Mini review

Strategies for structure elucidation of small molecules based on LC–MS/MS data from complex biological samples

Zhitao Tian a,b, Fangzhou Liu c, Dongqin Li a, Alisdair R. Fernie d, Wei Chen a,b,*

a National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China
b Hubei Hongshan Laboratory, Wuhan 430070, China
c College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China
d Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm 14476, Germany

ABSTRACT

LC–MS/MS is a major analytical platform for metabolomics, which has become a recent hotspot in the research fields of life and environmental sciences. By contrast, structure elucidation of small molecules based on LC–MS/MS data remains a major challenge in the chemical and biological interpretation of untargeted metabolomics datasets. In recent years, several strategies for structure elucidation using LC–MS/MS data from complex biological samples have been proposed, these strategies can be simply categorized into two types, one based on structure annotation of mass spectra and for the other on retention time prediction. These strategies have helped many scientists conduct research in metabolite-related fields and are indispensable for the development of future tools. Here, we summarized the characteristics of the current tools and strategies for structure elucidation of small molecules based on LC–MS/MS data, and further discussed the directions and perspectives to improve the power of the tools or strategies for structure elucidation.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

1. Introduction ................................................................. 5086
2. Acquisition and curation of LC–MS/MS data from complex biological samples ........................................ 5086
3. Strategies for structure annotation of mass spectra. ........................................................................ 5086
   3.1. Structure elucidation based on authentic standard compounds ............................................. 5087
   3.2. Structure elucidation based on public/commercial reference spectral libraries ....................... 5087
   3.3. Structure elucidation based on an in-silico approach ............................................................ 5088
      3.3.1. Generation of in-silico spectral or fragmental libraries ................................................. 5088
      3.3.2. Molecular fingerprints prediction for ESI-based spectra ............................................... 5089
      3.3.3. Network-based strategies in structure elucidation for spectrum ..................................... 5089
      3.3.4. Structure elucidation for mass spectra with generative methods .................................... 5090
      3.3.5. Other in-silico methods assisting the structure elucidation .............................................. 5090
4. Retention time prediction methods assisting structure elucidation based on LC–MS/MS data .................. 5090
5. Fusion tools for metabolite identification based on LC–MS/MS data .................................................. 5092
6. Conclusions and perspectives ................................................................................................. 5092
CRediT authorship contribution statement ................................................................. 5093
Declaration of Competing Interest .................................................................................. 5093
Acknowledgments ........................................................................................................... 5093
References ....................................................................................................................... 5093
1. Introduction

The metabolome, first introduced in 1998, refers to the complement of small molecules in biological samples [1]. Small molecules have been studied extensively, as many of them have special biological significance for cell biology, physiology, and medicine [2]. For plants, small molecules can act as defense compounds (glucosinolates in Brassicaceae, gossypol in cotton, etc.), and plant developmental and growth regulators [3]. In addition, small molecules can also function as signaling molecules, immune modulators, endogenous toxins, and environmental sensors [4]. As the chemicals and physical properties of small molecules are very diverse, the detection and identification of small molecules accordingly becomes a major bottleneck for metabolomics [5–7].

Liquid chromatography–tandem mass spectrometry (LC–MS/MS) is one of the major analytical platforms used in the small molecule identification process [8]. Liquid chromatography separates mixtures with multiple components can be mainly divided into three parts: reversed-phase liquid chromatography (RPLC), hydrophilic interaction liquid chromatography (HILIC) and ion chromatography (IC) [9]. Tandem mass spectrometry provides mass spectral information that can be used for identifying separated components [10]. The mass spectrometry can be classified into in-time (ion traps, FTICR) and in-space (quadrupoles, TOFs) mass analyzers based on the principle of different platforms [11]. The mass spectrometry for LC–MS/MS commonly produces weak or absent in-source fragment ions (MS1) with electrospray ionization, compared to the mass spectrometry for GC–MS/MS using electron ionization [12]. The precursor ions are selected to generate MS/MS spectra in collision-induced dissociation (CID) or higher energy collisional dissociation (HCD) modes which can produce complementary fragments for further detection and structure annotation [10,13]. Compared to nuclear magnetic resonance (NMR) and gas chromatography–tandem mass spectrometry (GC–MS/MS), LC–MS/MS can produce more data and requires relatively simple extraction steps, which makes it more popular for exploring the metabolites in complex biological samples, especially for metabolomics [3,14]. For example, many features, defined as unique ions with MS1 and retention time information [15], can be routinely detected in biological samples, and follow-up studies were conducted to uncover the regulatory mechanism of some identified metabolites [16–24]. Regrettably, most features reported in these studies remains unknown due to the vast diversity of metabolites in biological samples, the lack of corresponding spectra in MS/MS spectral libraries, the disunity of collision energy with different platform, the existence of noise signal, the complexity of LC conditions optimization and the lack of large and diverse RT training sets [8,14]. For the past twenty years, many tools and methods have been developed for annotating the features from LC–MS/MS analysis [10,25]. These tools all involved in the metabolite identification are based on the information of mass spectra and retention time from LC–MS/MS [10,14,25]. In this review, we will discuss the characteristics of the various tools and strategies for metabolite identification based on LC–MS/MS data, and analyzes the ways to further improve the power of the tools or strategies for structure elucidation.

2. Acquisition and curation of LC–MS/MS data from complex biological samples

The features, with information of chromatographic peak (retention time) and mass spectral peak (m/z), can be detected by multiple software packages or frameworks such as MetAlign [26], OpenMS [27], MZmine [28], XCMS [29] and CAMERA [30]. In a biological sample, over ten thousand features can be detected, while there are a large number of artifactual peaks, chemical contaminants, and signal redundancies, which hamper the identification of “true” metabolites [31–35]. Those peaks are mainly arising from background, isotopes, adducts, homodimers, heterodimers and in source fragments [31–33,36], which can be identified based on retention time grouping, correlations between features, features clustering by retention time and calculating pairwise correlations, chromatographic peak-shape similarity, relative adduct frequency, isotope detection and specifying common adducts and neutral loss events [30,32,36–48]. The filtered features can be then identified via MS/MS spectra [36,38]. LC–MS/MS data from complex biological samples is produced by liquid chromatography coupled with tandem mass spectrometry in targeted/untargeted MS and-metabolomics [10]. Targeted metabolomics involves multiplexed analysis of a set of defined metabolites using multiple reaction monitoring (MRM) and is limited with metabolites coverage compared to untargeted metabolomics. In untargeted metabolomics, data-independent acquisition (DIA) and data-dependent acquisition (DDA) are two approaches to acquire MS/MS spectra [49]. In DDA acquisition workflows, the precursor ions exceed a predefined threshold of intensity or other predefined criteria are selected from a full scan analysis for further fragments obtaining [50]. DDA generally produces relatively good quality tandem mass spectra, which are useful to the subsequent structure elucidation [50]. In DIA acquisition workflows, precursor ions and fragments information are obtained from alternating scans acquired at either low or high collision energy in the collision chamber [51]. For conventional DIA, including All-Ion Fragmentation (AIF), MS\textsuperscript{ALL} and MS\textsuperscript{2} approaches, it is much difficulty to deduce the physical relationship between the precursor ions and their fragments for the wide scan range and the diversity of metabolites in biological samples [49,51]. To enhance selectivity, SWATH (sequential window acquisition of all theoretical fragment-ion spectra) narrows the precursor ion selection range to 20–50 Da consecutive isolation windows and gives a higher quality spectra [52]. SWATH can obtain a similar quantitative result with MRM, while deconvolution algorithms or tools are also indispensably to deduce the assignment of precursors to the corresponding fragment ions and improve the MS\textsuperscript{2} spectral quality, such as MS-Dial, Progenesis QI (Waters Co., Manchester, UK) and MasterView (AB Sciex, USA) [53,54].
of query mass spectra and only requires a molecular structure library, rather than reference spectral libraries. This strategy can provide a large number of annotations, however, the accuracy of the identification is relatively low to ‘Level 3’ annotations (tentative structure candidates or putatively characterized compound classes), or even equal to ‘Level 4’ annotations (formula determined) [8,25,56].

3.1. Structure elucidation based on authentic standard compounds

Authentic standard compounds can be used for targeted metabolomics, which focuses on several metabolites or a specific category of metabolites. As a comparison of retention times with references limits the range of candidates, a low-resolution LC–MS/MS analyzer is sufficient for targeted metabolomics. In addition, the metabolites in untargeted metabolomics can also be identified with unique 2D structures based on authentic standard compounds [57,58]. The key step of this strategy is to compare the query mass spectra and retention time to that of purified authentic standards, which can be obtained through purchases from chemical companies or isolation from complex biological samples or via enzyme-based synthesis from other purified metabolites by enzymes [25].

Purchasing authentic standard compounds is the most convenient way to conduct targeted metabolomics based on LC–MS/MS, while this strategy limits targeted metabolomics to common metabolites. For example, some amino acids, catecholamines, lipids and steroids were detected in urine samples in a targeted manner [59]. In addition to common metabolites, some species-characteristic metabolites were also identified and detected by this strategy. As an example, glucosinolates are distinctively present in nearly all members of the plant order Capparales, and are well studied as a model for research on secondary metabolism. The qualitative and quantitative analysis of glucosinolates is relatively easier than that of some other secondary metabolites [60]. In recent years following the development of synthesis and separation technology, some vendors can provide thousands of standards for researchers, such as IROA Technologies LLC (https://www.irotech.com/), Sigma–Aldrich (https://www.sigmaaldrich.cn/), and Agilent (https://www.agilent.com.cn/).

3.2. Structure elucidation based on public/commercial reference spectral libraries

The query mass spectra can be additionally identified by searching the public/commercial reference spectral libraries, which are built with authentic reference standards by institutions and companies around the world. Thanks to the great progress that mass spectrometry technology and chemical synthesis/isolation techniques have made in the past two decades, many mass spectral databases have been established and developed to extend millions of reference mass spectral data for diverse instruments and different collision energies, such as MassBank of North America (MoNA) (https://mona.fiehnlab.ucdavis.edu/), the Colm Metabolome database (https://gmd.mpimp-golm.mpg.de/), MassBank (https://massbank.eu/MassBank/), METLIN (https://metlin.scripps.edu/), mzCloud (https://www.mzcloud.org/), GNPS (https://gnps.ucsd.edu/), NIST 20 (https://www.nist.gov/) and Wiley Science Solutions (https://sciencesolutions.wiley.com/) [61]. In addition, some special spectral libraries were established for particular researches, such as ReSpect (https://spectra.psc.riken.jp/) for phytochemicals-related researches, HMDB (https://hmdb.ca/) for the human metabolomics.

MassBank is the first public repository of mass spectra of small chemical compounds for life sciences, and the mass spectra it houses are from different instruments and multiple contributors [62]. With the efforts that global chemists and computer scientists have made, many more sharing platforms have been established for research from various fields and have greatly promoted the development of metabolomics. MoNA is a centralized repository with 695,425 spectra, including 145,316 MS/MS spectra, with the spectra contained is mainly being contributed by MassBank, ReSpect, HMDB, GNPS, LipidBlast, Vaniya/Fiehn Natural Products Library and RIKEN PlaSMA. In addition to the free public spectral libraries, several commercial spectral libraries were established with well curated spectra and enriched contents. HMDB is a comprehensive database on small molecules from Homo sapiens and contains 64,923 experimental MS/MS spectra of 4,064 metabolites and 1,440,324 in-silico MS/MS spectra of 217,920 metabolites [63]. GNPS is a web-based mass spectrometry ecosystem, which also collects the MS/MS spectra from public spectral libraries, including Massbank, ReSpect, HMDB, CASMI [64]. METLIN is the largest library of mass spectra among all of the public and commercial spectral libraries, and hosts over 850,000 molecular standards with over 4,000,000 curated high-resolution tandem mass spectra [65]. NIST 20 is another commercial spectral library contributed by many researchers currently containing 1,320,389 spectra of 185,608 precursor ions from 30,999 chemical compounds. Among all of the spectra in the library, both High-Resolution, Accurate-Mass (HRAM) MS/MS (1,026,712 spectra for 27,840 chemical compounds) and Low-Resolution MS/MS (215,649 spectra for 28,559 chemical compounds) are represented. In addition to the small molecule, 90,244 spectra of 6,803 precursor ions from 1,504 peptides were also included in this library. mzCloud is specialized with high quality spectral trees of MSn spectra, which are generated with various collision energies. As each tree represents a molecule in this library, 19,515 molecules are contained with 2,310,148 positive mass spectra, and 7,875 molecules are contained with 850,580 negative mass spectra.

In addition to the size and quality of spectral libraries, the algorithms of the search systems employed affect the outcomes of the structure annotation of querying mass spectra against public/commercial reference spectral libraries. In contrast to EI-based spectra (produced by GC–MS), ESI-based spectra (produced by LC–MS/MS) tend to be less reproducible, especially for the cross-instrument comparisons, which raise higher requirements for the search system [66–68]. In this regard, the traditional mass spectral library search algorithms were firstly utilized for EI-based spectra, such as dot-product, Euclidean distance and probability-based matching (PBM) system. For dot-product and Euclidean distance, each mass spectrum can be considered as a point in a multidimensional hyperspace, with the axis presented by the mass (m/z) and the position of the point presented by the intensities of those masses [69]. Dot-product, as the name implies, calculates the dot product of two vectors, of which one is the vector from the coordinate origin to the point of a query mass spectrum and the other one is a library of mass spectrum. By analogy, Euclidean distance is the algorithm that calculates the Euclidean distance between the two points of the query mass spectrum and a library mass spectrum. Comparing the two algorithms mentioned above, the PBM system is relatively complex and cannot be described as an analytic function. However, the PBM system can to some extent avoid mismatch results when the metabolite with the query spectrum is not in the reference library [70]. Among those algorithms, dot-product derived algorithm has gained widespread use for ESI-based spectra [71,72]. To overcome the great differences of ESI-based spectra from a range of instruments, some sophisticated matching algorithms different from the routine algorithms were created [73–75]. In an independent approach, X-Rank is based on statistical relations between mass over charge values, ordered by intensities, rather than taking into account absolute or relative intensities, which makes it more effective in supporting cross-platform identi-
fication [75]. Recently, spectral entropy similarity has been developed in addition to the previous approaches, and has been proven to outperform all classic algorithms by adapting concepts from information theory [76].

3.3. Structure elucidation based on an in-silico approach

The range of metabolites in complex biological samples is daunting, with hundreds of thousands of metabolites representing the lower end of estimates, and this number is continually growing [2, 77–79]. PubChem (https://pubchem.ncbi.nlm.nih.gov/) [80] and ChemSpider (https://www.chemspider.com/) [81] are the world’s two largest collections of freely accessible chemical information. While the number of compounds contained in PubChem and ChemSpider has reached 111 million and 114 million, respectively, the number of metabolites with MS data collected in the reference library is considerably more limited, with the METLIN database, which is the largest experimental spectral library, covers a mere 850,000 metabolites, which represents less than 1% of the compounds found in PubChem or ChemSpider [65]. To overcome this difficulty, in-silico qualitative tools have been developed in the last decade [82–87]. Early tools were mainly developed by commercial companies, such as Mass Frontier by Thermo Fisher Scientific. Afterwards, many open-source mass spectrometry identification software programs appeared, and most of those tools could be divided into four categories (Fig. 1).

3.3.1. Generation of in-silico spectral or fragmental libraries

The earliest strategy for structure elucidation using MS/MS data based on an in-silico approach is to predict mass spectra or fragments from structures of chemical species subjected to a fragmentation process. Generally, a SMILES string or other molecular string is utilized for representing the structure of an individual chemical [88–91]. The fragmentation process can be distinguished as quantum chemistry, heuristic-based methods, and machine learning [8, 82].

Fragmentation methods based on quantum chemistry (QC) were first established for structure annotation of EI-based spectra [92–94]. Then, a quantum mechanical and molecular mechanical combined approach was built for the structure annotation of ESI-based spectra of large polypeptides [95]. Similarly, QCMS2 is another approach that combines QC with a heuristic approach,
and can also simulate and understand the mass spectra of peptides [96], QC-FPT takes the dominant fragment peaks and 3D candidate structures from a particular experiment and attempts to model the collision-induced dissociation [97]. In contrast to the QC-FPT, ChemFrag supports fragment ion annotations of an entire spectrum, rather than predicting the dominant mass spectrum [98]. Although fragmentation methods based on quantum chemistry have been developed for nearly twenty years, a large amount of computational power is required to implement this strategy, especially for large molecules, compared with tools based on other strategies, such as machine learning approaches and heuristic-based approaches [82].

Heuristic-based methods generate in-silico fragments relying on a collection of general heuristic rules of fragmentation. For this strategy, tools were developed for structure elucidation based on the match scores between in-silico fragments and experimental peaks. Among those tools, Mass Frontier (HighChem Ltd., Bratislava, Slovakia) was one of the earliest developed tools and involves cleaving the bond in the structure based on reactions described in the literature. In addition, MetFrag first generates all possible fragments of the candidate structures using simple bond-breaking rules and combinatorial fragmentation, and then the fragmentation trees were traversed by breadth-first search (BFS) [99]. Compared to MetFrag, MIDAS traversed the fragmentation trees by depth-first search (DFS), which was more memory-efficient than BFS [100]. MAGMA performed the fragmentation of the fragmentation trees were traversed by breadth-first search (BFS) [100]. Compared to MetFrag, MIDAS traversed the fragmentation trees by depth-first search (DFS), which was more memory-efficient than BFS [100]. MAGMA performed the fragmentation of a structure or a substructure by removing each heavy atom sequentially (i.e. non-hydrogen atoms), and can be used for structure annotation of MS6 data [101,102]. As hydrogens always rearrange during bond cleavage in low-energy CID, MS-FINDER applied nine hydrogen rearrangement (HR) rules to generate in-silico fragments [103]. DEREPLICATOR + produced in-silico fragments by disconnecting bridges and 2-cuts at N–C, O–C and C–C bonds, and could identify many variants within spectral networks [104]. It was recently demonstrated that DEREPLICATOR + improves the structure elucidation of general metabolites and natural products above that of the previous developed DEREPLICATOR, which was a Heuristic-based algorithm that generated fragments for pepticid natural products [104,105]. Heuristic-based methods are also utilized for structure elucidation for a group of substances. For example, the two heuristic-based tools, LipidBlast and Lipid-Match, are popular tools for lipid identification [106,107]. The in-silico spectra of LipidBlast contains 212,516 spectra covering 119,200 compounds from 26 lipid compound classes, were calculated with this strategy [108]. In addition, heuristic-based methods can be worth "adjusting" the relative abundances of fragment ions within in-silico generated tandem mass spectra to facilitate metabolite identification [109]. Thus, heuristic-based methods are considered to be very helpful strategies for the structure annotation of fragment ions in the field of biochemistry.

The advent of machine learning has expedited the structure annotation of experimental mass spectra. ISIS is one of the earliest tools based on machine learning to find accurate bond cleavage rates for collision-induced dissociation in an ESI-based spectrum, and targets the identification of lipids [110]. CFM-ID, arguably the most popular machine learning approach based on the Markov module, was published to identify multiple collision energy spectra [111]. The results showed that CFM-ID obtained substantially better rankings for the correct candidate than existing methods (MetFrag and FingerID) on tripeptide and other metabolite data, when querying PubChem [80] or KEGG [112] for candidate structures of similar mass [111]. A new version of CFM-ID (version 4.0) has been published recently, and it has been proven to be significantly more accurate than previous versions [113]. With the advance of artificial neural networks (ANNs), deep learning has been utilized for the structure annotation of a considerable amount of experimental mass spectra, such as NEIMS for EI-based spectra [114]. By contrast, there has not been any deep learning method for directly predicting ESI-based spectra of a given molecular. However, given the rapid development of public ESI-based spectral libraries, it should be possible to develop such tools in the future.

### 3.3.2. Molecular fingerprints prediction for ESI-based spectra

Predicting molecular fingerprints for ESI-based spectra is another strategy for metabolite identification based on in-silico methods. Commonly, the molecular fingerprint can be represented by a vector, in which each number represents the possibility for the presence of the molecular property. The predicted fingerprint of the unknown compound can be compared against the fingerprint of each candidate molecular structure to produce a similarity score, and then candidate structures are sorted according to the similarity scores. In the case of a molecular fingerprint, the prediction of the substructures or fingerprints for each query spectrum is the key step. For example, FingerID is the first molecular identification tool that predicts the molecular fingerprints for each query spectrum with a support vector machine (SVM) model trained by a large set of tandem mass spectra in MassBank [115]. In addition, CSI:FingerID performs molecular fingerprint prediction using multiple kernel learning with fragmentation trees inputted [116–118]. Instead of fragmentation trees, SIMPLE formulates a sparse interaction model for metabolite peaks to predict fingerprints and is lighter and more readily interpretable than CSI:FingerID [119]. The IOKR reverse model maps molecular structures into the MS/MS feature space and then solves a pre-image problem to find the molecule with the most similar fingerprint [120,121]. By contrast, ADAPTIVE is an IOKR-derivative tool that learns a model to generate fingerprints for metabolites [122]. As a result, all of these mentioned fingerprints are specific to both data and the task of metabolite identification and are therefore nonredundant [122]. In addition to the above mentioned SVM-based methods, SF-Matching achieved similar performance to CSI:FingerID with a random forest model used for fingerprint prediction [123]. Compared to common machine learning, deep learning is another strategy for predicting the fingerprints of query mass spectra. MetFID applied an artificial neural network with two hidden layers to predict a composite vector comprising of 528 binary entries [124]. McSearch applied core structure-based search (CSS) algorithm based on hypothetical neutral loss values to predict the core substructure of the query MS [125].

### 3.3.3. Network-based strategies in structure elucidation for spectrum

Network-based strategies could be not only used for predicting the functional activity or category of metabolites directly from spectral features [126,127], but also for metabolite identification per se [128,129]. iMet was the first metabolite identification tool based on spectral similarity and structure similarity for MS data [130]. Regarding this method, neighbor metabolites (with similar MS/MS spectra) share structural similarities, so the unknown metabolites could be identified according to the identification of the neighbor metabolites. This strategy can be easily integrated with other methods to improve the accuracy of metabolite identification. As an example, NAP integrated MetFrag into network-based strategy, which improved the ranking of the correct spectra from a mean ranking position of 14.7 to a mean ranking position of 4.7 [131]. This network-based strategy embraces the construction of two kinds of networks: one is a network usually based on spectral similarity, the other is a network usually based on structural similarity. Compared with NAP, MetDNA uses the reaction in KEGG instead of the fingerprint similarity to represent the structure similarity and cumulatively annotated approximately 2000 metabolites from one experiment [132]. DeepPeptide and MS2DeepScore both train deep learning models to predict structural similarity.
scores for spectral pairs instead of other spectral similarity scores, such as dot-product, and then construct a network for metabolite identification [133,134]. MS similarity scores can also be presented by the fingerprint similarities predicted from mass spectra. For example, Spec2Vec is a novel spectral similarity score inspired by Word2Vec, which is a natural language processing algorithm. Word2Vec learned fragmental relationships within a large set of spectral data to derive abstract spectral embeddings that can be used to assess spectral similarities [135]. The feature information (e.g., isotope patterns, adduct formation, chromatographic retention times, and fragmentation patterns) can also be used for aiding the construction of metabolite networks [44,136,137]. Compared to the above-mentioned methods, NetI.D connected MS peaks based on mass differences reflecting adduction, fragmentation, isotopes, or feasible biochemical transformations, and performed the global network optimization to produce an optimal and consistent network annotation by linear programming [138].

3.3.4. Structure elucidation for mass spectra with generative methods

The above-mentioned strategies for metabolite identification based on in-silico approach are seriously dependent on the structure libraries, and the identification results must be included in those libraries. For the diversity of chemical modification in complex biological individuals, especially for plants, there are still many metabolites not included in public libraries. Meanwhile, it is immensely time-consuming to predict the in-silico spectra or calculate all of the fingerprints for all metabolites. The direct approach is to translate the query high-resolution mass spectral peak to a representation (for instance, in SMILES) of the annotated metabolite rather than to predict the in-silico spectrum from the structure or intermediate fingerprints from the query spectrum, which requires another comparison process. In a similar approach MassGenie utilizes a transformer-based deep neural network coupled with VAE-Sim, a variational autoencoder (VAE)-based model, to directly predict the structure of a molecule from the query spectrum [139]. VAE-Sim is a variational autoencoder that is the backbone of MassGenie, which is used to generate ‘true’ molecules [140]. This strategy for direct prediction of a molecule from the query MS has been studied recently, proving it is convincing to provide valuable clues to expedite structure annotation of experimental mass spectra. However, in some ways, the accuracy of identification could be further improved. For instance, SMILES, a regular molecular representation, could be replaced by SELFIES [141] and DeepSMILES [142], which are more convenient methods for representing a valid molecule.

3.3.5. Other in-silico methods assisting the structure elucidation

Besides the above four classes of methods, some other machine learning approaches were developed to assist the metabolite identification. MS2LDA adopted Latent Dirichlet Allocation, an algorithm originally used for text mining, to extract the co-occurring molecular fragments and neutral losses [143]. With structure annotation for the extracted motifs by additional methods, the compounds can be identified with candidate structures or classified. MESSAR (MEtabolite Substructure Auto-Recommender) extracted spectral features of spectra and generated substructures of metabolites in the spectral library, then generated the association rules linking spectral features (with exact masses) with specific substructures based on the concept of association rule mining (ARM) [144]. With the association rules applied, the structure elucidation or classification of metabolites can be conducted automatically. It is a very complex and time-consuming process to identify metabolites with a unique structure, while it can be a solution to perform automatic classification of compounds into compound classes based on machine learning methods [145,146]. In particular, CANOPUS used a deep neural network to classify the metabolites based on mass spectrometry [147].

4. Retention time prediction methods assisting structure elucidation based on LC–MS/MS data

In addition to the MS information, chromatographic behavior, reflecting the physicochemical properties (molecular weight, hydrophobicity, polarity, molecular shape etc.) of metabolites, should also provide structural information. As an example, many isomeric compounds are indistinguishable in both MS+1 and tandem MS analyses, and must be resolved chromatographically if separate quantitation and identification is desired. Retention time prediction commonly are not utilized for structure elucidation separately, while the results of retention time prediction can be combined to screening the candidates and/or improve candidates rankings based on mass spectra information [148,149]. Empirically, the octanol/water partition coefficients (log P values) of each metabolite were believed to determine the chromatographic retention time, and the linear solvation energy relationships (LSERs) equation was utilized for predicting retention data in early researches [150,151]. Such technique, relating the variations in one response variables (chromatographic retention time) to the variations of several descriptors, is called Quantitative structure–retention relationships (QSRRs) [151,152]. The chromatographic retention time always varies considerably depending on the experimental conditions such as column packing, flow rate, elution gradient, and pH of the mobile phase. Afterwards, much more complex models and larger training sets were used for the prediction of LC retention time (Table 1). Among those methods, Molecular descriptors (MDs) are the most common variables for LC retention time prediction, as MDs encode structural information and chemical information, such as the type of atoms and bonds, number of rings, charge, and stereochemical configuration, through mathematical and statistical approaches [153]. MDs have been widely utilized for training common machine learning models, such as multiple linear regression (MLR), random forest (RF) regression, support vector machine (SVM), and partial least squares (PLS) regression. With the advance of artificial neural networks (ANNs), it has been proven that deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs) and graph neural networks (GNNs) models (neural network types of ANNs) show robust performance for RT prediction in both RP and HILIC chromatography, compared to XGBOOST, BRNN, RF and LIGHTGBM [154–159]. For the METLIN small molecule retention time (SMRT) dataset, the artificial neural networks (ANNs) had lower mean absolute error (MAE, MAE = ~0.5 min) than the traditional machine learning model (MAE = ~1 min) [156,160]. While the training datasets of thousand metabolites were still not sufficient for the ANNs models, and with more MDs and more complex models considered, a larger database was needed to overcome the overfitting problem [154]. As the chromatographic retention time always varies considerably depending on the experimental conditions, it is a priority to make community sharing of RT information possible across laboratories and chromatographic systems [161,162]. To obtain more databases, an approach directly mapping RTs between different systems was developed, and a sufficient number of common compounds were needed in both systems for this approach [161]. In addition to directly mapping RTs, a retention order prediction model can also be trained using retention time measurements from different LC systems and configurations, and this can be an effective way to learn the retention behavior of molecules from heterogeneous retention time data [148]. In addition to the large databases, transfer learning in combination with self-supervised
| Publication                        | Year | LC type | Model type                      | Size of training data | Molecular type | Variables                                             |
|-----------------------------------|------|----------|---------------------------------|-----------------------|----------------|--------------------------------------------------------|
| Hagiwara et al. [175]             | 2010 | RP-LC    | SVR and MLR                     | 150 authentic compounds |                | 9 MDs                                                  |
| Creek et al. [176]                | 2011 | HILIC    | MLR                             | 120 authentic compounds |                | 6 MDs                                                  |
| D'Archivio, Maggi and Ruggeri [177] | 2014 | RP-LC    | MLR and PLS regression         | 47 authentic compounds | butyl esters of 47 acylcamitines | 73 MDs       |
| Kouskoura, Hadjipavlou-Litina and Markopoulou [178] | 2014 | RP-LC    | PLS regression                   | 100 authentic compounds |                | 66 MDs                                                  |
| D'Archivio et al. [179]           | 2014 | RP-LC    | DNNs                            | 24 authentic compounds | s-triazines    | 5 MDs                                                  |
| Cao et al. [180]                  | 2015 | HILIC    | MLR and RF                      | 93 authentic compounds |                | 346 MDs                                                 |
| Aicheler et al. [181]             | 2015 | RP-LC    | SVR                             | 201 authentic compounds | lipid          | 11 MDs                                                  |
| Munro et al. [182]                | 2015 | RP-LC    | DNNs                            | 166 authentic compounds | drugs          | 17 MDs                                                  |
| Falchi et al. [183]               | 2016 | RP-LC    | Four combined (fingerprints + ordinary) KPLS models | 1383 authentic compounds |                | molecular and fingerprints descriptors (CN) and (DB) number |
| Ovcacevka et al. [184]            | 2016 | RP-LC    | The second degree polynomial regression | 400 authentic compounds | lipid          | 28 MDs                                                  |
| Alazadeh et al. [185]             | 2016 | RP-LC    | MLR, DNNs, and SVM               | 528 and 298 compounds for positive and negative electrospray ionization mode respectively | 6 MDs | 5 MDs                                                  |
| Wolfer et al. [186]               | 2016 | RP-LC    | Combination of RF and SVR models | 442 authentic compounds |                | 97 MDs                                                  |
| Bubnik and Wiczing [187]          | 2016 | RP-LC    | Lasso, Stepwise and PLS regressions | 115 authentic compounds | drugs          | 50 MDs                                                  |
| Barron and McEneff [188]          | 2016 | RP-LC    | DNNs                             | 1,117 authentic compounds |                | 16 MDs                                                  |
| Randazzo et al. [189]             | 2016 | RP-LC    | PLS regression                   | 91 authentic compounds | steroids       | 97 MDs                                                  |
| Taraji et al. [190]               | 2017 | HILIC    | PLS regression                   | 16 authentic compounds | β-adrenergic agonists and related compounds | 321 MDs |
| Taraji et al. [191]               | 2017 | HILIC    | PLS regression                   | 98 authentic compounds | pharmaceutical compounds | 321 MDs |
| Zhang et al. [192]                | 2017 | RP-LC    | MLR                              | 24 authentic compounds | 16-membered ring macrolides | 8 MDs |
| Park et al. [193]                 | 2017 | RP-LC    | MLR                              | 41 authentic compounds |                | 10 MDs                                                  |
| Wen et al. [194]                  | 2018 | RP-LC    | PLS regression                   | 148 authentic compounds | drugs          | 126 MDs                                                 |
| Wen et al. [195]                  | 2018 | RP-LC    | PLS regression                   | 191 authentic compounds |                | 128 MDs                                                 |
| McEachran et al. [196]            | 2018 | RP-LC    | PLS regression                   | 97 authentic compounds |                | 7 MDs                                                   |
| Hall et al. [197]                 | 2018 | RP-LC    | DNNs                             | 1,955 authentic compounds |                | 47 MDs                                                  |
| Bousmeester, Martens and Degroeve [198] | 2019 | RP-LC    | Bayesian Ridge Regression (BRRE), Least Absolute Shrinkage and Selection Operator (LASSO), DNNs, Adaptive Boosting (AB), Gradient Boosting (GB), RF and SVR | 6,759 authentic compounds |                | 151 MDs                                                 |
| Bonini et al. [154]               | 2020 | HILIC & HILIC (33) & (3) | XGBoost, Bayesian-regularized Neural Network (BRNN), RF, Light Gradient-Boosting Machine (LightGBM), DNNs | 1,023 (HILIC) & 494 (RP-LC) authentic compounds |                | 286 MDs                                                  |
| Ju et al. [163]                   | 2021 | HILIC & RP-LC | DNNs + TL                 | 77,898 authentic compounds (DNNs), and 17 data sets (Transfer Learning) |                | 1,470 MDs                                                |
| Osipenko et al. [159]             | 2021 | HILIC & RP-LC | RNNs + TL                     | 1 million molecules (pre-training) and 269–457 authentic compounds (transfer Learning) |                | SMILES                                                  |
| Kessert et al. [156]              | 2021 | HILIC & RP-LC | Graph Convolutional Networks (GCNs) | 77,980 (SMRT, 852(RIKEN) and 1,400 (Fiehn HILIC) authentic molecules |                | Graph and 25 atom and bond features (continued on next page) |
| Yang et al. [157]                 | 2021 | HILIC & RP-LC | GNNs + TL                     | in silico HILIC RT dataset with about 306 K molecules for GNNs, 100--200 molecules for TL |                | 16 kinds of atoms and 4 kinds of bonds |
pre-training is another option to overcome the limitation of the training data required for training ANNs models [159,163,164].

5. Fusion tools for metabolite identification based on LC–MS/MS data

Metabolite identification based on LC–MS/MS data is a sophisticated subject that involves analysis of authentic standard compound LC–MS/MS data, comparison between query spectra and reference/in-silico spectra, sub-annotating of the features, and prediction of retention time of metabolites. In this regard, the development of fusion tools, in the form of client software or web servers, has greatly boosted the utilization of LC–MS/MS in the metabolism of complex biological samples (Table 2). ChemDistiller is a fusion software that combines a fingerprint prediction algorithm (FingerScorer) inspired by CSI:FingerID with an in-silico fragmentation algorithm (FragScorer) inspired by CFM-ID and MetFrag, and can retrieve and rank candidates from multiple target databases [165]. SIRIUS is a java-based software framework integrating a collection of tools, including ‘SIRIUS’ (the core function of SIRIUS), CSI:FingerID (with COSMIC), ZODIAC and CANOPUS [166–168]. Besides CSI:FingerID (a fingerprint predictor) and CANOPUS (a metabolites classifier) mentioned above, COSMIC [167] provides a confidence score for every structure annotated by CSI:FingerID and ZODIAC [166] performs de novo molecular formula annotation. Additionally, MS-DIAL combines MS-FINDER, LipidBlast with reference library search engine to solve the comprehensive identification mass spectrometry data (MS/MS) [64]. Combined with mass spectrometry data, the RT can also be utilized as additional and orthogonal information for the putative identification of small molecules [170]. MetFrag combines compound database searching, retention times prediction, and in-silico fragments prediction for small molecule identification from tandem mass spectrometry data [99,171,172]. Although, the fusion tools involving RT prediction are still limited compared to those tools based on both in-silico and experimental spectral libraries, retention time prediction and utilizing artificial neural networks, was proven to reduce structure isomer hit lists when used prior to in-silico spectral prediction software [149]. Retention order prediction and spectrum-based scores can also be combined for more accurate metabolite identifications in a LC–MS/MS experiment [148]. All these studies have thus triggered new initiatives for developing fusion tools based on RT to promote the use of LC–MS/MS.

6. Conclusions and perspectives

Commonly, the strategy based on authentic standard compounds can provide relative credible structure elucidation, whereas the number of the structure annotations is almost too small compared to the number of metabolites in complex biological extracts. Recently, the strategy based on public/commercial reference spectral libraries is developing rapidly, as the reference spectral libraries can be collected from the institutions and compa-
nies around the world in a simple way. The in-silico approaches can produce numerous complement structure annotations for the results of the above strategies, whereas the accuracy of the annotation based on in-silico approaches is quite lower than that of strategies based on authentic standard compounds or public/commercial reference spectral libraries. Although many tools for structure elucidation based on LC–MS/MS data have been developed in recent years, most of them have been shown to lack a degree of reliability that needs to be evaluated by a third-party organization. In particular, the Critical Assessment of Small Molecule Identification (CASMI) is an open contest on the identification of small molecules from mass spectrometry data and has been held five times since 2012 (https://casmi-contest.org/) [83,173]. With top or top n candidates used for evaluating the approaches, the individual in-silico approaches were proved to produce structure elucidation with low accuracy (17–25 %), while the combination of the strategy based on public/commercial reference spectral libraries and in-silico approaches can correctly identified up-to 87–93 % [83].

I t is important to note that a much larger, comprehensive and user-friendly library will be needed for complex deep learning models, like transfer learning, could accurately predict RT across different platforms.

In addition to the algorithms for structure elucidation of metabolites based on LC–MS/MS data, the construction and preservation of public data are also crucial to metabolomics. Although a large number of reference spectral databases have been built, most of the databases are still insufficient to train a complex model, especially for deep learning models. With the rapid development of public MS libraries, like CNPS and MassBank, we are convinced that a much larger, comprehensive and user-friendly library will be established for the researchers in the world. Such a resource, alongside faithful adherence to recently published standards for metabolomics reporting [174], allow us to be confident that the large coverage gap between annotated and non-annotated metabolites will ultimately be closed.

Zhitao Tian: Writing – original draft. Fangzhou Liu: Data curation. Dongqin Li: Data curation. Alisdair R. Fernie: Writing – review & editing. Wei Chen: Funding acquisition, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Natural Science Foundation for Distinguished Young Scientists of Hubei Province (No. 2021CFA058), the Young Top-notch Talent Cultivation Program of Hubei Province, the Fundamental Research Funds for the Central Universities (No. 2662019PY008).

References

[1] Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. Trends Biotechnol 1998;16(9):373–8. https://doi.org/10.1016/S0167-7799(08)01214-4
[2] Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 2012;13(4):263–9. https://doi.org/10.1038/nrm3314
[3] Fang C, Fernie AR, Luo J. Exploring the diversity of plant metabolism. Trends Plant Sci 2019;24(1):83–98. https://doi.org/10.1016/trplants.2018.09.006
[4] Wishart DS. Metabolomics for investigating physiological and pathophysiological processes. Physiol Rev 2019;99(4):1819–75. https://doi.org/10.1152/physrev.00028.2018
[5] da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. PNAS 2015;112(41):12459–50. https://doi.org/10.1073/pnas.1516879112
[6] Johnson SR, Lange BM. Open-access metabolomics databases for natural product research: present capabilities and future potential. Front Biogeo Biotechnol 2015;3: https://doi.org/10.3189/fbto.2015.00027.
[7] Bocker S. Searching molecular structure databases using tandem MS data: are we there yet? Curr Opin Biotechnol 2017;36:1–8. https://doi.org/10.1016/j.copbio.2016.12.010.
[8] Blazevovic I, Kind T, Ji J, Fiehn O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. Metabolites 2018;8(2). https://doi.org/10.3390/metabo8020022
[9] Haddad PR, Taraji M, Szucs R. Prediction of analyte retention time in liquid chromatography. Anal Chem 2021;93(1):228–56. https://doi.org/10.1021/acs.analchem.0c04190
[10] Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, et al. Identification of small molecules using accurate mass MS/MS search. Mass Spectrom Rev 2018;37(4):513–32. https://doi.org/10.1002/mas.21535
[11] Chen GL, Burunsky DR. Hybrid mass spectrometers for tandem mass spectrometry. J Am Soc Mass Spectrom 2018;19(2):161–72. https://doi.org/10.1016/j.jasms.2017.11.013.
[12] Werner E, Hellier JF, Ducruix C, Ezan E, Junot C, et al. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. J Chromatogr B-Anal Technol Biomed Life Sci 2008;871(2):143–63. https://doi.org/10.1016/j.chrombi.2007.07.004.
[13] Ichou F, Schwarzenberg A, Lesage D, Alves S, Junot C, et al. Comparison of the activation time effects and the internal energy distributions for the CID, PEP, and HCD excitation modes. J Mass Spectrom 2014;49(6):498–508. https://doi.org/10.1002/jms.3365.
[14] Chaleclis K, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC-MS/MS-based metabolomics. Curr Opin Biotechnol 2019;55:44–50. https://doi.org/10.1016/j.copbio.2018.07.018.
[15] Tautenhahn R, Bottcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinf 2008;9. https://doi.org/10.1186/1471-2105-9-554.
[16] Zhu G, Wang S, Huang Z, Zhang S, Liao Q, et al. Rewiring of the fruit metabolism in tomato breeding. Cell 2018;172(1–2):249–261 e12. https://doi.org/10.1016/j.cell.2017.12.019.
[17] Wen W, Li D, Li X, Gao Y, Li W, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun 2014;5:3438. https://doi.org/10.1038/ncomms4388.
[18] Chen W, Gao Y, Xie W, Gong L, Lu K, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 2014;46(7):714–21. https://doi.org/10.1038/ng.3007.
[19] Chen J, Hu X, Shi T, Yin H, Sun D, et al. Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. Plant Biotechnol J 2020;18(8):1722–35. https://doi.org/10.1111/pbi.13335.
[20] Shi T, Zhu A, Jia J, Hu X, Chen J, et al. Metabolomics analysis and metabolite–agronomic trait associations using kernels of wheat (Triticum aestivum) recombinant inbred lines. Plant J 2020;103(1):279–92. https://doi.org/10.1111/tpj.13777.
[21] Marie B, Gallet A. Fish metabolome from sub-urban lakes of the Paris area (France) and potential influence of noxious metabolites produced by cyanobacteria. Chemosphere 2022;296. https://doi.org/10.1016/j.chemosphere.2022.134035.
[22] Zhalina K, Louie KB, Hao Z, Mansoori N, da Rocha UN, et al. Dynamic root exudate chemistry and microbial substrate preferences drive patterns in rhizosphere microbial community assembly. Nat Microbiol 2018;3(4):470–80. https://doi.org/10.1038/s41564-018-0129-3.
[23] Li JX, Wang ZZ, Zhai GT, Chen CL, Zhu KZ, et al. Untargeted metabolic profiling identifies disease-specific and outcome-related signatures in chronic rhinosinusitis. J Allergy Clin Immunol 2022. https://doi.org/10.1016/j.jaci.2022.04.006.
[24] Moreau R, Claria J, Aguilar F, Fenaille F, Lozano JJ, et al. Blood metabolomics undercribes inflammation-associated mitochondrial dysfunction as a potential mechanism underlying ACLF. J Hepatol 2020;72(4):688–701. https://doi.org/10.1016/j.jhep.2019.11.009.
[25] Tsugawa H, Rai A, Saito K, Nakabayashi R. Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. Nat Prod Rep 2021;38(10):1729–59. https://doi.org/10.1039/d1np00014d.
[176] Creek DJ, Janekevics A, Breitling R, Watson DG, Barrett MP, et al. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography-mass spectrometry: improved metabolite identification by retention time prediction. Anal Chem 2011;83(22):8703–10. https://doi.org/10.1021/ ac2018237.

[177] D’Archivio AA, Maggi MA, Ruggieri F. Modelling of UPLC behaviour of acylcarboxylates by quantitative structure-retention relationships. J Pharm Biomed Anal 2014;96:224–30. https://doi.org/10.1016/j.jpba.2014.04.006.

[178] Kouskoura MC, Hadjipavlou-Litina D, Markopoulou CK. Elucidation of the retention mechanism on a reverse-phase cyano column by modeling. J Sep Sci 2014;37(15):1915–20. https://doi.org/10.1002/jssc.201400957.

[179] D’Archivio AA, Maggi MA, Ruggieri F. Prediction of the retention of s-triazines in reversed-phase high-performance liquid chromatography under linear gradient-elution conditions. J Sep Sci 2014;37(15):1930–6. https://doi.org/10.1002/jssc.201400946.

[180] Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, et al. Predicting retention Ovcacikova M, Lisa M, Cifkova E, Holcapek M. Retention behavior of lipids in Wolfer AM, Lozano S, Umbdenstock T, Croixmarie V, Arrault A, et al. UPLC-MS Falchi F, Bertozzi SM, Ottonello G, Ruda GF, Colombano G, et al. Kernel-based, Kubik L, Wiczling P. Quantitative structure-(chromatographic) retention relationship models to support nontarget high-structure-retention relationship models for dissociating compounds. J Pharm Biomed Anal 2016;127:176–83. https://doi.org/10.1016/j.jpba.2016.02.050.

[181] Aicheler F, Li J, Hoene M, Lehmann R, Xu G, et al. Retention time prediction in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics 2015;11 (3):696–706. https://doi.org/10.1007/s11306-015-0888-2.

[182] Munro K, Miller TH, Martins CP, Edge AM, Cowan DA, et al. Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data. J Chromatogr A 2015;1396:34–44. https://doi.org/10.1016/j.chroma.2015.03.063.

[183] Falchi F, Bertozzi SM, Ottolino G, Ruda GF, Colombano G, et al. Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: A useful tool for metabolite identification. Anal Chem 2016;88(19):9510–7. https://doi.org/10.1021/acs.analchem.6b02075.

[184] Ovcacikova M, Lisa M, Cifkova E, Holcapek M. Retention behavior of lipids in reversed-phase ultrahigh-performance liquid chromatography-electrospray ionization mass spectrometry. J Chromatogr A 2016;1450:76–85. https://doi.org/10.1016/j.chroma.2016.04.082.

[185] Ailazadeh R, Thomaidis NS, Bletsou AA, Cago-Ferrero P. Quantitative structure-retention relationship models to support nontarget high-resolution mass spectrometric screening of emerging contaminants in environmental samples. J Chn Inf Model 2016;56(7):1384–98. https://doi.org/10.1016/j.cheminf.2016.07.002.

[186] Wolfer AM, Lozano S, Umbdenstock T, Croixmarie V, Arrault A, et al. UPLC-MS retention time prediction: a machine learning approach to metabolite identification in untargeted profiling. Metabolomics 2016;12(1). https://doi.org/10.1007/s11306-015-0884-2.

[187] Kohk L, Wiczling P. Quantitative structure-(chromatographic) retention relationship models for dissociating compounds. J Pharm Biomed Anal 2016;127:176–83. https://doi.org/10.1016/j.jpba.2016.02.050.

[188] Barron LP, McEneff GL. Gradient liquid chromatographic retention time prediction for suspect screening applications: A critical assessment of a generalised artificial neural network-based approach across 10 multi-residue reversed-phase analytical methods. Talanta 2016;147:261–70. https://doi.org/10.1016/j.talanta.2015.09.065.

[189] Randazzo GM, Tonoli D, Hambye S, Guillerme D, Janneret F, et al. Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. Anal Chim Acta 2016;918:8–16. https://doi.org/10.1016/j.aca.2016.02.014.

[190] Taraji M, Haddad PR, Amos RI, Talebi M, Szucs R, et al. Prediction of retention in hydrophilic interaction liquid chromatography using solute molecular descriptors based on chemical structures. J Chromatogr A 2017;1486:59–67. https://doi.org/10.1016/j.chroma.2016.12.025.

[191] Zhang X, Li J, Wang C, Song D, Hu C. Identification of impurities in macrolides by liquid chromatography-mass spectrometric detection and prediction of retention times of impurities by constructing quantitative structure-retention relationship models for hydrophilic interaction liquid chromatographic systems. J Chromatogr A 2017;1507:53–62. https://doi.org/10.1016/j.chroma.2017.05.044.

[192] Wen Y, Talebi M, Amos RJJ, Szucs R, Dolan JW, et al. Retention prediction in reversed phase high performance liquid chromatography using quantitative structure-retention relationships applied to the Hydrophobic Subtraction Model. J Chromatogr A 2018;1541:1–11. https://doi.org/10.1016/j.chroma.2018.01.053.

[193] Wen Y, Amos RJJ, Talebi M, Szucs R, Dolan JW, et al. Retention index prediction using quantitative structure-retention relationships for improving structure identification in nontargeted metabolomics. Anal Chem 2018;90 (15):9434–40. https://doi.org/10.1021/acs.analchem.8b02084.

[194] McEachran AD, Mansouri K, Newton SR, Beverly BEJ, Sobus JR, et al. A comparison of three liquid chromatography (LC) retention time prediction models. Talanta 2018;182:371–9. https://doi.org/10.1016/j.talanta.2018.01.027.

[195] Hall LM, Hill DW, Bugden K, Cawley S, Hall LH, et al. Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds. J Chem Inf Model 2018;58(3):591–604. https://doi.org/10.1021/acs.jcim.8b00496.

[196] Bouwmeester R, Martens L, Degroeve S. Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. Anal Chem 2019;91(5):3694–703. https://doi.org/10.1021/acs.analchem.8b02085.

[197] Souhi A, Mohai MP, Falm E, Malm L, Krute A. MultiConditionRT: Predicting liquid chromatography retention time for emerging contaminants for a wide range of eluent compositions and stationary phases. J Chromatogr A 2022,1666.: https://doi.org/10.1016/j.chroma.2022.462867.

[198] Liapikos T, Zisi C, Kodra D, Kademoglou K, Diamantidou D, et al. Quantitative structure-retention relationship (QSSR) modelling for Analytes’ retention prediction using quantitative structure-retention relationships in reversed phase high performance liquid chromatography. J Sep Sci 2022;1191:. https://doi.org/10.1002/jssc.2022.123132.