weight and other information, predicts the virtual spectra of only the candidate compounds each time, thereby forming a dynamic "candidate spectrum set" and then ranks them based on similarity scores. From an application perspective, a precomputed *in-silico* MS/MS library substantially extends the coverage of conventional experimental spectral libraries by performing virtual fragmentation prediction for large-scale chemical spaces in advance, but they are associated with relatively high costs in construction and maintenance. In contrast, an on-the-fly candidate-based *in-silico* MS/MS framework generates virtual fragment spectra on demand, placing greater emphasis on computational efficiency and sample specificity.

From an algorithmic perspective, the virtual fragmentation strategy is based on chemical bond breaking rules to iteratively simulate the MS fragmentation process, enabling spectrum prediction[32]. Classic algorithms include: MetFrag[33] generates a fragment tree by iteratively breaking molecular bonds, assigns intensity weights, and outputs mass-to-charge ratio-intensity pairs; MS-FINDER[30] identifies cleavage sites using 228 literature-validated cleavage rules and combines hydrogen rearrangement to predict non-adjacent cleavages; CFM-ID[34] models the fragmentation process as a

Markov process, learns bond cleavage probabilities from data using an EM algorithm, and generates probabilistic spectra. Subsequently, the Allard team[29] applied CFM-ID to construct the ISDB, a virtual spectral library for natural products with a capacity exceeding 220,000 entries; ICEBERG[32] builds a generation-scoring framework based on GNN and Transformer to achieve deep integration between physical cleavage processes and neural networks; FIORA[35] predicts bond cleavage events through local molecular neighborhoods, modeling MS cleavage as an edge-level prediction task in RGCN (Relational Graph Convolutional Network)[tm1]. In addition, end-to-end DL models are increasingly being applied in this field, enabling direct learning of the mapping relationship between compound structures and MS/MS spectra without explicit chemical cleavage rules[36]. For example, MassFormer[37] uses a Graph Transformer to model the global structure of molecules, enabling joint prediction of MS/MS spectra, retention time (RT), and collision cross section (CCS); NEIMS[38] constructs a bidirectional gated MLP architecture, simulating the real fragmentation process through forward small fragment prediction and reverse neutral loss modeling; DeepCDM[31] transforms NEIMS into a dedicated predictor for chemically derivatized molecules *via* transfer learning; 3DMolMS[39] introduces a 3D point cloud convolution architecture (3DMolConv) to capture interatomic spatial relationships and achieve spectrum prediction. Our team also developed MassKG[3], an algorithm that integrates a knowledge-based fragmentation strategy and an RNN-based molecular generation model, thereby constructing a virtual fragment library for compound matching and comparison. In summary, by accurately predicting MS/MS spectra based on molecular structures, this approach expands the experimental reference standard material library in the field, thereby providing a breakthrough for the structural analysis of TCM components[32].

Structural annotation independent of library matching

The advantage of this virtual spectrum prediction method is that it only requires a list of candidate molecules or a structural database, without the need for a complete and authentic reference spectrum library. However, due to the limited coverage of the candidate molecule list, this method still cannot achieve *de novo* identification of unknown compounds, that is, it cannot identify molecules outside the list[34]. This contradiction is particularly evident in the study of complex TCM systems, because the systematic analysis of the chemical composition of TCM requires not only high-throughput characterization of known components, but also the establishment of new analytical paradigms for the discovery of unknown structures. Therefore, many ML methods that do not rely on spectral library matching have become effective alternatives for the elucidation of natural product structures[16].

A common approach is to convert MS/MS spectra into molecular fingerprints to bypass complex library matching. CSI-FingerID[40] is based on fragment tree guidance, converting mass spectra into molecular attribute fingerprints *via* multi-core SVM, followed by similarity matching in PubChem, overcoming the limitation of library size, but it still cannot predict unknown structures; Hoffmann et al.[41] combined the prediction capabilities of CSI:FingerID with a dual confidence mechanism to successfully identify 12 unreported natural bile acid structures from a virtual structure library; MSNovelist[42] pioneered a two-stage framework combining "fingerprint decoding (CSI:FingerID) and sequence generation (RNN)" to achieve the conversion from MS data to molecular fingerprints and then to molecular structure sequences, demonstrating the feasibility of *de novo* construction of molecular structures from MS/MS spectra.

Annotation method based on MSI

A standard procedure for metabolite annotation in MSI involves comparing experimental m/z values against a database of known molecular masses, constrained by a predefined mass error tolerance[43–44]. However, this integrated method has inherent limitations: since a single MS scan cannot distinguish between isotopes of the same mass or structurally similar isomers, multiple candidate molecules often match the same signal[45]. Therefore, relying solely on precise mass matching is insufficient for reliable molecular identification. Another approach is to combine MSI with tandem MS to obtain both MS/MS data and MSI data, thereby making the identification results more definitive[45]. However, this method is often limited by experimental conditions, such as the difficulty of achieving effective fragmentation of single-charge ions generated by MALDI, or the risk of signal loss and molecular migration when obtaining high-quality MS/MS data from the same tissue section[45]. Therefore, molecular identification methods based on MSI data have become a major challenge in this field.

With the development of AI technology, some bioinformatics tools have been developed that hold promise for solving this problem. The European Molecular Biology Laboratory developed the first automated annotation workflow for high-resolution imaging MS, pySM[46], and implemented it as a free and open-source annotation engine called METASPACE[47]. Its innovative features are as follows: (1) Original scoring system (MSM Score): Integrates spatial disorder, isotope spatial co-localization, and spectral similarity to jointly screen positive results; (2) Target-decoy false discovery rate (FDR) strategy: Constructs a decoy library using "impossible additive ions" to quantify and control false positives. Building on this, the team published METASPACE-ML[48] in 2023, replacing the original rule-based MSM scoring system with a GBDT model to achieve a data-adaptive annotation method. LipostarMSI[49] is the first software to cover the entire workflow from "raw data → preprocessing → statistical analysis → molecular identification → visualization". It combines precise mass matching, MS/MS fragment spectrum verification, spatial co-localization, and interactive identification to achieve high-confidence molecular identification. HIT-MAP[15] is specifically designed for spatial proteomics MSI data, utilizing an FDR-controlled two-tier validation framework to enable direct identification of MSI proteins without the need for LC-MS/MS assistance.

In summary, the fragment ion information provided by tandem MS spectrometry can significantly improve

the reliability of compound identification, and annotation tools for MS/MS have also formed a relatively systematic method system. However, most of these tools rely on high-quality reference spectrum libraries, and algorithms for compound design from scratch are still relatively scarce[16]. Therefore, in the study of complex TCM systems, there is still a lack of reference data, and issues such as the difficulty of annotating "dark matter" remain. In addition, it should be noted that TCM formulations typically consist of multiple chemically related constituents with markedly different abundances, which inevitably leads to component co-existence and signal interference during mass spectrometric acquisition and interpretation. However, most existing algorithms are still developed under a "single-molecule–single-spectrum" modeling assumption and do not explicitly account for synergistic effects or signal interference in multi-component systems. Therefore, in complex natural product systems such as TCM, how to incorporate inter-component interaction information at the MS annotation stage and achieve holistic modeling of mixture spectra remains a major challenge for AI-driven mass spectrometric analysis. Although MSI technology can simultaneously provide molecular spatial distribution and signal characteristics, the number of mature annotation tools is currently limited, the approaches are single-minded, and there is a lack of real molecular standard libraries[14]. Overall, although these two methods each have their advantages, their application in the complex system of TCM still faces significant limitations. The development of reference spectrum libraries and the deep integration of AI are urgently needed to advance high-precision annotation research of natural products. At the same time, it should be emphasized that structural annotation is not an end in itself, but rather provides the necessary foundation for functional annotation and for uncovering the bioactivity and therapeutic potential of TCM[50].

Integration and construction of MS resources

The vast amounts of data generated by MS experiments cannot realize their full value without effective organization and sharing. To address this, researchers have established multiple MS databases for the systematic storage, standardization, and reuse of MS data, enabling broad applications such as molecular identification and bioactivity exploration. However, the coverage and characteristics of different databases vary depending on specific research objectives and application scenarios; no single database encompasses all known compounds[51]. This review will outline commonly used MS databases in recent years, providing guidance for appropriate selection during research.

The Global Natural Products Social Molecular Networking (GNPS)[51] is an MS ecosystem integrating diverse data resources and analytical tools, designed to serve as an open-access knowledge repository for organizing and sharing raw, processed, or annotated fragment MS (MS/MS) data within the scientific community. Through diversified toolchains including MN, metadata integration (ReDU), and retrieval algorithms (MASST), it systematically supports the entire research workflow from known compound annotation to unknown molecule discovery. MassBank[52] is the first distributed public MS database focused on small-molecule compounds, covering data from multiple ionization techniques and instrument types. METLIN[53], publicly accessible since 2005, has become one of the largest secondary MS databases in metabolomics, housing over 431,000 high-resolution MS/MS spectra. The Human Metabolome Database (HMDB)[54] is the world's largest dedicated human metabolome database, covering over 98% of known human metabolites. As of 2022, it has integrated 37,589 experimental MS/MS spectra and predicted 1.44 million virtual spectra using CFM-ID 4.0, filling gaps in experimental data. Additionally, a small number of specialized databases focused on natural products or TCM have been developed. BMDMS-NP[55] is a high-precision MS/MS spectral library focused on plant-derived natural products, containing 288,000 spectra for 2,739 natural metabolites. It aggregates spectra of the same compound under different conditions into "blocks", enhancing cross-platform matching robustness. However, it only supports positive ion mode data and requires access *via* an SQL client, presenting a certain technical barrier. Thermo Fisher Scientific collaborated with Tsinghua University to release the Orbitrap Traditional Chinese Medicine Library (OTCML). Using the herbal medicines listed in the 2015 edition of the Chinese Pharmacopoeia (Part I) as a reference, it completed the acquisition of primary and fragment mass spectra for over 1,200 reference standards of TCM compounds, yielding more than 7,000 high-quality secondary mass spectra. This enables rapid and accurate characterization of TCM and natural product components. However, this database requires a fee for use. Other commonly used MS databases are summarized in Table 1.

The close integration of experimental data with bioinformatics tools has driven the development of resource repositories, with the breadth and richness of MS databases continually expanding and being updated[50]. However, the lack of unified standards across databases regarding data formats, annotation methods, and retrieval criteria has created obstacles for cross-database data integration and sharing. Therefore, storing diverse data in a standardized format has become the primary approach for achieving efficient retrieval. In 2004, the standardized format mzXML[56] was proposed. It employs Base64 binary encoding and an index scanning mechanism, achieving uniformity across multi-vendor instrument outputs through an XML schema. Subsequently, mzML[57] integrated features from both mzXML and mzData. While maintaining full compatibility with mzXML's functionality, it addressed issues of technical rigidity, ambiguous metadata, and ecosystem fragmentation, becoming the standard format promoted by HUPO-PSI. mzMLb[58] is a variant of mzML, enhancing read/write speed and storage efficiency for large datasets. Beyond data storage and format standardization, researchers have recently proposed standardized methods for data retrieval. For instance, MassQL[59] is a standardized query language enabling unified syntax across different spectral libraries to retrieve substructure features, fragment ions, or neutral losses.

The development and standardization of MS databases not only address data storage and sharing challenges but

Table 1

Summary of mainstream mass spectral databases

Database	Compound source	Pros and cons	Website
GNPS	General	1. Support for molecular networking analysis 2. Community-driven development and open-access data 3. Predominantly positive ion mode data	https://gnps.ucsd.edu/
MassBank	General	1. Spectra from multiple ionization modes and instrument types 2. Limited coverage of plant metabolites	https://massbank.eu/MassBank/
mzCloud	General	1. Spectral tree structure for multi-level MS^n data visualization 2. Plugin installation required for functionality	https://www.mzcloud.org/
NIST Mass Spectral Library	General	1. Authoritative GC-MS data 2. Paid access required	https://www.nist.gov/programs-projects/tandem-mass-spectral-library
Wiley Registry MS/MS	General	1. Equipped with the dedicated MS for ID search algorithm 2. Paid access required	https://msforid.com/
METLIN	General	1. Extensive compound coverage 2. Paid access required 3. Data download not supported	https://metlin.scripps.edu/auth-login.html
HMDB	Animals	1. Comprehensive coverage of human metabolites 2. Heavy reliance on predicted spectra	https://hmdb.ca/
ReSpect	Plants	1. Specialized MS^n spectral data for phytochemicals 2. Discontinued service	http://spectra.psc.riken.jp/
BMDMS-NP	NPs	1. Focused ESI-MS/MS spectra of NPs 2. Access *via* SQL client required	http://bmdms.bmdrc.org
OTCML	TCM	1. A collection comprising over 1,200 TCM compounds with more than 7,000 high-quality MS/MS spectra 2. Access available through a paid subscription	Access-restricted
METASPACE	General	1. Specializes in MSI 2. Enables structural annotation *via* FDR-controlled workflows	https://metaspace2020.org/annotations

ESI-MS/MS: Electrospray ionization tandem mass spectrometry; FDR: False discovery rate; GC-MS: Gas chromatography–mass spectrometry; MS/MS: Tandem mass spectrometry; MSI: Mass spectrometry imaging; NPs: Natural products; TCM: Traditional Chinese medicine.

are also increasingly becoming the core resources driving the digital and intelligent transformation of TCM (Figure 2)[60]. At the data level, MS resources provide large-scale, well-annotated spectral data and structural information, serving as the foundation for training and evaluating machine learning or deep learning models. At the application level, database searches and matching enable the retrieval of required spectral data or annotation of unknown spectra. Based on MS MN, it helps establish clustering relationships between compounds and achieve chemical space visualization (Figure 2). When integrated with other omics data, these resources support research into drug mechanisms of action and key target identification. At the knowledge level, fragmentation patterns deduced from massive datasets are applied in developing algorithmic tools for structural elucidation and activity prediction.

It is evident that the standardization and open sharing of MS data form the cornerstone for characterizing the structure and activity of compounds[60]. However, in the field of TCM research, even though the chemical composition of many single-ingredient or compound TCM preparations has been partially analyzed *via* MS, their MS data remain scattered across different laboratories or published articles. Dedicated databases for TCM are severely limited in both quality and scale, significantly hindering the progress of digital and intelligent analysis of TCM[61]. Consequently, constructing a high-quality, systematic MS database of the complete chemical composition of TCM has become a core task requiring urgent attention. In the process of database construction, it is necessary to introduce unified standards at the levels of data acquisition, annotation, and sharing. This includes the standardization of raw and processed MS data formats, systematic recording of key metadata such as compound sources, sample preparation conditions, MS acquisition parameters, and annotation confidence levels, as well as adherence to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles to enhance data discoverability and reusability[62].

MS-based metabolomics research

Metabolomics research involves a comprehensive analysis of metabolites within biological organisms, aligning well with the characteristic of TCM that emphasizes holistic regulation of human metabolic activities[63]. MS, with its high sensitivity and accuracy in identifying metabolites and endogenous compounds, has become a key tool in metabolomics[64–65]. Consequently, MS-based metabolomics has been extensively applied in TCM research, spanning multiple domains including TCM component

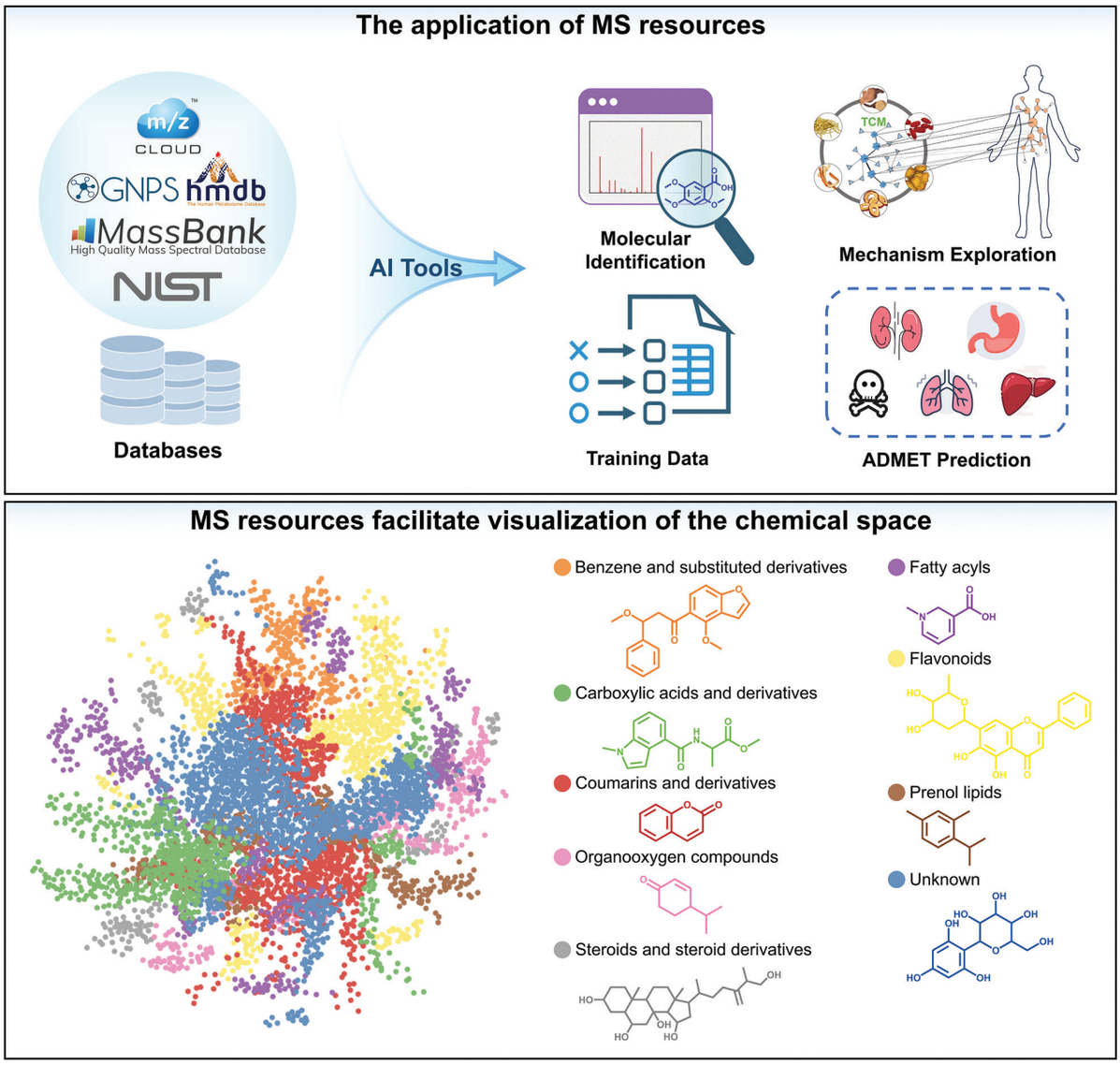

Figure 2. Application of mass spectrometry resources in TCM research. Top panel: MS databases integrated with AI tools enable molecular identification and activity prediction through training data analysis. Bottom panel: MS-based chemical space visualization reveals the distribution of major compound classes, highlighting structural diversity and potential for systematic discovery. AI: Artificial intelligence; MS: Mass spectrometry; TCM: Traditional Chinese medicine.

analysis, molecular target identification, and mechanism-of-action studies[63]. Owing to the high dimensionality, strong inter-variable correlations, and pronounced system-level characteristics of metabolomics data, computational approaches—particularly AI-based methods—have been increasingly introduced to address these challenges. A classic research approach integrates metabolomics with network pharmacology through a Dual-Track Strategy. As illustrated in Figure 3, Pathway 1 emphasizes top-down chemo-biological reasoning: First, MS data are used to analyze TCM components, combined with network pharmacology methods to infer potential active compounds. Pathway two employs a bottom-up systems biology approach: starting from animal models or cell experiments, MS-driven multi-omics technologies screen for differentially expressed molecules, thereby identifying key pathways and potential targets. Ultimately, the two pathways converge to achieve cross-validation between active components and their targets, thereby establishing a component-target-pathway-phenotype network and forming a closed-loop framework for elucidating the pharmacological mechanisms of TCM.

Several studies have adopted the aforementioned dual-track strategy, where network pharmacology provides upstream target predictions while metabolomics reveals downstream metabolic phenotypes. For instance, based on this approach, Zhang et al.[66] elucidated the mechanism by which Xiaoyao San alleviates depression-related inflammatory responses in CUMS mice; Shu et al.[67] systematically investigated the mechanism by which Tongmai Yangxin Pill alleviates doxorubicin-induced cardiac toxicity; Pan et al.[68] combined targeted energy metabolomics with network pharmacology to capture energy metabolism dysregulation associated with non-alcoholic fatty liver disease (NAFLD).

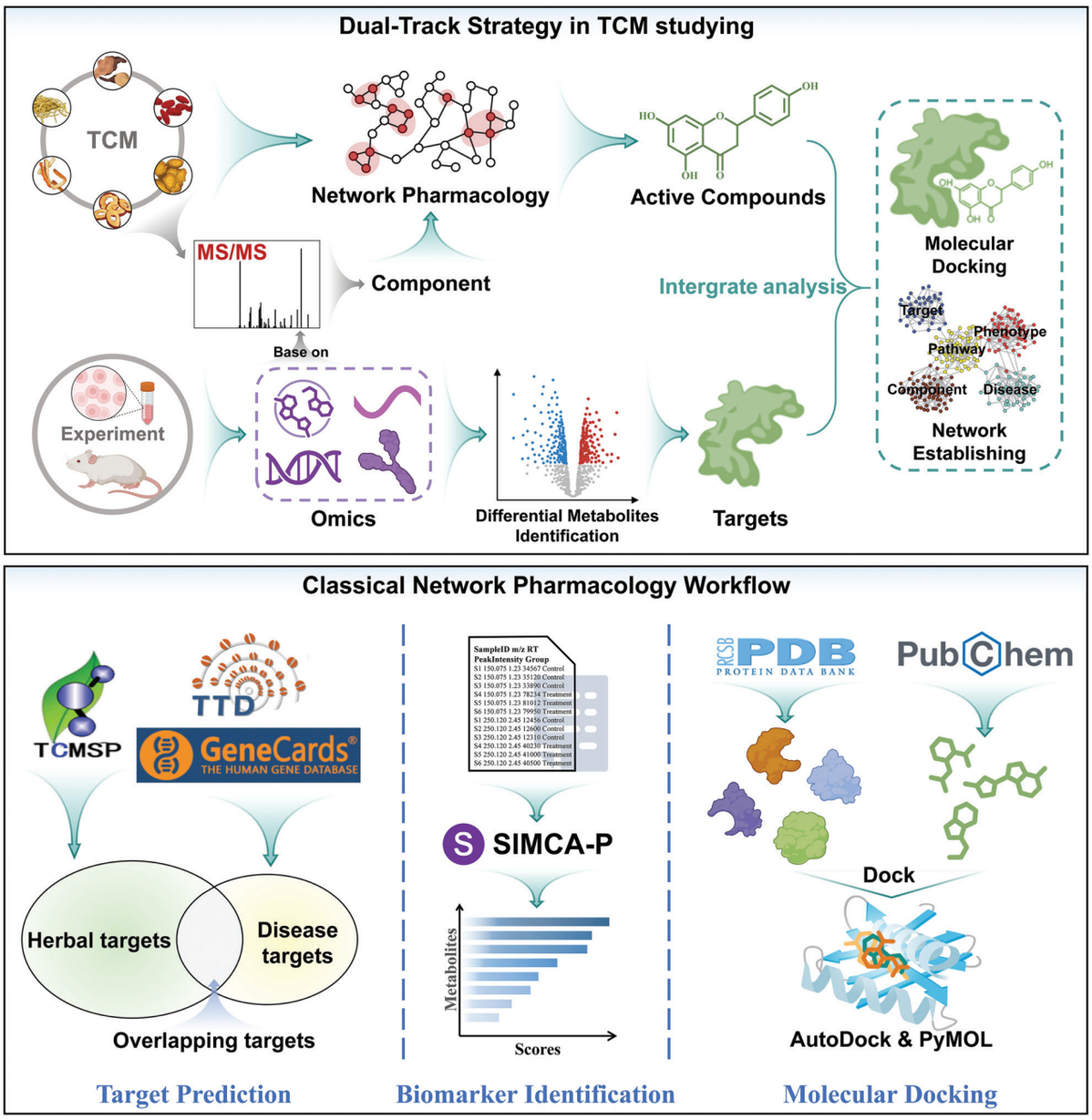

Figure 3. Classical framework and methodologies for elucidating the mechanisms of TCM. The mechanisms integrate network pharmacology (top panel) with multi-omics data to identify active compounds and their putative targets, and are further complemented by target prediction, biomarker identification, and molecular docking to enable comprehensive mechanistic validation. TCM: Traditional Chinese medicine.

Concurrently, to overcome the limitations of single-omics approaches, the academic community increasingly emphasizes multi-omics integration methods. For instance, Zhang et al.[69] integrated network pharmacology, traditional metabolomics, and transcriptomics to investigate the mechanism by which Coptis-Monascus formula improves NAFLD. Traditional metabolomics can only provide holistic results and struggles to reveal spatiotemporal and intercellular metabolic heterogeneity, leading to the emergence of spatial metabolomics and single-cell metabolomics[63]. Fan et al.[70] combined MSI-based spatial metabolomics with network pharmacology. On one hand, they employed AFADESI-MSI technology to perform high-spatial-resolution imaging of brain tissue in Alzheimer disease (AD) model mice, visualizing metabolite distribution (e.g., in the hippocampus and cortex) and identifying 28 AD-associated biomarkers along with their involved pathways. On the other hand, they predicted the active components and their targets in the ginseng-schisandra (RS) drug pair, constructing a "component-target-disease" network to enrich key pathways. By integrating these two pathways, we ultimately revealed the pharmacologically active substances and mechanisms by which RS alleviates AD through multi-component synergistic regulation of multiple pathways. Currently, the application of single-cell metabolomics in TCM efficacy research remains limited, with most studies still concentrated in pharmaceutical drug development to elucidate differential drug effects across cell types[71–72].

However, significant challenges persist across various stages of the dual-track strategy—particularly in target

Table 2

Application patterns of AI in the dual-track strategy

Application	Tool	Core methodology	Key features
Compound annotation	Details in Part 1 (e.g., MetFrag, MS-FINDER, CFM-ID)		
Target prediction	DrugBAN[79]	BAN	Prediction of drug–target interaction
	DeepDTAGen[80]	FetterGrad	Binding affinity prediction of known drug–target pairs and target-specific drug generation.
	DrugMAN[81]	Transformer & GAT	GAT and MAN for heterogeneous data integration and drug–target interaction modeling.
	IIFDTI[82]	Transformer & GAT & CNN	Extracting and fusing the independent and interactive features of drugs and targets.
Molecular docking	GNINA[83]	CNN	To enhance prediction accuracy by utilizing a CNN as the scoring function while inheriting the sampling algorithm from AutoDock Vina.
	Interformer[75]	Graph-Transformer & MDN	Explicitly modeling non-covalent interactions between protein and ligand atoms.
	Umol[76]	EvoFormer	Predicting all-atom protein–ligand complex structures end-to-end, solely from protein sequences and ligand SMILES.
Biomarker identification	DeepMSProfiler[77]	CNN	Directly processing raw LC-MS data to distinguish metabolic profiles of different diseases and reveal key metabolic signals.
	OmicLearn[78]	Scikit-learn & XGBoost	Machine learning–based interactive analysis platform
	CRANK-MS[84]	NN & SHAP	Eliminates the need for prior feature selection; primarily addresses Parkinson disease while being theoretically applicable to other diseases.

prediction, molecular docking, and biomarker validation—where conventional database- and rule-based methods remain dominant. As shown in Figure 3, the vast majority of current studies employ similar workflows and tools. For target prediction, databases like SwissTargetPrediction and Genecards are commonly used to identify potential targets for compounds and diseases, respectively, with their intersection serving as potential interaction targets. In molecular docking, 3D structures of target proteins and active compounds are typically obtained from PDB and PubChem databases. AutoDock Vina is then employed for molecular docking, with PyMOL used to visualize interactions. For biomarker validation, SIMCA-P performs PCA and OPLS-DA analysis on MS/MS data, screening differential metabolites based on variable importance in projection (VIP) > 1.5 and $P < 0.05$. However, these mainstream approaches suffer from limitations including heavy database dependency and high false-positive rates. Currently, the rapid advancement of AI technologies is systematically optimizing and reshaping this research paradigm by introducing more powerful computational models. For instance, in target prediction, FRoGS[73] employs deep learning techniques to project gene signatures into biological functional spaces, significantly enhancing the accuracy and sensitivity of compound-target predictions; Lu et al.[74] integrated drug–target interaction (DTI), binding affinity (DTA), and mechanism of action (MoA) prediction into an end-to-end framework called DTIAM, which performs exceptionally well in cold-start scenarios. In molecular docking, Interformer[75] is a Graph-Transformer–based interaction-aware model that precisely captures non-covalent interactions between protein and ligand atoms, thereby enhancing docking pose accuracy. Umol[76] predicts all-atom 3D structures of protein–ligand complexes directly from protein sequences and ligand SMILES information, without relying on known protein structures or predefined binding pockets. For biomarker identification, DeepMSProfiler[77] integrates CNNs and ensemble learning strategies to directly classify diseases from raw LC-MS data and extract disease-specific metabolic profiles. OmicLearn[78] is a web-based machine learning platform specifically designed for biomarker discovery from proteomics and other omics data. Additionally, other representative algorithms and their features are summarized in Table 2.

In summary, the dual-track strategy provides a robust framework for systematically elucidating the mechanisms of action of TCM. However, this strategy and its technical workflow still present certain limitations. First, the reliability of network pharmacology predictions heavily depends on the completeness and accuracy of the databases employed[63]. Second, technologies such as spatial omics and single-cell omics face challenges, including high technical barriers, complex data interpretation, and relatively late adoption in TCM research[64–65]. Finally, despite demonstrating significant potential, AI models face major obstacles to integration into TCM research paradigms due to interpretability issues stemming from their "black box" nature and the scarcity of high-quality training data[85]. Future research must address these challenges to ultimately achieve precise and comprehensive elucidation of TCM mechanisms of action.

Fragment-based structural characterization: from chemical language to biological function

In recent years, fragment-based drug discovery (FBDD) has garnered significant attention from the academic community. Unlike traditional high-throughput screening, which directly screens vast libraries of drug-like molecules, FBDD starts with a fragment library composed of low-molecular-weight (typically <300 Da) and structurally simple "chemical fragments" to identify lead compounds[86].

Table 3

Applications of fragment-based structural representation in AI-driven biological function prediction

Application	Methodology	Reference
DTI Prediction	Fragmentation: Computational (e.g., BRICS) or MS-based fragmentation AI Model: GNNs or Attention-based models	Huang et al.[97]
ADMET Properties Prediction	Fragmentation: Chemically-aware fragmentation AI model: Multi-task transformers or DNNs	Xie et al.[99]
Drug Synergy & Combination Therapy	Fragmentation: Dual representation of drug pairs as sets of fragments AI model: Siamese networks or dual-encoder architectures	Liu et al.[104]
De Novo Molecular Design	Fragmentation: Use of fragment vocabularies (e.g., from t-SMILES) as building blocks AI model: Generative models (e.g., VAEs, GANs) combined with RL	Wu et al.[105]
Bioactivity & Functional Class Prediction	Fragmentation: Generation of fragment-based fingerprints (e.g., ECFP) or learned fragment embeddings AI model: Classic ML (e.g., SVM, RF) or DNNs	Li et al.[106]

The methods used in each field of application of fragment-based structural representation are highlighted.

FBDD is highly favored for its exceptional screening efficiency, higher success rate, and the fact that the lead compounds obtained typically exhibit more desirable drug metabolism and pharmacokinetics (DMPK) properties and lower toxicity[87–88]. However, the profound impact of this approach extends beyond drug design, offering a new paradigm for molecular characterization and activity prediction in the era of AI[89]. In this paradigm, molecular fragments are no longer merely regarded as "structural units" but are elevated to "knowledge units" that carry rich biological activity information.

To convert complex molecular structures into information-rich units that can be learned by AI models, the academic community has developed various molecular fragmentation strategies, which can be broadly categorized into two main types. The first type is fragmentation based on organic synthesis rules. These methods generate fragment libraries suitable for virtual screening and combinatorial library design, essentially following the chemical synthesis experience and logic summarized by humans[90]. Classic algorithms such as BRICS[91] and RECAP[92] decompose molecules into common synthetic building blocks or structural skeletons by breaking predefined, chemically feasible bonds (e.g., amide bonds, ester bonds, etc.)[93]. The second category is an emerging strategy that is more closely aligned with the intrinsic physical-chemical nature of molecules—fragmentation based on MS cleavage rules. Unlike methods that follow synthetic logic, this approach is guided by the physical-chemical principles governing molecular behavior under energy excitation in a mass spectrometer, with cleavage preferentially occurring at sites with lower bond energies or those capable of forming stable radicals/charged fragments[94]. This fragmentation pattern can be directly observed experimentally, such as identifying characteristic neutral losses corresponding to specific functional groups in tandem MS, and thus used for high-throughput reaction screening; it can also be simulated using computational models[95]. In recent years, with the development of machine learning, various *in-silico* fragmentation prediction tools have emerged, such as the MolDiscovery[96] algorithm, which constructs graphical models to predict molecular fragmentation sites and fragments, and uses probabilistic models to match experimental mass spectra, significantly improving the automation and accuracy of small-molecule structure identification.

After fragmenting molecules into information-rich representational units, researchers have been able to use AI models to learn complex mappings from chemical structures to biological functions, achieving progress in multiple applications of drug discovery. In DTI prediction, fragment-based strategies have demonstrated strong performance. For example, Huang et al.[97] developed a knowledge-inspired substructure mining algorithm that effectively captures deep semantic relationships between fragments through a sophisticated interactive modeling module, significantly improving the accuracy of DTI prediction. Researchers have further combined MS fragmentation data with AI. Wang et al.[98] developed an innovative "fragmentation analysis-multi-source data fusion" algorithm to dynamically associate MS fragments with inflammatory targets, successfully elucidating the synergistic anti-inflammatory mechanism of the TCM compound Zhishi Xiebai Guizhi Decoction (ZXG) and predicting the regulatory functions of its active fragments on inflammatory pathways. Additionally, fragment-based characterization is equally critical in predicting the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of drugs. For example, Xie et al.[99] proposed the CAFÉ-MPP model, which employs a fragment-based strategy informed by chemical knowledge and advanced adversarial contrastive learning for pretraining, demonstrating outstanding performance on multiple ADMET public benchmarks. Additionally, our team developed the MSformer-ADMET model based on MS-derived fragment pretraining, achieving state-of-the-art accuracy across numerous ADMET benchmarks. This model not only outperforms existing approaches but also offers enhanced interpretability[100–101]. Similarly, this method has been successfully applied to predicting drug–drug interactions (DDI)[102], and screening inhibitors targeting specific targets[103]. To more clearly demonstrate the application of fragment-based structural characterization in various fields, its core methods are summarized in Table 3. These applications fully demonstrate that fragment-based characterization has become a bridge connecting chemical structures and biological activities.

The "fragment-based structural characterization" strategy bridges the gap between fragment structures and biological activity by decomposing molecules into "knowledge units" that contain biological information. The paradigm of converting chemical structures into functional

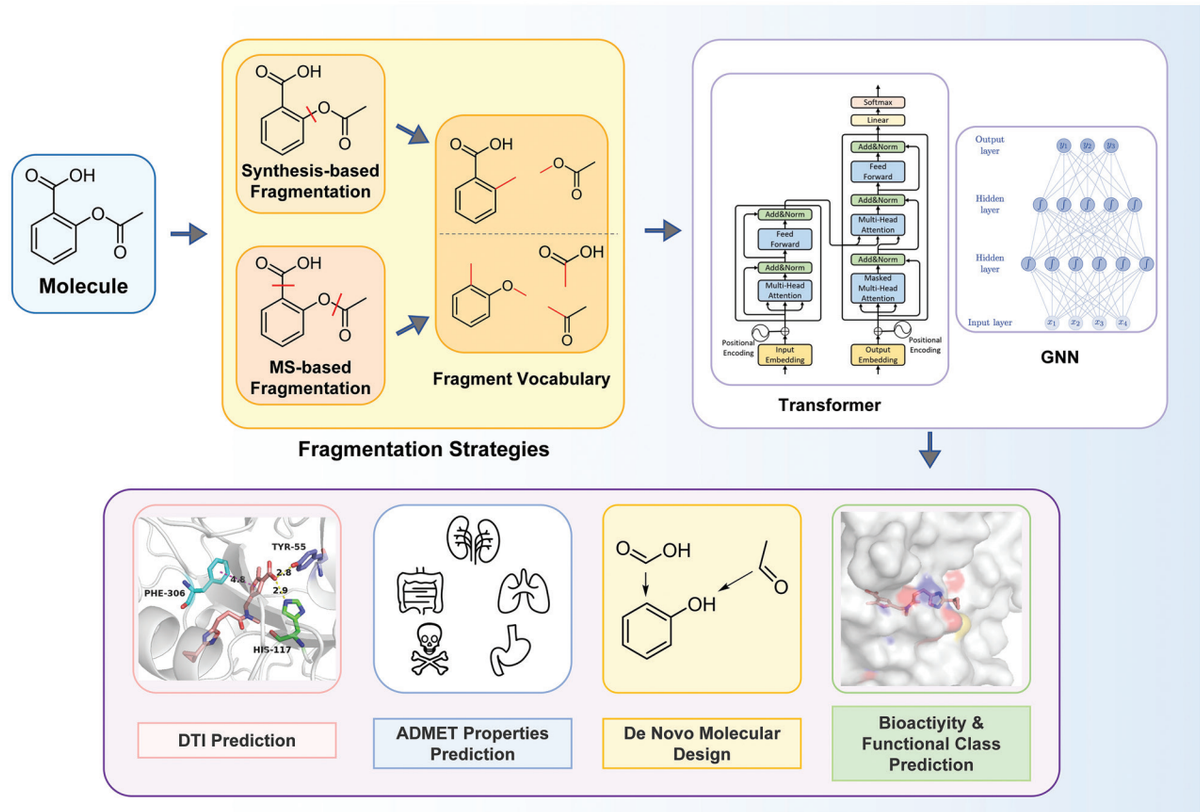

Figure 4. Workflow of fragment-based structural representation for biological function prediction. (1) The process begins with a molecule of interest. (2) The molecule is decomposed into a "vocabulary" of chemical fragments using two strategies: synthesis-based rules or principles derived from MS fragmentation. (3) This fragment vocabulary is then used to train an AI model. (4) Finally, the trained model can be applied to many drug discovery tasks. AI: Artificial intelligence; MS: Mass spectrometry.

predictions through fragment-based representation can be conceptualized as a workflow, visually summarized in Figure 4. The significance of this strategy lies in its provision of a "chemical language" that AI models can learn[105]. However, this field still faces challenges, primarily including: current methods heavily rely on predefined fragmentation rules, which are limited in representing complex polycyclic or chiral systems and may overlook novel chemical patterns[105,107]; additionally, most models exhibit bias in predicting biologically active compounds that are sensitive to stereochemistry[107]. Future research will focus on two key directions: first, developing adaptive fragmentation algorithms that can understand "chemical grammar" to achieve dynamic fragmentation decomposition[105]; second, introducing geometric deep learning algorithms (such as isomorph neural networks) to incorporate three-dimensional conformational information into the representation, thereby enhancing the model's ability to perceive molecular stereochemistry (Figure 4)[107].

Conclusion and perspective

The integration of AI with mass MS has catalyzed a profound paradigm shift in TCM research, transforming it from a largely descriptive discipline into a predictive and mechanistic science. This review has highlighted how AI-driven methodologies are revolutionizing every stage of the analytical workflow—from the structural elucidation of complex TCM components to the mapping of their mechanisms. The field has evolved from rudimentary spectral library matching to de novo structural elucidation and the visualization of metabolite distribution *via* MS imaging. Foundational to this remarkable progress are the increasingly standardized, large-scale MS resource libraries, which serve as the essential fuel for training and validating sophisticated AI models. This powerful integration has forged two transformative pathways for decoding TCM's inherent complexity. The first is the "dual-pathway pincer strategy" for mechanistic elucidation. The second, and perhaps more fundamental, is the emergence of fragment-based structural representation, which redefines molecular fragments—not merely as fragment structures—but as core "knowledge units" that encode rich information on chemical structure and biological function. This paradigm provides a learnable "chemical language" for AI, bridging the big gap between fragment structures and biological activity. To provide an overview of the current landscape and future directions, Figure 5 presents a conceptual framework.

Looking ahead, the future trajectory of this field hinges on our ability to address foundational data limitations and to advance the sophistication, transparency, and biological fidelity of AI models[14]. The paramount challenge remains the scarcity of a high-quality, centralized, and openly accessible MS database dedicated to the complete chemical profiles of TCMs[16]. Such a resource is not merely desirable but essential to overcome the "dark matter" problem and to provide the ground truth needed for model development. Concurrently, research must push beyond

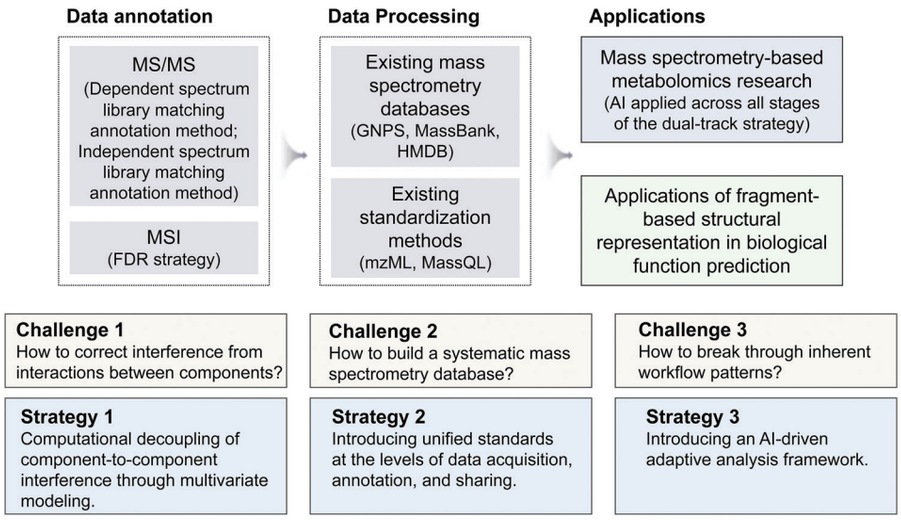

Figure 5. Conceptual framework for AI-driven mass spectrometry in TCM research. The framework outlining the integration of AI with mass spectrometry in TCM research, highlighting key data annotation and processing steps, major applications, and strategic responses to three core challenges: component interference, database standardization, and workflow innovation. AI: Artificial intelligence; FDR: False discovery rate; MS: Mass spectrometry; MSI: Mass spectrometry imaging; TCM: Traditional Chinese medicine.

current algorithmic boundaries. This includes developing more transparent and Explainable AI (XAI) frameworks to demystify the "black box" nature of deep learning predictions, thereby fostering trust and enabling hypothesis generation[85]. It also demands the creation of adaptive fragmentation algorithms that can learn the intrinsic "chemical grammar" of molecules, moving beyond static, rule-based cleavage to dynamic, context-aware decomposition[105]. Furthermore, integrating geometric deep learning to incorporate crucial three-dimensional conformational and stereochemical information will be vital for accurately predicting bioactivity, as many TCM targets are highly sensitive to molecular shape[107]. These advancements will not occur in isolation. The most exciting opportunities lie at the intersection of disciplines: merging AI/MS with single-cell and spatial omics to reveal the spatial heterogeneity of TCM component action at the single-cell level[108], leveraging LLMs in MS data interpretation and TCM knowledge graph construction[11]. By embracing these interdisciplinary synergies and tackling the core challenges of data and model interpretability, we can unlock a new generation of tools that are not only more accurate but also deeply insightful, ultimately fulfilling the immense, yet still latent, potential of TCM.

Conflict of interest statement

The authors declare no conflict of interest.

Funding

This work was supported by Zhejiang Provincial Natural Science Foundation of China (LD25H280002, J.L.), the National Natural Science Foundation of China (Grant No. U23A20513, X.F.), and the Key Project of Zhejiang Provincial Administration of Traditional Chinese Medicine (Grant No. GZY-KJS-ZJ-2025-071, J.L.), the Starlit South Lake Leading Elite Program (Grant No. 2023A303005, X.F.), and the Fundamental Research Funds for the Central Universities (226-2025-00009, X.F.).

Author contributions

Tianyi Ma and Zhaoxin Xu contributed to writing the review and editing and visualization. Tianyi Ma, Zhaoxin Xu, and Yugang Lin contributed to the literature collection. Jie Liao and Xiaohui Fan contributed to supervision. All authors reviewed and approved the published version of the manuscript.

Ethical approval of studies and informed consent

Not applicable.

Acknowledgments

None.

Data availability

All data can be available from the corresponding author with a reasonable request.

Declaration of generative AI in scientific writing

The conception and writing of this article were carried out collaboratively by all authors. Artificial intelligence was utilized exclusively for refining sentence coherence and amending punctuation. All citations in this manuscript were manually searched for and incorporated. After utilizing the aforementioned tools/services, the authors thoroughly reviewed and revised the content and take full responsibility for the entirety of this publication.

References

[1] Zhang B, Yang S, Guo D-a. The quest for the modernization and internationalization of traditional Chinese medicine. *Engineering* 2019;5(1):1–2.
[2] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod* 2020;83(3):770–803.
[3] Zhu B, Li Z, Jin Z, et al. Knowledge-based in silico fragmentation and annotation of mass spectra for natural products with MassKG. *Comput Struct Biotechnol J* 2024;23:3327–3341.

[4] Yang X-X, Gu W, Liang L, et al. Screening for the bioactive constituents of traditional Chinese medicines—progress and challenges. *RSC Adv* 2017;7(6):3089–3100.
[5] Wang X, Zhang A, Yan G, et al. UHPLC-MS for the analytical characterization of traditional Chinese medicines. *TRAC Trends Anal Chem* 2014;63:180–187.
[6] Pang B, Zhu Y, Lu L, et al. The applications and features of liquid chromatography-mass spectrometry in the analysis of traditional Chinese medicine. *Evid Based Complement Alternat Med* 2016;2016:3837270.
[7] Wang R, Zhang L, Li X, et al. Single-cell sequencing and mass spectrometry imaging reveal the multicellular compartmentalization map of plant triterpenes: a ginsenoside in Panax ginseng example. *Plant Biotechnol J* 2025;23(10):4461–4476.
[8] Duan Y-y, P-r L, T-t H, et al. Application and development of intelligent medicine in traditional Chinese medicine. *Current Med Sci* 2021;41(6):1116–1122.
[9] Zhang J, Chen X, Huang L, et al. Traditional Chinese medicine + artificial intelligence: Wuzhen consensus. *Acupunct Herb Med* 2025;5(2):134–135.
[10] Chen Z, Wang H, Li C, et al. Large language models in traditional Chinese medicine: a systematic review. *Acupunct Herb Med* 2025;5(1):57–67.
[11] Xu Y, Ma Y, Xu W, et al. A large language model for deriving spectral embeddings for accurate compound identification in mass spectrometry. *Commun Chem* 2025;8(1):326.
[12] Bilbao A. The future of a myriad of accelerated biodiscoveries lies in AI-powered mass spectrometry and multiomics integration. *J Mass Spectrom* 2025;60(8):e5157.
[13] Xing S, Huan T. Radical fragment ions in collision-induced dissociation-based tandem mass spectrometry. *Anal Chim Acta* 2022;1200:339613.
[14] Alexandrov T. Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence. *Ann Rev Biomed Data Sci* 2020;3(1):61–87.
[15] Guo G, Papanicolaou M, Demarais NJ, et al. Automated annotation and visualisation of high-resolution spatial proteomic mass spectrometry imaging data using HIT-MAP. *Nat Commun* 2021;12(1):3241.
[16] Hu G, Qiu M. Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Nat Prod Rep* 2023;40(11):1735–1753.
[17] Kim S, Kato I, Zhang X. Comparative analysis of binary similarity measures for compound identification in mass spectrometry-based metabolomics. *Metabolites* 2022;12(8):694.
[18] Huber F, Ridder L, Verhoeven S, et al. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol* 2021;17(2):e1008724.
[19] Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA* 2012;109(26):E1743–E1752.
[20] Wang M, Jarmusch AK, Vargas F, et al. Mass spectrometry searches using MASST. *Nat Biotechnol* 2020;38(1):23–26.
[21] Huber F, van der Burg S, van der Hooft J J, et al. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminf* 2021;13(1):84.
[22] de Jonge N F, R LJJ, Chekmeneva E, et al. MS2Query: reliable and scalable MS(2) mass spectra-based analogue search. *Nat Commun* 2023;14(1):1752.
[23] Shen X, Wang R, Xiong X, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun* 2019;10(1):1516.
[24] Ji H, Xu Y, Lu H, et al. Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification. *Anal Chem* 2019;91(9):5629–5637.
[25] Wang X, Li C, Li Z, et al. A structure-guided molecular network strategy for global untargeted metabolomics data annotation. *Anal Chem* 2023;95(31):11603–11612.
[26] Kim S, Thiessen PA, Bolton EE, et al. PubChem substance and compound databases. *Nucleic Acids Res* 2016;44(D1):D1202–D1213.
[27] da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci U S A* 2015;112(41):12549–12550.
[28] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28(1):27–30.
[29] Allard PM, Péresse T, Bisson J, et al. Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication. *Anal Chem* 2016;88(6):3317–3323.
[30] Lai Z, Tsugawa H, Wohlgemuth G, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* 2018;15(1):53–56.
[31] Chen B, Li H, Huang R, et al. Deep learning prediction of electrospray ionization tandem mass spectra of chemically derived molecules. *Nat Commun* 2024;15(1):8396.
[32] Goldman S, Li J, Coley CW. Generating molecular fragmentation graphs with autoregressive neural networks. *Anal Chem* 2024;96(8):3419–3428.
[33] Wolf S, Schmidt S, Müller-Hannemann M, et al. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinf* 2010;11:148.
[34] Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 2015;11(1):98–110.
[35] Nowatzky Y, Russo FF, Lisec J, et al. FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nat Commun* 2025;16(1):2298.
[36] Litsa EE, Chenthamarakshan V, Das P, et al. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun Chem* 2023;6(1):132.
[37] Young A, Röst H, Wang B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell* 2024;6(4):404–416.
[38] Wei JN, Belanger D, Adams RP, et al. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Cent Sci* 2019;5(4):700–708.
[39] Hong Y, Li S, Welch CJ, et al. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* 2023;39(6):btad354.
[40] Dührkop K, Shen H, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* 2015;112(41):12580–12585.
[41] Hoffmann MA, Nothias LF, Ludwig M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol* 2022;40(3):411–421.
[42] Stravs MA, Dührkop K, Böcker S, et al. MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;19(7):865–870.
[43] Bergman HM, Lundin E, Andersson M, et al. Quantitative mass spectrometry imaging of small-molecule neurotransmitters in rat brain tissue sections using nanospray desorption electrospray ionization. *Analyst* 2016;141(12):3686–3695.
[44] Zhang Y, Buchberger A, Muthuvel G, et al. Expression and distribution of neuropeptides in the nervous system of the crab Carcinus maenas and their roles in environmental stress. *Proteomics* 2015;15(23–24):3969–3979.
[45] Zhang H, Lu KH, Ebbini M, et al. Mass spectrometry imaging for spatially resolved multi-omics molecular mapping. *NPJ Imaging* 2024;2(1):20.
[46] Palmer A, Phapale P, Chernyavsky I, et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nat Methods* 2017;14(1):57–60.
[47] Alexandrov T, Ovchinnikova K, Palmer A, et al. METASPACE: a community-populated knowledge base of spatial metabolomes in health and disease. *BioRxiv* 2019;539478.
[48] Wadie B, Stuart L, Rath CM, et al. METASPACE-ML: context-specific metabolite annotation for imaging mass spectrometry using machine learning. *Nat Commun* 2024;15(1):9110.
[49] Tortorella S, Tiberi P, Bowman AP, et al. LipostarMSI: comprehensive, vendor-neutral software for visualization, data analysis, and automated molecular identification in mass spectrometry imaging. *J Am Soc Mass Spectrom* 2020;31(1):155–163.
[50] Yang F, Liang Z, Zhao H, et al. Mass spectral database-based methodologies for the annotation and discovery of natural products. *Chin J Nat Med* 2025;23(4):410–420.
[51] Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 2016;34(8):828–837.
[52] Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 2010;45(7):703–714.
[53] Xue J, Guijas C, Benton HP, et al. METLIN MS(2) molecular standards database: a broad chemical and biological resource. *Nat Methods* 2020;17(10):953–954.
[54] Wishart DS, Guo A, Oler E, et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;50(D1):D622–D631.
[55] Lee S, Hwang S, Seo M, et al. BMDMS-NP: a comprehensive ESI-MS/MS spectral library of natural compounds. *Phytochem* 2020;177:112427.
[56] Pedrioli PG, Eng JK, Hubley R, et al. A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotechnol* 2004;22(11):1459–1466.

[57] Martens L, Chambers M, Sturm M, et al. mzML—a community standard for mass spectrometry data. *Mol Cell Proteomics* 2011;10(1):R110.000133.
[58] Bhamber RS, Jankevics A, Deutsch EW, et al. mzMLb: a future-proof raw mass spectrometry data format based on standards-compliant mzML and optimized for speed and storage requirements. *J Proteome Res* 2021;20(1):172–183.
[59] Damiani T, Jarmusch AK, Aron AT, et al. A universal language for finding mass spectrometry data patterns. *Nat Methods* 2025;22(6):1247–1254.
[60] Jarmusch SA, van der Hooft JJ, Dorrestein PC, et al. Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Nat Prod Rep* 2021;38(11):2066–2082.
[61] Kang KB, Jeong E, Son S, et al. Mass spectrometry data on specialized metabolome of medicinal plants used in East Asian traditional medicine. *Sci Data* 2022;9(1):528.
[62] Paulhe N, Canlet C, Damont A, et al. PeakForest: a multi-platform digital infrastructure for interoperable metabolite spectral data and metadata management. *Metabolomics* 2022;18(6):40.
[63] Zhai Y, Liu L, Zhang F, et al. Network pharmacology: a crucial approach in traditional Chinese medicine research. *Chin Med* 2025;20(1):8.
[64] Wu L, Wu Q, Du X, et al. Revolutionizing pharmacological research of traditional Chinese medicine with single-cell omics technologies. *Fitoterapia* 2025;186:106846.
[65] Hui T, Zhou J, Yao M, et al. Advances in spatial omics technologies. *Small Methods* 2025;9(5):e2401171.
[66] Zhang Y, Li XJ, Wang XR, et al. Integrating metabolomics and network pharmacology to explore the mechanism of Xiao-Yao-San in the treatment of inflammatory response in CUMS mice. *Pharmaceuticals (Basel)* 2023;16(11):1607.
[67] Shu L, Wang Y, Huang W, et al. Integrating metabolomics and network pharmacology to explore the mechanism of Tongmai Yangxin pills in ameliorating doxorubicin-induced cardiotoxicity. *ACS Omega* 2023;8(20):18128–18139.
[68] Pan M, Deng Y, Qiu Y, et al. Shenling Baizhu powder alleviates non-alcoholic fatty liver disease by modulating autophagy and energy metabolism in high-fat diet-induced rats. *Phytomedicine* 2024;130:155712.
[69] Zhang X, Zhang J, Zhou Z, et al. Integrated network pharmacology, metabolomics, and transcriptomics of Huanglian-Hongqu herb pair in non-alcoholic fatty liver disease. *J Ethnopharmacol* 2024;325:117828.
[70] Fan Y, Wang A, Liu Z, et al. Integrated spatial metabolomics and network pharmacology to explore the pharmacodynamic substances and mechanism of Radix ginseng-Schisandra chinensis Herb Couple on Alzheimer's disease. *Anal Bioanal Chem* 2024;416(19):4275–4288.
[71] Ali A, Abouleila Y, Shimizu Y, et al. Single-cell metabolomics by mass spectrometry: advances, challenges, and future applications. *TrAC Trend Anal Chem* 2019;120:115436.
[72] Minakshi P, Ghosh M, Kumar R, et al. Single-cell metabolomics: technology and applications. *Single-Cell Omics* 2019;1(1):319–353.
[73] Chen H, King FJ, Zhou B, et al. Drug target prediction through deep learning functional representation of gene signatures. *Nat Commun* 2024;15(1):1853.
[74] Lu Z, Song G, Zhu H, et al. DTIAM: a unified framework for predicting drug-target interactions, binding affinities and drug mechanisms. *Nat Commun* 2025;16(1):2548.
[75] Lai H, Wang L, Qian R, et al. Interformer: an interaction-aware model for protein-ligand docking and affinity prediction. *Nat Commun* 2024;15(1):10223.
[76] Bryant P, Kelkar A, Guljas A, et al. Structure prediction of protein-ligand complexes from sequence information with Umol. *Nat Commun* 2024;15(1):4536.
[77] 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(1):7136.
[78] Torun FM, Winter S V, Doll S, et al. Transparent exploration of machine learning for biomarker discovery from proteomics and omics data. *J Proteome Res* 2023;22(2):359–367.
[79] Bai P, Miljković F, John B, et al. Interpretable bilinear attention network with domain adaptation improves drug–target prediction. *Nat Mach Intell* 2023;5(2):126–136.
[80] Shah PM, Zhu H, Lu Z, et al. DeepDTAGen: a multitask deep learning framework for drug-target affinity prediction and target-aware drugs generation. *Nat Commun* 2025;16(1):5021.
[81] Zhang Y, Wang Y, Wu C, et al. Drug-target interaction prediction by integrating heterogeneous information with mutual attention network. *BMC Bioinf* 2024;25(1):361.
[82] Cheng Z, Zhao Q, Li Y, et al. IIFDTI: predicting drug-target interactions through interactive and independent features based on attention mechanism. *Bioinformatics* 2022;38(17):4153–4161.
[83] McNutt AT, Francoeur P, Aggarwal R, et al. GNINA 1.0: molecular docking with deep learning. *J Cheminform* 2021;13(1):43.
[84] Zhang JD, Xue C, Kolachalama VB, et al. Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease. *ACS Cent Sci* 2023;9(5):1035–1045.
[85] Li Y, Liu X, Zhou J, et al. Artificial intelligence in traditional Chinese medicine: advances in multi-metabolite multi-target interaction modeling. *Front Pharmacol* 2025;16:1541509.
[86] Holvey RS, Erlanson DA, de Esch I JP, et al. Fragment-to-lead medicinal chemistry publications in 2023. *J Med Chem* 2025;68(2):986–1001.
[87] Xu W, Kang C. Fragment-based drug design: from then until now, and toward the future. *J Med Chem* 2025;68(5):5000–5004.
[88] Guo ZR. FBDD and drugs originated from FBDD. *Acta Pharm Sin* 2023;58(12):3490–3507.
[89] Zhang K, Yang X, Wang Y, et al. Artificial intelligence in drug development. *Nat Med* 2025;31(1):45–59.
[90] Yang Z, Shi S, Fu L, et al. Matched molecular pair analysis in drug discovery: methods and recent applications. *J Med Chem* 2023;66(7):4361–4377.
[91] Seo S, Lim J, Kim WY. Fragment-based molecular generative model with high generalization ability and synthetic accessibility. *arXiv preprint arXiv:2111.12907* 2021.
[92] Vásquez AF, Muñoz AR, Duitama J, et al. Non-extensive fragmentation of natural products and pharmacophore-based virtual screening as a practical approach to identify novel promising chemical scaffolds. *Front Chem* 2021;9:700802.
[93] Jinsong S, Qifeng J, Xing C, et al. Molecular fragmentation as a crucial step in the AI-based drug development pathway. *Commun Chem* 2024;7(1):20.
[94] Wang F, Liigand J, Tian S, et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;93(34):11692–11700.
[95] Hu M, Yang L, Twarog N, et al. Continuous collective analysis of chemical reactions. *Nature* 2024;636(8042):374–379.
[96] Cao L, Guler M, Tagirdzhanov A, et al. MolDiscovery: learning mass spectrometry fragmentation of small molecules. *Nat Commun* 2021;12(1):3718.
[97] Huang K, Xiao C, Glass LM, et al. MolTrans: molecular interaction transformer for drug-target interaction prediction. *Bioinformatics* 2021;37(6):830–836.
[98] Wang Q, Wang Q, Huang Q, et al. Five-layer-funnel filtering mode discovers effective components of Chinese medicine formulas: Zhishi-Xiebai-Guizhi decoction as a case study. *Phytomedicine* 2024;129:155678.
[99] Xie A, Zhang Z, Guan J, et al. Self-supervised learning with chemistry-aware fragmentation for effective molecular property prediction. *Brief Bioinform* 2023;24(5):bbad296.
[100] Liu H, Zhu B, Nie S, et al. Advancing ADMET prediction through multiscale fragment-aware pretraining with MSformer-ADMET. *Brief Bioinform* 2025;26(5):bbaf506.
[101] Zhu B, Liao J, Liu H, et al. MSformer: a meta-structure based interpretable framework for representation learning of natural products. *Anal Chem* 2025;97:25144.
[102] Lin S, Wang Y, Zhang L, et al. MDF-SA-DDI: predicting drug-drug interaction events based on multi-source drug fusion, multi-source feature fusion and transformer self-attention mechanism. *Brief Bioinform* 2022;23(1):bbab421.
[103] Lanka G, Banerjee S, Adhikari N, et al. Fragment-based discovery of new potential DNMT1 inhibitors integrating multiple pharmacophore modeling, 3D-QSAR, virtual screening, molecular docking, ADME, and molecular dynamics simulation approaches. *Mol Divers* 2025;29(1):117–137.
[104] Liu Y, Zhang P, Che C, et al. SDDSynergy: learning important molecular substructures for explainable anticancer drug synergy prediction. *J Chem Inf Model* 2024;64(24):9551–9562.
[105] Wu JN, Wang T, Chen Y, et al. t-SMILES: a fragment-based molecular representation framework for de novo ligand design. *Nat Commun* 2024;15(1):4993.
[106] Li M, Zeng M, Zhang H, et al. Biological activity predictions of ligands based on hybrid molecular fingerprinting and ensemble learning. *ACS Omega* 2023;8(6):5561–5570.
[107] Chen B, Pan Z, Mou M, et al. Is fragment-based graph a better graph-based molecular representation for drug design? A comparison study of graph-based models. *Comput Biol Med* 2024;169:107811.
[108] Chen YJ, Zeng HS, Jin HL, et al. Applications of mass spectrometry imaging in botanical research. *Adv Biotechnol (Singap)* 2024;2(1):6.