

Machine learning methods for compound annotation in non-targeted mass spectrometry—A brief overview of fingerprinting, *in silico fragmentation* and *de novo* methods

Francesco F. Russo¹ | Yannek Nowatzky² | Carsten Jaeger³ | Maria K. Parr⁴  | Philipp Benner² | Thilo Muth⁵ | Jan Lisec¹ 

¹Department of Analytical Chemistry and Reference Materials, Organic Trace Analysis and Food Analysis, Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany

²eScience, Bundesanstalt für Materialprüfung und -forschung, Berlin, Germany

³Department of Analytical Chemistry and Reference Materials, Environmental Analysis, Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany

⁴Institute of Pharmacy, Pharmaceutical and Medicinal Chemistry (Pharmaceutical Analyses), Freie Universität, Berlin, Germany

⁵Department MF 2, Domain Specific Data Competence Centre, Robert Koch Institut, Berlin, Germany

Correspondence

Jan Lisec, Department of Analytical Chemistry and Reference Materials, Organic Trace Analysis and Food Analysis, Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany.

Email: jan.lisec@bam.de

Non-targeted screenings (NTS) are essential tools in different fields, such as forensics, health and environmental sciences. NTSs often employ mass spectrometry (MS) methods due to their high throughput and sensitivity in comparison to, for example, nuclear magnetic resonance-based methods. As the identification of mass spectral signals, called annotation, is labour intensive, it has been used for developing supporting tools based on machine learning (ML). However, both the diversity of mass spectral signals and the sheer quantity of different ML tools developed for compound annotation present a challenge for researchers in maintaining a comprehensive overview of the field.

In this work, we illustrate which ML-based methods are available for compound annotation in non-targeted MS experiments and provide a nuanced comparison of the ML models used in MS data analysis, unravelling their unique features and performance metrics. Through this overview we support researchers to judiciously apply these tools in their daily research. This review also offers a detailed exploration of methods and datasets to show gaps in current methods, and promising target areas, offering a starting point for developers intending to improve existing methodologies.

1 | INTRODUCTION

Mass spectrometry (MS) comprises a variety of analytical methods that ultimately yield ion intensities or mass spectra representing molecules contained in the processed samples. The assignment of chemical identity to the mass spectral data, known as annotation, is crucial in many scientific domains, particularly in environmental and health sciences. The achievement of such assignments strongly depends on the type and structure of the mass spectra, including factors such as resolution and complexity. Except for the most straightforward cases, annotation is a time-consuming process that requires expert

knowledge. This situation has sparked an interest in the use of machine learning (ML), a field that explores the use of algorithms capable of “learning” from data, making their development more cost effective, faster and more precise compared to conventional, human-designed algorithms. However, ML-based approaches require and depend on training data, which can be quite diverse in the field of spectral annotation. Throughout this introduction, we, therefore, will present commonly used MS techniques and methods, focusing on the properties of the generated data that are relevant for both conventional and ML-based annotation approaches (Figure 1). Frequently used data repositories are presented in Table S1.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](#) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Rapid Communications in Mass Spectrometry* published by John Wiley & Sons Ltd.

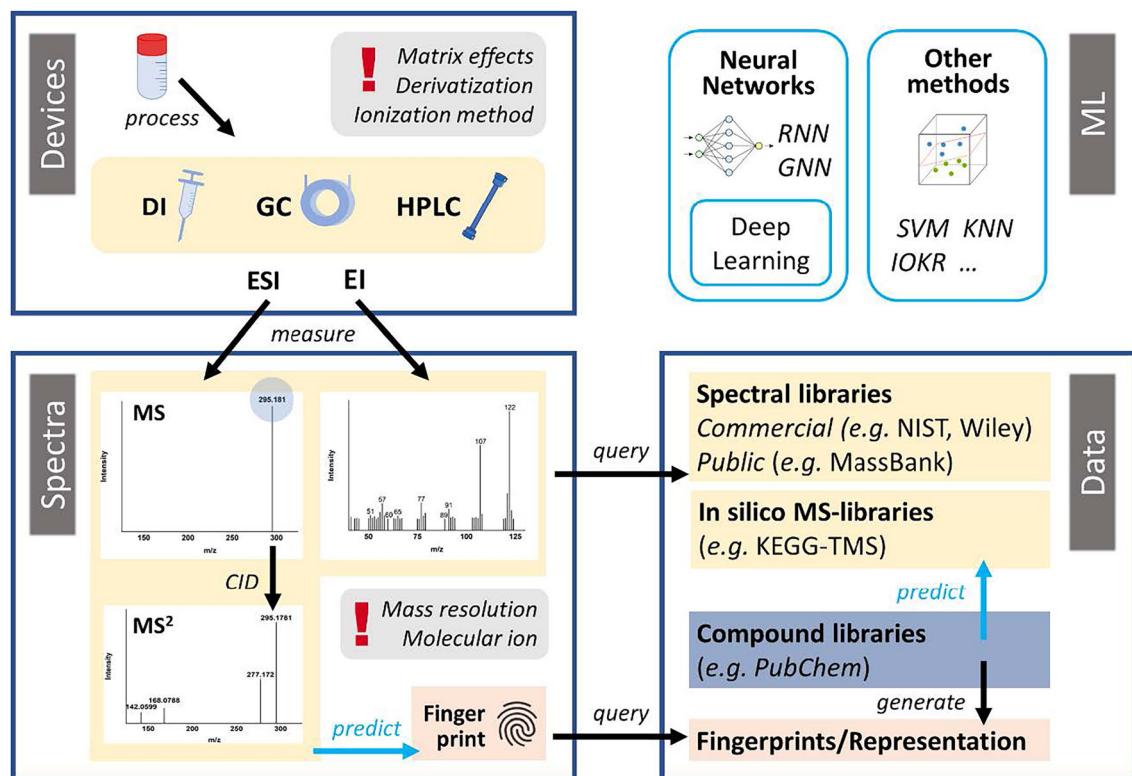


FIGURE 1 Graphic overview of relevant components in compound annotation of mass spectrometry data. The data obtained from the experiment can be highly heterogeneous depending on the preparation (e.g., derivatisation), the type of measurement (direct injection [DI], coupling with gas chromatography [GC] or high-performance liquid chromatography [HPLC], supercritical fluid chromatography [SFC], capillary electrophoresis [CE], etc.) and instrumental factors such as the ion source (e.g., electron ionisation [EI], electrospray ionisation [ESI] and chemical ionisation [CI]) and device resolution. Spectral annotation is often attempted by querying a spectral database. However, spectral databases have limited coverage due to the slow and laborious process of measuring reference spectra. More comprehensive databases are generated using in silico fragmentation tools, which were steadily improved with the use of machine learning (ML). Alternatively, the molecular ion (if present) can be used to assign a chemical formula and query a spectral library. The direct querying of compound libraries can return overwhelming quantities of candidates, thus fuelling the development of fingerprinting methods. Fingerprinting methods are ML-based tools that predict a molecular fingerprint from a mass spectrum, offering a more specific query for spectral databases. ML-based steps are indicated by blue arrows, and ML methods discussed in this review are indicated in the upper right panel. [Color figure can be viewed at wileyonlinelibrary.com]

MS experiments can be conducted in either a targeted or non-targeted fashion. Traditionally, targeted approaches are used to identify and quantify compounds of interest within a sample. These approaches require each substance of interest to be known and a standard to be available for confirming the identity of the compound and enabling quantitation. As a consequence, only a limited number of compounds are monitored, even though well more than 100 may be integrated. However, in fields such as metabolomics or monitoring of narcotics, where novel compounds are of interest, targeted approaches are of limited use. In this case, an alternative is offered by non-targeted methods. The predecessor of modern non-targeted methods involved gas chromatography coupled with electron ionisation mass spectrometry (GC-EI-MS). This approach offers advantages, as EI-MS fragmentation patterns exhibit good reproducibility, and extensive spectral libraries are available.^{1,2} However, the restriction to volatile compounds and the occurrence of extensive fragmentation, which can result in the loss of the molecular ion and ambiguous interpretation, have limited the applications of EI

in non-targeted MS. As an alternative, liquid chromatography in combination with soft ionisation methods, such as electrospray ionisation (ESI), enables the analysis of a broader range of substances. Soft ionisation techniques, when combined with high-resolution mass spectrometry, often allow the deduction of the molecular formula from the molecular ion and enable querying a compound database (e.g., PubChem³) for potential candidates in a successive step.

As previously stated, various MS techniques have been employed in non-targeted approaches. The type of instrument used for data generation is an important question for ML as it influences mass spectra and therefore the potential application of ML. Briefly, a mass spectrometer can be schematically divided into three components: the ion source, the mass analyser and the detector. The ion source generates ions from a typically neutral analyte, allowing us to separate analyte ions based on their mass-to-charge (m/z) ratio. Concerning the resulting mass spectra, the ion source often determines which analyte is amenable for analysis and plays an important role in how intensive the signal of a given analyte

is. Furthermore, the ionisation process determines whether the analyte is visible as positive or negative species and determines if radical cations, protonated or deprotonated molecular ions, alkali or similar adduct, clusters or only fragments are detected.

Due to their widespread use and the availability of data, EI and ESI have been at the core of ML method development. EI sources ionise analytes in the gas phase, with electrons accelerated at 70 eV. The energy transferred to the analyte molecule causes fragmentation of the molecule, resulting in a mass spectrum characterised by multiple signals representing the fragments of the molecule, with no guaranteed presence of the ionised intact molecule, that is, the molecular ion.

In contrast, ESI sources typically produce spectra using a lower number of signals. The ESI source transfers charged ions from a solution into the gas phase.⁴ Spectra resulting from ESI sources usually present signals for protonated or deprotonated analytes, complexes with ions, adducts and/or multimers. In most cases, the unfragmented analyte is observed. Although this is an advantage compared to EI, it comes at the cost of losing structural information that may be deduced from the fragmentation pattern. A solution to this conundrum is offered by tandem MS (abbreviated as MS/MS or MS²). In an MS² experiment, an ion of interest, referred to as the precursor ion, is fragmented into so-called product ions. The most common method of fragmentation is collision-induced dissociation (CID), where collision with neutral gas molecules is used to induce fragmentation of the precursor ion.

After ionisation, the ions are transferred to the mass analyser, where they are separated based on their *m/z* ratio. The mass analyser determines the resolution of the resulting spectrum. The resolution of the mass spectra ranges from nominal mass (NM) spectra to high-resolution (HR) spectra (mass accuracy level <5 ppm and mass resolution >10 000 full width at half maximum).⁵ This diversity of ionisation methods (EI, ESI) and MS techniques (MS, MS²) is reflected in highly heterogeneous MS data, leading to time-consuming data evaluation and annotation for non-targeted experiments. Consequently, various (bio)informatics tools for automated batch processing have been developed.

Annotation is often achieved through laborious expert evaluation, making automation approaches particularly intriguing but complex. Promising strategies for automating and enhancing annotation often rely on ML. Indeed, applications of ML have surged over the past decades and become prominent in bioinformatics⁶ and life sciences,⁷ in general. In ML, mathematical models are estimated (trained) on observables (training data) to learn patterns, enabling the programme to make meaningful predictions about new observations.⁸ With the recent increase in computational power and the availability of massive datasets, especially deep learning has gained enormous attention in the scientific community.

In the broader context of non-targeted MS, a variety of ML applications have emerged, successfully addressing various data processing tasks to automate and refine compound annotation and interpretation. These include fragment ion identification, prediction of fragmentation pathways from molecular structures and fingerprinting

methods to characterise molecules. Algorithms for these tasks range from conventional ML and statistical methods, such as Markov processes and kernel methods, to deep neural networks (DNN). Kernel methods express similarities between data points, with a kernel function serving as the basis for high-dimensional learning techniques.⁹ Examples are support vector machines (SVM), which optimise a linear boundary between training classes in the high-dimensional feature space, resulting in a nonlinear boundary in the original feature space by computing only the kernel function.^{9,10} Nevertheless, kernel methods have fallen out of favour due to the resource-intensive estimation of kernel parameters. Artificial neural networks (ANN) are a diverse class of nonlinear statistical models that can be efficiently estimated on large datasets. ANNs extract a linear combination of their inputs to model the desired output as a nonlinear function of the derived feature combinations.⁹ This is accomplished by stochastic gradient descent (SGD) methods that allow training on specialised hardware (e.g., graphical processing unit) where only limited memory is available. Several reviews offer a comprehensive overview of current computational methods for metabolomics. We refer to Liebal et al.,¹¹ Nguyen et al.,¹² Petrick and Shomron¹³ and Pomyen et al.¹⁴ Liebal et al.,¹¹ Petrick and Shomron¹³ and Pomyen et al.¹⁴ offer broad overviews of the use of ML for various steps of the metabolomics data processing pipeline, like peak picking, quantification and data interpretation. On the contrary, Nguyen et al.¹² focuses on structure annotation and ML-based annotation in particular. The described methods enable the processing of substantial quantities of non-targeted MS data, reducing the need for manual evaluation. ML methods provide advantages over more traditional methods. For instance, predicting mass spectra using ML models is faster than ab initio quantum mechanical simulations. Additionally, fingerprinting methods (described in Section 2.1) reduce dependence on spectral libraries. *De novo* methods generate new, previously unreported structures and extend compound libraries more rapidly than usual.

However, the ML methods and data types used among the various software are not the focus of aforementioned reviews. Our contribution to this review is to shed light on various state-of-the-art ML applications for non-targeted metabolomics, clarifying the methods utilised for different types of MS and the involved datasets. We also aim to identify gaps in current software and potential target areas for the development of future applications. We focus on tools for *in silico* fragmentation, the prediction of representations from mass spectra to query compound libraries (i.e., chemical formulas, molecular fingerprinting, representation-based methods) and *de novo* methods. We would nevertheless like to mention that the use of ML in MS also includes the prediction of orthogonal properties (e.g., collision cross section, chromatographical properties) and the development of scoring functions for querying and networking (e.g., MS2DeepScore,¹⁵ DeepMass,¹⁶ MS2Query¹⁷). Neither of these topics will be treated in this review, and we would refer interested readers to Liebal et al.¹¹ and Petrick and Shomron¹³ instead.

Following a formal systematic review process, we queried Web of Science (Clarivate) using the keyword search terms “(‘machine

learning' OR ML) AND (MS or 'mass spectrometry') AND annotation" and "('machine learning' OR ML) AND (MS or 'mass spectrometry').'" Additionally, the results were filtered to include entries with the citation topic "mass spectrometry" and to exclude the topic "proteomics." The filtered entries were exported to files, the files were merged, and duplicate entries were removed. In a first step the titles were manually evaluated to sort out publications which did not directly fall in the focus of this manuscript. Finally, we conducted a full-text analysis of the resulting 54 publications that were re-evaluated for relevance of the present study. A graphical overview of their fields and applicability for the two considered ion source types is shown in Figure 2.

2 | REPRESENTATION-BASED METHODS

Based on the acquired mass spectrum of an unknown compound, the assignment of the corresponding chemical identity is often achievable by querying a spectral library. As a result, candidate compounds are ranked based on a similarity score between the query spectrum and the library spectrum. However, building spectral libraries by measuring the spectra of pure substances is a time-consuming and expensive process, resulting in libraries that grow slowly and often remain incomplete. The incomplete coverage of spectral libraries has prompted the search for alternatives in compound identification. Of particular interest is the use of compound databases, such as PubChem,³ which comprise collections of molecules orders of magnitude larger than spectral databases. The primary objective of compound databases, however, lies in the collection of compounds and their properties, rather than on mass spectra. Therefore, compound databases require a query other than a mass spectrum,

such as the molecular mass or a chemical formula. Queries with little specificity, such as molecular mass, can yield a large number of candidates, requiring additional information for further filtering and ranking. An approach to tackle this challenge is the use of molecular fingerprints or an intermediate representation, which can be easily generated for a given molecule. In their most common form, molecular fingerprints are vectors of fixed length that encode the presence or absence of molecular features (i.e., topological, physico-chemical or electrical properties of a compound¹⁸). The molecular structure of a compound can be encoded into a molecular fingerprint, allowing compounds in compound databases to be converted into molecular fingerprints. ML has been employed to predict molecular fingerprints from mass spectra. The predicted molecular fingerprints can be used to query compound databases similar to mass spectral libraries. This approach offers the advantage of efficiently computing the similarity between molecular fingerprints, making it advantageous for querying large databases. Nevertheless, molecular fingerprints were not designed for the task of identifying compounds from MS spectra and thus are unspecific and in some case might contain redundant features.

In the following text, we will discuss methods that predict molecular fingerprints, substructures and embeddings for querying and scoring molecules from a database. Two groups of fingerprinting methods can be distinguished: methods based on supervised learning and methods based on unsupervised learning. Supervised methods require a labelled training dataset, that is, a dataset consisting of inputs and the desired outputs, to train the algorithm. In most cases considered in this manuscript, algorithms were trained to predict molecular fingerprints (the label). The training set consists of mass spectral libraries (the input) and the molecular fingerprints for each compound in the library. Unsupervised methods, on the contrary, do not require labelled data; that is, only the input data (without the desired output) are required. These methods aim to identify patterns and groupings, such as neutral losses or functional groups, that appear in a given set of spectra. To be more precise, the unsupervised methods reported in this work are inspired by natural language modelling and interpret mass spectra as "documents" containing "words" (e.g., peaks, mass differences). These methods then model the relationship between "words" and their respective "distributions," allowing the identification of similar compounds based on the co-occurrence of similar MS patterns or by their "semantic" distance.

The field of supervised fingerprinting is dominated by the SIRIUS¹⁹ suite, which is the most cited tool of all those considered in this section. The SIRIUS suite and CSI:FingerID²⁰ not only are efficient as useful tools but also offer a user-friendly graphical user interface (GUI) which facilitates operation for users inexperienced with command line interfaces (CLI) and/or without knowledge of programming languages.

CSI:FingerID²⁰ has emerged from kernel-based fingerprinting methods and differentiates itself from other similar methods through the use of fragmentation trees in addition to MS^2 spectra as input. The computation of fragmentation trees is time consuming and has contributed to the interest in alternative methods. ANN-based

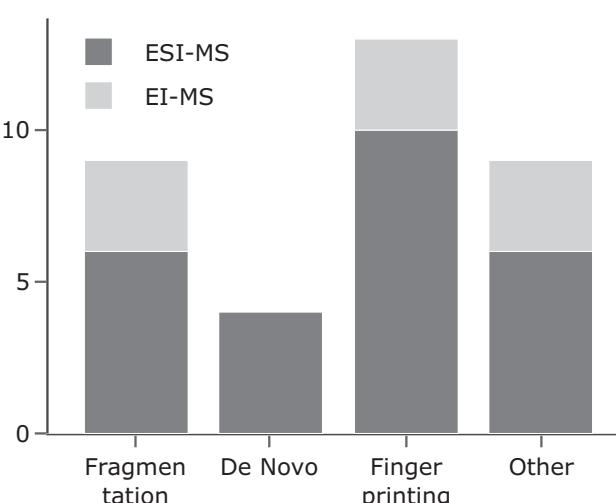


FIGURE 2 Overview of publications categorised by the task they solve and the type of mass spectrometry (MS) data they can be used on (electron ionisation [EI], electrospray ionisation [ESI]). After review papers were excluded, we found that most of the publications analysed were concerned with ESI-MS and only a few with EI-MS.

alternatives offer the possibility of predicting all bits of a fingerprint at once, instead of requiring one estimator per bit (like with SVMs). The use of ANNs allowed the exploration of novel approaches based on the use of embeddings. Spec2Vec²¹ and ADAPTIVE both use embeddings as fingerprints, which should result in fingerprints which are more specific for the task and faster calculation of similarity scores. At the moment of writing, Spec2Vec is, together with MS2LDA,²² one of the most cited tools after CSI:FingerID and the SIRIUS^{19,23,24} suite. The two unsupervised fingerprinting tools, Spec2Vec and MS2LDA, serve a different use than supervised fingerprinting tools and use different approaches. Spec2Vec²¹ generates a “fingerprint” which can be used similar to supervised fingerprinting methods, whereas MS2LDA²² works like a networking method that uncovers relationships between compounds in a dataset. When developing a new method, accessibility to the end user should be considered. Of the four most popular tools, only Spec2Vec is provided as CLI python package and does not offer a GUI.

2.1 | Fingerprinting methods

Heinonen et al.²⁵ used SVMs to predict fingerprints from MS² spectra. The authors examined linear and quadratic integral mass kernels and HR mass kernels (probability product kernel previously described by Kondor and Jebara²⁶) separately or in combination with mass difference kernels and neutral loss kernels. Three datasets were used in the study: the triple quadrupole (QqQ) dataset, which consisted of 514 compound spectra, recorded in ESI positive mode at NM resolution, using five different collision energies between 10 and 50 eV; the Ltq dataset consisting of 293 compound spectra recorded in positive mode at HR; and the lipid dataset, consisting of HR negative-mode spectra of 403 phosphatidylethanolamines. Both HR datasets were measured on Orbitrap instruments. The authors found that the HR kernel on all features resulted in the best performance on average. The performance of FingerID was compared with MetFrag²⁷ on 20 spectra randomly removed from the QqQ dataset and 20 spectra randomly removed from the Ltq dataset. The two datasets were used to query both PubChem^{3,28} and KEGG (Kyoto Encyclopedia of Genes and Genomes).²⁹⁻³² The performance was reported as the top 10 recall rate, indicating how often the correct molecule was ranked within the first 10 candidates. The reported top 10 recall rates are summarised in Table S2, where reported recall rates from all the reviewed manuscripts can be found.

Building on the work by Heinonen et al.,²⁵ Brouard et al.³³ developed a method based on input output kernel regression (IOKR) to map MS² spectra to molecular fingerprints. The method enables the prediction of a fingerprint for a given ESI-MS², which was subsequently used to query a compound database for candidates. The authors trained a linear combination of input kernels with multiple kernel learning (MKL) to map the input spectra to an intermediary representation. Three output kernels were used to map the intermediary representation to a molecular fingerprint of 4138 compounds. Candidates were queried from PubChem and ranked

based on the distance from the predicted feature vector. CSI:FingerID, a state-of-the-art annotation tool within the SIRIUS suite, was used for comparison with respect to the top *n* recall rates.

ADAPTIVE³⁴ is a fingerprinting tool developed using ESI-MS² data, which was proposed to overcome the problem with redundancy and lack of specificity present in manually curated fingerprints. To improve the specificity, a custom fingerprint, which the authors refer to as molecular vector, was generated using a message passing neural network (MPNN). In a second step, IOKR was used to learn a mapping from MS spectra to the vector representation. The MPNN was trained to generate vectors from molecular graphs while maximising the correlation between the generated molecular vector and the ESI-MS² spectra. The performance was compared with IOKR,³³ FingerID²⁵ and CSI:FingerID²⁰ on a dataset of 4138 ESI-MS² spectra from the Global Natural Product Social (GNPS) spectral library regarding the top *n* recall rates.

Similarly, MetFID³⁵ uses a DNN to predict the fingerprint of a compound from ESI-MS² spectra. The DNN that generates the molecular fingerprint was trained on 11 748 spectra of 5667 compounds from MoNA and 122 481 spectra of 10 731 compounds from NIST 2017. The training set consisted of only ESI positive-mode (ESI+) spectra within a mass range of 100–1010 Da measured on QqQ, Orbitrap or quadrupole time-of-flight instruments. Furthermore, only spectra of H⁺ and NH₄⁺ adducts were kept in the training set. Multiple entries for the same compound were merged, with exception of the MS² spectra of different collision energies. The spectra were normalised to relative intensities, scaled between 0 and 100, denoised by removing peaks with relative intensities lower than 10 and spectra with less than five peaks with a relative intensity higher than 2%. Finally, the spectra were binned to NM vectors. The performance of MetFID was compared to MetFrag and ChemDistiller³⁶ on a test set of 482 compounds removed from the training data, where it performed significantly better than the other methods. Additionally, MetFID was compared with CSI:FingerID,²⁰ which was trained on data from the NIST library, on the CASMI 2016³⁷ dataset. Again, recall rates are presented in Table S2.

IDS_MINT³⁸ is a tool developed to allow a simpler use and development of ML tools for non-targeted ESI-MS². IDS_MINT allows to easily train fingerprinting models based on the transformer architecture. The models can be trained on a custom library focused on the end users' needs. To demonstrate the performance of the pipeline, the authors trained two transformer models, one for negative-mode and one for positive-mode ESI-MS² spectra. Each model consisted of four hidden encoder and decoder layers and two attention heads. The models were trained to predict ECFP2 fingerprints of a subset of the LIPID MAPS library. The subset was derived from the MoNA and GNPS libraries and was cleaned by removing in silico spectra and spectra with over 10% of the peaks outside the *m/z* range of 50–1000. The resulting training sets for positive and negative modes contained mass spectra of 2617 unique lipids and 1722 unique compounds, respectively. The publicly available metabolomics study ST002044 from the Metabolomics WorkBench database was used as the test set, containing 3386 and

1901 unique mass spectra in positive and negative modes, respectively. When the LIPID MAPS database were queried with fingerprints generated from the test set, the trained models reached a top 1 recall rate of 35.4% and 35.1% for positive- and negative-mode spectra, respectively.

As part of the SIRIUS suite/web service,¹⁹ SIRIUS is a state-of-the-art tool for chemical formula prediction and fragmentation tree computation of ESI-MS² spectra. To identify the most likely chemical formula, SIRIUS iterates candidate formulas matching the precursor peak and computes the fragmentation tree for each, solving the maximum colourful subtree problem using integer linear programming.³⁹ Uncommon elements in the molecular formulas are detected from isotope patterns using a DNN.²³ Unfeasible formulas are filtered out, and unlikely ones are identified using an SVM.²³ Next, SIRIUS determines the posterior probability for each tree using Bayesian statistics. This is done by estimating the prior probability from the size of the tree, neutral losses and fragment formulas and the likelihood of the MS² spectrum given the fragmentation tree. Molecular formulas are ranked according to the posterior probability of their best fragmentation tree. SIRIUS 3.0 was trained and tested on GNPS and AFT library datasets and reportedly outperformed all competing methods and previous versions with a top 1 recall rate of 76% and a top 5 recall rate of 91%.

CSI-FingerID,²⁰ which is available as part of the SIRIUS suite, is a fingerprinting tool that uses the MS² spectra and the fragmentation trees generated by SIRIUS as input to predict molecular fingerprints. In the training phase a library of 4138 compounds from the GNPS⁴⁰ library and 2120 compounds from the AFT library were used as the training dataset. Multiple fragmentation tree kernels and multiple spectrum kernels were used to compute the similarities for each pair of compounds in the training dataset. SVMs, one for each bit in the target fingerprint, are trained on the kernel similarities to discriminate between compounds that exhibit a given bit of the fingerprint and those that do not. CSI:FingerID was compared with FingerID²⁵ (retrained on the same training data as CSI:FingerID), CFM-ID, MAGMa, MIDAS and MetFrag.²⁷ For comparison, PubChem was queried with 3868 compounds from GNPS and 2055 from the AFT library. The top 1 recall rate of the three best-performing methods was reported for both libraries individually and for a mixed library of both. An overview of the reported top 1 recall rates is presented in Table S2, and recall rates up to 20 for all tools can be found in the figures of the original publication.³⁵

MIST is a fingerprint tool developed by Goldman et al.⁴¹ Inspired by CSI:FingerID,²⁰ MIST uses chemical formulae (C, H, N, O, P, S, F, Cl, Br, I, Si, B, Se, Fe, Co and As) for each peak and calculates the pairwise distance for all peaks. The chemical formulae and the respective intensities of the peaks are encoded using an multilayer perceptron (MLP). The embedded chemical formulae, the intensities and the pairwise formula differences between peaks are used as input for a chemical formula transformer. The pairwise difference is used to model the relationship between peaks and is featurised in the modified attention mechanism developed by the authors. The final representation of the transformer is used to predict

the molecular fingerprint. The prediction of the fingerprint is performed by stepwise unfolding. The unfolding is achieved by a model trained to reverse stepwise halving in size of the molecular fingerprint. Furthermore, a custom distance metric was learned by fine-tuning MIST with a contrastive learning objective. MIST was trained on 31 145 positive-mode ESI-MS² spectra of 27 797 compounds from the NIST, MoNA and GNPS spectral libraries. The resulting dataset was augmented with a purpose-built *in silico* fragmentation ANN. During training a sub-module to predict MAGMa substructures was used as an additional signal. The substructure annotation module used the final representation of the transformer to predict the 512-bit Morgan fingerprint for each sub-fragment while using a clipped cosine as loss function. The unfolding module is trained with the binary cross-entropy loss function calculated at each unfolding step. The loss function for training MIST was the sum of the binary cross-entropy of the final fingerprint, the sum of the losses at each unfolding step and the loss function of the substructure annotation module. Both the unfolding loss and the substructure annotation module were weighted with factors determined during hyperparameter tuning. For fine-tuning, the model weights were used as a contrastive space. For each compound, the 256 closest isomers by Tanimoto score were retrieved from PubChem as decoys and sampled in each batch proportionally to their Tanimoto similarity. A single-layer projection to map the fingerprints to the contrastive space was learned. The training objective was to minimise the distance of the true fingerprint to the latent transformer representation, while maximising the distance to the decoys. MIST was compared with CSI:FingerID on three separate splits of 20% data holdouts. Using an ensemble of five separately trained models with different random initialisation, MIST fingerprints reached higher cosine similarity to the ground truth spectra than CSI:FingerID in 11 994 of 18 700 predictions. Interestingly, the top 1 recall rate is higher for CSI:FingerID, whereas MIST has higher recall rates for $k > 20$.

In regard to EI-MS, Ljioncheva et al.⁴² developed a tool, which they call CSI:IOKR, to identify trimethylsilyl (TMS) derivatives of contaminants of emerging concern (CEC) using GC-EI-MS. CSI:IOKR consists of a product kernel as input kernel and a linear kernel as output kernel. The training data consisted of 4648 TMS derivative NM EI-MS spectra from the NIST 2017 *mainlib* and *replib* libraries. The fingerprint to be predicted was the combination of four fingerprints, with features being always present or duplicate removed. The test set consisted of GC-EI-MS spectra measured by the authors. For the test set 100 CECs were selected, derivatised and measured, resulting in spectra for 104 derivatives, and recall rates, compared with MetExpert,⁴³ are reported. On the test set CSI:IOKR demonstrates the viability of kernel methods for GC-EI-MS, which might see further development for a more general-use scenario or the use on HR-EI-MS spectra.

An alternative approach is offered by DeepEI, a deep-learning-based approach to fingerprinting using an ANN to extract the fingerprint from EI-MS spectra. The predicted fingerprint was obtained by concatenating six different fingerprints and removing bits

except those coding for features present in 10%–90% of compounds. For each bit of the fingerprint, an ANN was trained on NM EI-MS spectra from the NIST 2017. Compounds with molecular weight over 2000 Da and containing elements other than C, H, N, O, P, S, F, Cl, Br and Si were removed from the dataset, resulting in a training set of 184 874 spectra, making it more useful for general-purpose analysis of EI-MS data. A test set consisting of spectra from the MassBank database was cleaned like the NIST set, resulting in 13 000 spectra, of which 5619 were also present in the training set. Additionally, the performance was measured by querying the NIST 2017 database using the MassBank test set. Candidates were queried using a 5-Da mass window from an augmented NIST library; that is, spectra for compounds not present in the NIST library were generated using Neural Electron-Ionization Mass Spectrometry (NEIMS).⁴⁴ Furthermore, a synthetic library entirely simulated using NEIMS was used as the reference library, and the observed recall rates were reported.

Supervised fingerprinting methods (an overview of which is available in Table 1) are a developed research field. The first methods used fragmentation spectra as input for kernel-based methods to predict binary fingerprints. From this research CSI:FingerID and the SIRIUS suite emerged as widely used tools with a user-friendly GUI and useful documentation. However, CSI:FingerID has long computation times, a reason why alternative approaches are still being investigated.

Research in the field of fingerprinting is vibrant and competitive. The development of tools which offer faster computation and of tools that work with EI spectra are two main topics of research. Furthermore, the development of novel strategies based on the

possibilities offered by ANNs, like the use of embeddings as fingerprints, as seen with ADAPTIVE and Spec2Vec, is an interesting development. Another important consideration during the development of new tools to ensure its success is user friendliness, because the final user might not be interested in CLI or packages and might expect a more familiar GUI, which is reflected in the citation numbers of the individual tools.

2.2 | Substructure-prediction and embedding-based methods

Spec2Vec is a method adopted from natural language processing. Here, Huber et al. implemented the Word2Vec algorithm by treating MS² spectra as text documents. Peaks and neutral losses are interpreted as words. By training a neural network to predict the context of each word (peak or neutral loss) and the word from its context, latent embeddings are learned for all peaks. Each spectrum is then represented as the weighted average of its peak embeddings, and spectral similarity scores are computed in the embedding space. Spec2Vec was trained and evaluated on a large GNPS⁴⁰ dataset, where it achieved a higher correlation to structural similarity (Tanimoto score) when compared with standard cosine scores. The authors also demonstrate Spec2Vec's application to query spectra of unknown compounds (not present in the library), being able to identify structurally similar candidates in 60% of cases. Although the two available Spec2Vec models were trained on positive-mode ESI-MS² spectra, the authors note that it is possible to train the model on specific datasets and, due to the higher computational efficiency

TABLE 1 Overview of supervised fingerprinting methods discussed in this manuscript. The full set of publications evaluated for this review with structured information is provided as Table S3.

Reference (year)	Method/tool	ML method	Highlights
Heinonen et al. ²⁵ (2012)	FingerID	SVM	First use of high-resolution mass kernel in fingerprinting. Three models trained on in-house datasets with at most 528 compounds.
Duhrkop et al. ²⁰ (2015)	CSI:FingerID	SVM	Uses both the ESI-MS ² spectra and the fragmentation tree from SIRIUS for fingerprinting.
Baygi and Barupal ³⁸ (2024)	IDSL_MINT	Transformer	First use of transformer model for fingerprinting, trained on two sets from MoNA and GNPS.
Brouard et al. ³³ (2016)	IOKR	IOKR	Reportedly faster than CSI:FingerID, trained on spectra from GNPS.
Nguyen et al. ³⁴ (2019)	ADAPTIVE	MPNN + IOKR	Uses a custom fingerprint from the MPNN for increased specificity. The fingerprint is predicted from the spectra using IOKR.
Fan et al. ³⁵ (2020)	MetFID	ANN	Uses an ANN to predict the fingerprint from ESI-MS ² spectra, trained on MoNA and NIST 2017.
Ljончева et al. ⁴² (2022)	CSI:IOKR	IOKR	Fingerprinting of EI-MS spectra of silylated compounds, trained on small in-house library.
Goldman et al. ⁴¹ (2023)	MIST	Transformer	Uses annotated spectra and peak differences, fingerprint reconstructed through “unfolding.”
Ji et al. ⁴⁵ (2020)	DeepEI	ANN	Uses an ANN for fingerprinting of EI-MS spectra, trained on NIST 2017.

Abbreviations: ANN, artificial neural network; EI, electron ionisation; ESI, electrospray ionisation; IOKR, input output kernel regression; ML, machine learning; MPNN, message passing neural network; MS, mass spectrometry; SVM, support vector machine.

when compared to cosine scores, would allow to search large libraries with all-against-all matching. This is particularly interesting in the case of GC-MS where filtering by precursor m/z is not a reliable method to reduce the number of scores to calculate.

Another tool inspired by a text-mining algorithm is MS2LDA,⁴⁶ which can recognise biochemically relevant substructures, from MS² data and group spectra based on shared structural patterns. The algorithm is based on latent Dirichlet allocation and decomposes fragmentation spectra into blocks of co-occurring peaks and losses, which the authors call “Mass2Motifs” (similar to assigning topics in text documents). MS2LDA learns these substructure motifs in an unsupervised manner (without the need for metabolite annotations) and enables grouping of spectra of structurally similar metabolites, regardless of their spectral similarity. Doing so aids structural de novo annotation and functional classification of unidentified compounds. Furthermore, “Mass2Motifs” can be annotated by querying MotifDB,²² a database of annotated “Mass2Motifs,” which can further increase the speed of analysis by reducing the need for manual annotation.

Similar to MS2LDA, MESSAR⁴⁷ is an ML tool for substructure annotation. Instead of decomposing spectra into mass to motives, MESSAR consists of 8378 rules that map ESI-MS² features to substructures. The substructures were mined from the target and decoy GNPS libraries as built by Scheubert et al.⁴⁸ Mass spectra for the same structure were merged, and duplicated fragments were removed, resulting in a training dataset of 3146 spectra. Target substructures were identified from the fingerprint of CSI:FingerID by taking all bits except the ECFP4 bits and by iteratively breaking bonds of the compounds in the training dataset and keeping all CHNO substructures with more than five carbon and oxygen atoms. MESSAR was tested on two MoNA test sets, “MASSBANK” and “MASSBANK_CASMI,” with structures present in the training set removed. In total, 4743 rules which had at least five true positives in the MASSBANK test set were evaluated, and 2364 of the evaluated rules had a recall over 0.6, whereas only 463 had a recall lower than 0.2. Additionally, MESSAR was compared to CSI:FingerID on the MASSBANK_CASMI test set. Although MESSAR correctly annotated fewer structures correctly under the top three candidates, concatenating the results of both CSI:FingerID and MESSAR and removing duplicates produced more correct annotations than either tool.

The methods described in this section are quite dissimilar. Spec2Vec uses unsupervised methods to generate a “fingerprint,” which is more specific to the task of finding compounds in spectral libraries and results more in a stronger correlation of fingerprint similarity score with the structural similarity of compounds. Nevertheless, like supervised fingerprinting methods it might require retraining for compounds that are too dissimilar or are absent from the training dataset. Retraining is not necessary for MS2LDA, which finds correlation inside a given dataset. Nevertheless, MS2LDA is considerably slower than supervised methods and does not offer a fingerprint for searching compound databases, instead identifying relationships between compounds in the dataset. Both tools are fairly

popular in regard to the number of papers citing them, with MS2LDA being by far the most cited tool of the two. Furthermore, MS2LDA and MESSAR are available as a web server, whereas Spec2Vec is available only as a Python package.

3 | IN SILICO FRAGMENTATION

In silico fragmentation methods are methods which computationally predict a mass spectrum of a specific type for a given compound, which subsequently can be used for library-based annotation. Four approaches to generate in silico spectra are rule-based methods, combinatorial methods, ML-based methods and ab initio methods.¹² Rule-based methods rely on expert knowledge to curate a collection of rules which are used to predict fragments. Combinatorial methods generate fragments by iteratively splitting bonds in the molecule. Ab initio methods use quantum mechanical simulations to generate a mass spectrum but are constrained by their low throughput. ML-based methods, which are covered in this review, are diverse in their approaches; some try to apply ML to a part of the problem, for example, predicting the bond dissociation probability, whereas others try to directly predict a mass spectrum from an input. The considered methods have different strengths and weaknesses. The direct prediction methods require huge quantities of high-quality homogenous data for training. The trained models are fast but can predict spectra only with the resolution of the training dataset and are less interpretable. These disadvantages can be addressed by models that simulate the fragmentation events, thus resulting in easily interpretable spectra which allow arbitrary precision. Nevertheless, models that simulate fragmentation tend to be conceptually more complicated and might include time-consuming fragmentation steps, like CFM-ID.^{49–53}

In silico identification software (ISIS)⁵⁴ is a fragmentation tool designed to predict ESI-MS² spectra of lipids. ISIS simulates the fragmentation processes in a linear ion trap through a kinetic Monte Carlo approach. The algorithm was trained on a set of 22 lipids measured in positive mode using a normalised collision energy of 30%. Under the same experimental conditions, a test set of 45 lipids was measured. A genetic algorithm, with the similarity between the simulated spectrum and the measured one as fitness function, was used to find the optimal weights of the ANN used to predict the bond-cleaving energies for CID. The ISIS algorithm was tested on a subset of 18 399 lipids from the LIPID MAPS^{55–57} database (mass \leq 1100 Da, only CHNOPS atoms). ISIS was used to generate an artificial library with 300 replicates of each lipid in the LIPID MAPS subset. The recall rates are presented in Table S2. A general-purpose tool is CFM-ID,^{49–52,58} available as a web server, developed for the annotation of MS² spectra of compounds not present in spectral databases. Since its initial release, the web server has offered three functionalities: predicting spectra, assigning fragments and identifying compounds. In the case of spectral prediction, the input (a SMILES string, an InChIKey or a list of SMILES strings) is used to generate the 10, 20 and 40-eV ESI-CID-MS spectra in both positive and negative

modes. The fragment assignment functionality uses a given SMILES or InChIKey structure and an input spectrum to annotate peaks with possible fragments ranked by their computed probabilities. The compound identification functionality allows putative identification of MS² spectra (ideally CID spectra acquired at 10, 20 and 40-eV collision energy) used as input. The scoring is based on the Jaccard score. Central to the function of CFM-ID is its ability to predict mass spectra. The spectral prediction is achieved by systematically breaking bonds, generating possible fragments and assigning a probability to each fragmentation event.

More precisely, the fragmentation process is simulated by a homogeneous Markov-chain process. A vector of features that characterise each fragmentation is used as input to a linear function to predict fragmentation events. The computed fragmentation probability is used to estimate the intensity of a signal in the spectrum. The weights of the linear function were learned on ESI-MS² data from the Metlin database⁵⁹ obtained using an Agilent 6520 Q-TOF spectrometer.

Functionality of CFM-ID was extended in version 3.0⁵¹ adding a rule-based fragmentation module for lipids, which improves computational time and predictive performance for *in silico* fragmentation of lipids. Furthermore, a chemical class classification tool, an improved scoring function and the inclusion of metadata and experimental spectra for annotation were implemented. Version 4.0⁵² further improved the rule-based fragmentation module, covering additional molecule classes and changing the ring fragmentation modelling, which simplified the feature vector describing the fragmentation event. The changes resulted in an improvement in CFM-ID 2.0/3.0 of on average 26.7% better dot product when predicting [M + H]⁺ spectra and of 20.6% when predicting [M-H]⁻ spectra. The performance of CFM-ID was compared with SIRIUS 4 on the CASMI 2016 dataset. CFM-EI⁴⁹ was implemented in CFM-ID 2.0 to predict EI-MS spectra from SMILES or InChI string. CFM-EI simulates the fragmentation of the input compound using a fixed-length stochastic Markov chain, with the transition between discrete fragment states sampled from a set containing all possible fragments. Additionally, the simulation function was adjusted to handle isotope peaks and odd electron peaks. Furthermore, an ANN with a 20-neuron and a 4-neuron hidden layer with ReLU activation function was implemented in the CFM transition function. The CFM transition function was trained on 70-eV EI-MS spectra from the NIST/EPA/NIH mass spectral library.

An *in silico* method specifically developed for EI-MS is NEIMS.⁴⁴ NEIMS is an ANN that predicts EI mass spectra of small molecules. Precisely, NEIMS uses an ANN to predict NM EI mass spectra from extended circular fingerprints (ECFP). The ANN was trained on 240 942 NM EI mass spectra from the *mainlib* of the NIST 2017. The ANN was adjusted with reverse prediction to improve the prediction performance in the high-mass region, which has a greater impact on the match score than the lower-mass region. To evaluate the performance of the trained ANN, the *replib* of the NIST 2017 spectral library was used. An augmented library was constructed by removing the spectra contained in the *replib* from the *mainlib* and replacing it

with artificial spectra generated by the NEIMS ANN. The augmented library was queried using the spectra from the *replib*, and a top 10 recall rate of 85.5% was observed, which further increased to 91.7% when prefiltered using a 5-Da mass filter. Nevertheless, the performance is inferior to the use of the not-augmented NIST 2017 spectral library, which was shown to achieve a top 10 recall rate of 98.8% and a vastly superior performance at top 1 and top 5 recall rates. Additionally, the authors compared the performance of NEIMS to CFM-EI. To make it more comparable, NEIMS was retrained on the NIST 2014 spectra database, and an augmented library containing only predicted spectra was queried using the NIST 2014 *replib* to obtain top *n* recall rates. The use of ANNs to directly predict mass spectra has the considerable advantage of being faster than CFM-ID, consequently allowing the generation of vast synthetic NM EI-MS libraries.

Further exploration in the direct prediction of EI mass spectra was conducted by Zhang et al.,⁵⁹ who examined the possibility of using graph neural networks (GNN). A significant difference compared to NEIMS is that, instead of coding the molecular structure into an ECFP, a molecular graph is used as input for the graph convolutional network (GCN). In the molecular graph, non-hydrogen atoms are encoded as nodes, whereas the chemical bonds are represented as edges. Unlike ECFPs, which are designed for general use, the GCN learns an intermediate molecular representation specific for predicting EI-MS spectra. Additionally, this avoids the problem of different features having the same value in the fingerprint due to bit overflow or the loss of relevant information. The generated molecular graphs were used as input for the GCN, consisting of a feature extraction module, which extracts structural features from the molecular graphs. The extracted molecular features are successively used in the spectral prediction module to predict the EI-MS spectrum. The GCN was trained on 143 989 spectra from the NIST 2005 *mainlib*, with spectra of compounds contained in the *replib* removed. A test of 22 316 spectra from the NIST 2005 *replib* was used for testing, together with 7462 spectra from the MassBank database. As previously done by Wei et al., the recall rate was measured by querying augmented libraries. The augmented libraries were generated by removing the spectra of compounds in the test set and replacing them with spectra generated by the GCN.

Another approach to *in silico* fragmentation is presented by Goldman et al.⁶⁰ in the form of ICEBERG, a tool which predicts ESI-MS² spectra in a two-step process. In the first step the most probable fragments are predicted by the ICEBERG Generate model, and in the second step the fragments are assigned an intensity by the ICEBERG Score module. Unlike CFM-ID, which generates fragments by removing bonds from the molecular graph, ICEBERG, like MAGMA,⁶¹ generates fragments by removing atoms. The fragmentation is simulated through the ICEBERG Generate module. The ICEBERG Generate module assigns each atom a fragmentation probability and retains only the most probable fragments. The fragmentation is repeated until the third generation of fragments. At each iteration, a GNN is used to encode the graph of the root molecule and the graph of the current fragment. The graph embedding of the root molecule,

the atom embeddings of the current fragment and a context vector containing metadata are concatenated and used as an input for an MLP that calculates the fragmentation probability for each atom. The resulting fragments are assigned intensities by the second model called ICEBERG Score. ICEBERG Score is a set transformer that predicts the intensities of the fragments. Isotope patterns and hydrogen shifts are modelled by predicting multiple intensities. For each fragment, the intensity of the fragment and the intensity of the fragment with the addition and loss of up to six hydrogen masses are predicted.

ICEBERG was trained on the NIST 20 MS² library and the NPLIB1 dataset (a subset of the GNPS library used to train CANOPUS⁶²). Consensus spectra were used for training, which means it cannot predict spectra for different collision energies (unlike tools like CFM-ID and FIORA⁶³). The consensus spectra were generated by merging all spectra at different collision energies, merging peaks within 0.1 mDa. The resulting spectra were normalised and filtered to keep up to the 50 most intense peaks with normalised intensity over 3%. Furthermore, the mass of the adduct ion was subtracted from all MS² spectra. The resulting test sets were split into structurally disjoint 90/10% train-test splits and used to compare ICEBERG with SCARF⁶⁴, CFM-ID, NEIMS and the GNN proposed by Zhang et al.⁵⁹ To train the ICEBERG Generate module, MAGMa was used to enumerate fragments up to a fragmentation depth of three and filtered to retain only fragments present in the mass spectra. The authors found ICEBERG to outperform the closest contender (SCARF) on average cosine similarity between simulated and real spectra for the NPLIB1 test set. ICEBERG was outperformed by SCARF on the NIST 20 MS² test set. Nevertheless, when comparing the top *n* recall rates, ICEBERG clearly outperforms all competing tools on both test sets.

An alternative method, developed by Zhu and Jonas,⁶⁵ is RASSP, a model to predict EI-MS spectra. Two versions of the prediction model were developed, a version based on the prediction of sub-formulae (RASSP:FN) and one based on the prediction of atom

subsets (RASSP:SN). Both models comprise an enumeration step followed by the prediction of the probability distribution over the enumerated sub-formulae/subsets. RASSP:FN enumerates the sub-formulae by recursively taking the set-wise Cartesian product of the possible sub-formulae of one element with the sub-formulae over the rest of the molecule. To estimate the probability of a sub-formula, a GNN is used to encode the molecular information into a per-atom feature matrix. The per-atom feature matrix and the sub-formula are used to calculate a context vector, which is concatenated to the sub-formula and used as input to an ANN to obtain the sub-formula probabilities. The probabilities are further scaled with weights derived from the per-atom feature matrix.

The RASSP:SN version, instead, enumerates subsets by iteratively breaking chemical bonds up to a depth of three. The enumerated sub-formulae are supplemented by considering hydrogen rearrangements. For a given subset, it calculates an embedding by calculating the average of the per-atom feature of atoms present in the subset. The embedding is used instead of the context vector to calculate the probability of a sub-formula. The models were trained with minibatch SGD and L2 loss function on spectra with 100 438 EI-MS spectra from the NIST 2017 library (≤ 48 atoms, ≤ 4096 maximum unique sub-formulae, $\leq 12\ 288$ subsets). The performance was evaluated by querying an augmented NIST 2017 *mainlib* with the NIST 2017 *replib*.

In silico fragmentation tools (an overview is presented in Table 2) have developed rapidly due to the evolution of ML methods. ML methods can predict intensities, instead of barcode spectra, in a fraction of the time required by ab initio methods. These ML-based in silico fragmentation methods can be useful to expand existing MS libraries, as an alternative to fingerprints to rank candidate annotation and, in the case of methods that predict fragments, to help elucidate the fragmentation mechanism.

The available ML-based in silico tools serve different purposes. Direct prediction tools like NEIMS or the GNN reported by Zhang et al.⁵⁹ can predict vast libraries of NM EI-MS spectra but are less useful in cases where annotation of the fragments is required or for

TABLE 2 Overview of in silico fragmentation methods described in this manuscript. The full set of publications evaluated for this review with structured information is provided as Table S3.

Reference	Tool name	ML method	Highlights
Kangas et al. ⁵⁴	ISIS	ANN	Monte Carlo simulation of fragmentation of lipids; trained on a small in-house lipid library (22 lipids).
Wang et al. ⁵²	CFM-ID	ANN	Modelling of CID-MS ² fragmentation as stochastic Markov-chain process; trained on Metlin database; available as web server and Docker container.
Wang et al. ⁵²	CFM-EI	ANN	CFM-ID model for EI spectra; trained on NIST 2014, available as CLI.
	NEIMS	ANN	ANN-based direct prediction of NM-EI-MS spectra; trained on NIST 2014.
Zhang et al. ⁵⁹	GCN	GNN	Use of a GNN to predict NM-EI-MS spectra from the molecular graph; trained on NIST 2005.
Goldman et al. ⁶⁰	ICEBERG	GNN + transformer	Models ESI-MS ² in two steps: one model predicts most probable fragments; the second model predicts the intensities.
	RASSP:SN	GNN	Use of GNN to predict fragments of EI-MS trained on NIST 2017.

Abbreviations: ANN, artificial neural network; CID, collision-induced dissociation; CLI, command line interface; EI, electron ionisation; ESI, electrospray ionisation; GCN, graph convolutional network; GNN, graph neural network; ISIS, in silico identification software; ML, machine learning; MS, mass spectrometry; NM, nominal mass.

the prediction of ESI-MS spectra, where CFM-ID is *de facto* without an alternative. The development of alternative methods has great potential, as is seen with EI-MS spectra, where RASSP is capable of producing easily interpretable spectra with annotated fragments, like CFM-ID, at much higher speeds.

4 | DE NOVO METHODS

Previously described methods are of limited use in the identification of *unknown unknowns*, that is, mass spectra of compounds that are not present in any public databases. Due to the necessity of an input for *in silico* methods and the dependence on compound databases of fingerprinting methods, both methods will fail to identify a genuinely unknown compound. A possible workaround is the *de novo* generation of candidate structures. The field of *de novo* molecular generation has been of particular interest for novel drug development, where they allow exploring a targeted chemical space by generating molecules with desired properties on demand. These methods are based on models used in linguistics, for example, long-short-term memory (LSTM) networks that learn an intermediate representation from which the novel molecule is generated via perturbation. Two methods will be described, both addressing different problems. The first paper treats compound library expansion and training with augmented datasets. The second publication handles inverse spectral prediction by dividing it into two sub-problems.

Building upon their previous study,⁶⁶ Skinner et al.⁶⁷ developed DarkNPS, a method of *de novo* structure generation for the purpose of identifying novel psychoactive substances (NPS). The chemical space of NPS is peculiar because it is characterised by a small number of structural motifs and a limited number of chemical transformations used to synthesise NPS. Based on their previous work, the authors trained a gated recurrent unit model and an LSTM model with augmented SMILES datasets with different degrees of augmentation. The augmentation of the datasets was achieved by including non-canonical SMILES, that is, SMILES obtained by varying the path by which the molecule is traversed to generate the SMILES string. The non-augmented training data consisted of 1753 unique NPS structures contained within the HighResNPS⁶⁸ dataset, with another 194 used as the test set. An LSTM model with an augmentation of factor 100 of the training set was selected based on the higher percentage of valid SMILES generated and the five metrics described in their previous work.⁶⁶ An artificial compound library was generated by sampling SMILES from the trained model and removing invalid SMILES and known NPS, resulting in an artificial library of 62 354 novel NPS. It was observed that some NPS appeared multiple times, which was hypothesised to correlate with the probability of the compound appearing on the grey market. Out of the molecules in the training set, 90.7% appeared at least once in the artificial library, with the 18 molecules not present showing a significantly lower similarity to any compound in the training set. After the 18 dissimilar entries were removed from the test set, 93.1% of compounds in the

TABLE 3 Overview of selected ML-based annotation tools.

Method	Typology	Input	Output
Heinonen et al. ²⁵	FP	ESI-MS ² spectra	Binary fingerprint
Brouard et al. ³³	FP	ESI-MS ² spectra	Binary fingerprint
ADAPTIVE ³⁴	FP	ESI-MS ² spectra	Molecular fingerprint
CANOPUS ⁶²	FP	ESI-MS ² spectra	Classy fire classes
SIRIUS ²³	AS	ESI-MS ² spectra	Molecular formula, molecular fingerprint, predicted structure
MIST	FP	ESI-MS ² spectra with annotated peaks	Molecular fingerprint
Ljioncheva et al. ⁴²	FP	ESI-MS ² spectra of TMS derivatives	Molecular fingerprint
MetFID ^{35,76}	FP	ESI-MS ² spectra	Molecular fingerprint
DeepEI ⁴⁵	FP	ESI-MS ² spectra	Molecular fingerprint
Spec2Vec ²¹	FP	ESI-MS ² , EI-MS ^a spectra	Embedding
MS2LDA-MotifDB ^{22,46}	FP	ESI-MS ² spectra	Annotated and unannotated mass spectral patterns
ISIS ⁵⁴	ISF	Molecular structure	ESI-MS ² spectrum
CFM-ID ^{49,50,52,53}	ISF	Molecular structure	HR-CID-MS ² , HR-EI-MS spectrum
NEIMS ⁴⁴	ISF	Molecular structure	Nominal mass EI-MS spectrum
Zhang et al. ⁵⁹	ISF	Molecular structure	Nominal mass EI-MS spectrum
DarkNPS ⁶⁷	DCG	Training set	Library of novel compounds
MSNovelist ⁶⁹	DCG	ESI-MS ² spectrum	Candidate structure

Abbreviations: AS, annotation suite; CID, collision-induced dissociation; DCG, *de novo* compound generation; EI, electron ionisation; ESI, electrospray ionisation; FP, fingerprinting; HR, high resolution; ISF, *in silico* fragmentation; ISIS, *in silico* identification software; ML, machine learning; MS, mass spectrometry; TMS, trimethylsilyl.

The column “method” provides references and the name of the published tool when available.

^aAlthough advantages of Spec2Vec for EI were mentioned in the manuscript, the user would need to retrain the model to use it on EI-MS spectra.

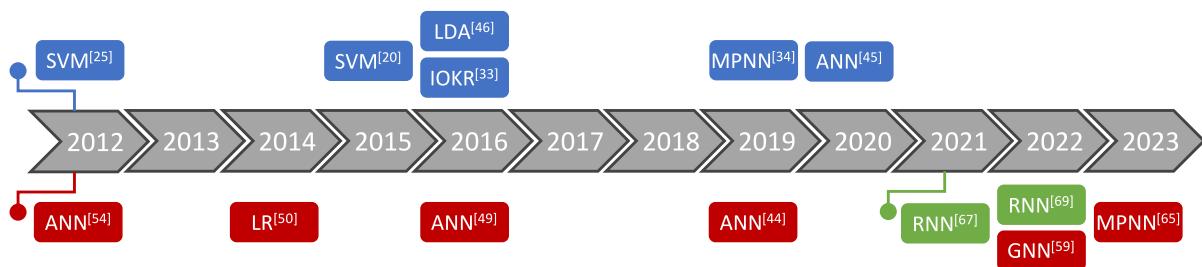


FIGURE 3 Graphical timeline of ML (machine learning) methods used in compound annotation. Fingerprinting methods are colour coded in blue, in silico fragmentation methods in red and de novo methods in green. The trend towards the adoption of artificial neural networks (ANN) can be observed for both in silico fragmentation and fingerprinting methods. The more recent de novo methods use recurrent neural networks (RNN). ANNs have been repeatedly investigated for in silico fragmentation tools, where in recent years new architectures, such as graph neural networks (GNN) and message passing neural networks (MPNN), have been implemented. In contrast, fingerprinting methods are dominated by kernel methods, for example, support vector machines (SVM) and input output kernel regression (IOKR), which still show competitive performance, even though methods based on ANN have been investigated in recent times. [Color figure can be viewed at wileyonlinelibrary.com]

test set were predicted by the model. Using only the accurate mass with a search window of ± 10 ppm, a top 1 recall of 33%, a top 3 recall of 48% and a top 10 recall of 72% were observed. Furthermore, based on the results obtained on a subset of 79 compounds with MS^2 data, the combination of CFM-ID⁵⁰ and DarkNPS with the sampling frequency of a molecule as ranking was analysed. A top 3 recall rate of 53% for the model alone, 1% for CFM-ID and 70% for the combination of CFM-ID and DarkNPS were observed.

Another de novo approach is the direct prediction of the molecular structure from mass spectra, as done by MSNovelist.⁶⁹ The method breaks the inverse spectral problem into two parts: the prediction of a fingerprint and the prediction of a SMILE from a fingerprint. The fingerprint is predicted using SIRIUS.¹⁹ The fingerprint and the molecular formula, predicted using CSI:FingerID²⁰ or manually given by the user, are used as input for an LSTM model to predict the SMILES of the compound. The ANN was trained on 1 232 184 compounds collected from HMDB,^{70–74} COCONUT and DSSTox.⁷⁵ The trained model was tested on a test set of 3863 compounds from the GNPS database. For each spectrum in the test set, the 128 highest-scoring SMILES were retrieved with a top 128 recall rate of 45% and a top 1 recall rate of 25%—for comparison, CSI:FingerID reaches a top 1 recall rate of 39% when searching against a database. The performance on the GNPS dataset was compared with a naive generation model, which lacked the fingerprint input. The naive model had a top 1 recall rate of 17% and retrieved 31% of all correct structures. Additionally, MSNovelist was tested on 127 positive ion mode spectra from the CASMI 2016 challenge, reaching a top 1 recall rate of 26% and a total correct recall rate of 57% (compared to, respectively, 24% and 52% for the naive model).

Of the discussed methods, de novo methods are the newest and as such the less investigated and less mature. The novelty of the models is only in part responsible for the low number of publications; more importantly the requirement of high quantities of data limits the development of de novo annotation methods. To overcome the limited data availability, data augmentation and breaking down

into multiple more manageable sub-problems have been deployed. In both cases it was shown that using adequate strategies, it is possible to utilise de novo methods to augment compound libraries and to predict the identity of a compound from MS^2 spectra. The relative novelty of the methods might limit their application right now, but the potential of expanding compound libraries and predicting molecular formulae directly from mass spectra will ensure further interest in the scientific community.

5 | CONCLUSION AND OUTLOOK

In recent years, significant progress has been made in using ML to annotate MS data in metabolomics. Multiple annotation approaches have been proposed (an overview is presented in Table 3).

The approaches studied in this article were grouped into three categories: fingerprinting, in silico fragmentation and de novo methods. Fingerprinting and in silico fragmentation can be considered as more developed fields and are, as such, dominated by several well-established tools, for example, the SIRIUS suite or CFM-ID, which are subject to steady improvement. The evolution of applied algorithms over time, as seen in Figure 3, shows trends similar to the general development of ML based approaches. Initially, methods were based on more traditional ML approaches, which in the case of fingerprinting are still highly relevant, followed by the adoption of ANNs and, especially in recent years, of novel specialised ANN architectures. More recently, GNNs were adopted in both, fingerprinting and in silico fragmentation approaches, and RNNs for de novo methods.

Fingerprinting methods have seen a broad use of classical ML methods and recently some utilisation of ANNs, which are capable of predicting whole fingerprints, resulting in fast inference. The problem with using ANNs is the requirement of considerable quantities of data for training and the quality of the mass spectra. Furthermore, the similarity of the training data to the application data determines how reliable the results are.

A similar scenario emerges for in silico fragmentation that has seen the emergence of ML-based direct prediction methods, which can generate huge synthetic spectral libraries due to their computational speed. Direct in silico fragmentation methods are based on ANNs, which require high quantities of data and strongly profit from homogeneous standardised datasets. The availability of vast, standardised NM libraries (NIST, Wiley) has resulted in promising algorithms for NM EI-MS. Although vast MS libraries are also available for ESI-MS, the high data heterogeneity has inhibited the development of direct in silico fragmentation algorithms so far. More complex physics-inspired algorithms like CFM-ID and RASSP, which predict molecular fragments, have shown the possibility to be trained with less data (CFM-ID) and the ability to predict spectra at arbitrary resolution. However, it is still necessary to use computationally intensive methods such as CFM-ID or quantum mechanics-based methods to simulate HR MS spectra, ESI-MS² spectra or compound spectra too dissimilar from the training set. Additionally, ML methods can facilitate the identification of “unknown unknowns,” that is, compounds that have not been previously described in the literature and are therefore absent from databases. This is possible either by identifying similar compounds in databases or by generating possible structures from the spectral data. More precisely, ANNs open novel possibilities such as de novo methods which can complement already-established tools by expanding their area of application to compounds not contained in compound libraries. The adoption of these methods is limited by their resource-intensive nature and the need for further fine-tuning to better fit the compounds of interest. Nonetheless, the continuous improvement in computer hardware and cloud computing, the availability of good-quality training data and the integration of ML methods into existing user-friendly packages might result in the popularisation of de novo methods for non-targeted MS annotation.

In summary, ML approaches are already substantially benefitting non-targeted MS analyses, and a variety of well-established tools exist and are being constantly improved. Naturally, a single universally applicable tool does not exist. Depending on the task and question being investigated, different tools are viable. Weaknesses of state-of-the-art tools and novel possibilities offered by ML approaches like ANNs are the main drivers of development. We observed that incorporating chemical knowledge into the architecture of a tool leads to improved performance, in addition to making models more interpretable for users. An additional contribution to the popularity of a tool is the ease of use, where a tendency of more popular tools offering a GUI can be observed. There is reason for some optimism as the accuracy, capabilities and accessibility of the methods are continuously improving. Simplifying and automating the annotation of non-targeted MS data is of great interest due to the potential of increased throughput, reproducibility and reduced costs. ML might catalyse a wider diffusion of non-targeted MS by suppressing costs and reducing the high time consumption of highly qualified labour. Beneficial for the development of the field are, on the one hand, the availability of high-quality training data, that is, annotated spectra with well-curated metadata, and, on the other hand, the use of well-defined test sets that allow a fair comparison between the different methods. Benchmarking

and comparison with existing methods is non-trivial and has been discussed by multiple authors (e.g., de Jonge et al.⁷⁷ and Hoffmann et al.⁷⁸). For a good comparison the test set needs to be representative of the tool use case. When multiple tools are compared, the test set should be structure disjoint from the training set. The lack of standardised, publicly available test sets can result in metrics which are comparable only within a single study. We are optimistic about the development of the field, as we observe an increased effort of authors to ensure fair and representative benchmarking.

Furthermore, we observe a steady growth of open-access mass spectral libraries, which hopefully will accelerate the progress and development in the field. ML approaches for compound annotation are already performing well. More competitive, dynamic approaches, as well as fruitful collaborative efforts, may result in the establishment of standard testing procedures, methods and datasets, which in turn will further improve ML-based approaches.

AUTHOR CONTRIBUTIONS

Francesco F. Russo: Writing—original draft; writing—review and editing; investigation. **Yannek Nowatzky:** Investigation; writing—original draft. **Carsten Jaeger:** Conceptualization; writing—review and editing. **Maria K. Parr:** Conceptualization; writing—review and editing; supervision. **Phillipp Benner:** Conceptualization; writing—review and editing; supervision. **Thilo Muth:** Writing—review and editing; conceptualization. **Jan Liseć:** Conceptualization; writing—review and editing; supervision; visualization.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/rcm.9876>.

ORCID

Maria K. Parr  <https://orcid.org/0000-0001-7407-8300>
Jan Liseć  <https://orcid.org/0000-0003-1220-2286>

REFERENCES

1. Hollender J, Schymanski EL, Singer HP, Ferguson PL. Nontarget screening with high resolution mass spectrometry in the environment: Ready to go? *Environ Sci Technol*. 2017;51(20):11505–11512. doi:[10.1021/acs.est.7b02184](https://doi.org/10.1021/acs.est.7b02184)
2. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nat Biotechnol*. 2000;18(11):1157–1161. doi:[10.1038/81137](https://doi.org/10.1038/81137)
3. Kim S, Chen J, Cheng T, et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res*. 2021;49(D1):D1388–D1395. doi:[10.1093/nar/gkaa971](https://doi.org/10.1093/nar/gkaa971)

4. Gross H Jr. *online resource (XXV, 968 pages 664 illustrations, 201 illustrations in color)*. 3rded. Springer International Publishing: Imprint: Springer; 2017:1.
5. Junot C, Fenaille F, Colsch B, Becher F. High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. *Mass Spectrom Rev*. 2014;33(6):471-500. doi:[10.1002/mas.21401](https://doi.org/10.1002/mas.21401)
6. Prompramate S, Chen Y, Chen Y-PP. In: Chen Y-PP, ed. *Bioinformatics Technologies*. Springer; 2005:117-153.
7. Cios KJ, Kurgan LA, Reformat M. Machine learning in the life sciences. *IEEE Eng Med Biol Mag*. 2007;26(2):14-16. doi:[10.1109/EMB.2007.335579](https://doi.org/10.1109/EMB.2007.335579)
8. Zhou Z-H. *SpringerLink, Machine Learning*. 1sted. Springer Singapore: Imprint: Springer; 2021.
9. Hastie T, Friedman JH, Tibshirani R. *The elements of statistical learning: data mining, inference, and prediction*. Springer; 2009. doi:[10.1007/978-0-387-84858-7](https://doi.org/10.1007/978-0-387-84858-7)
10. Hofmann T, Schölkopf B, Smola A. 2006.
11. Liebal UW, Phan ANT, Sudhakar M, Raman K, Blank LM. Machine learning applications for mass spectrometry-based metabolomics. *Metabolites*. 2020;10(6):243. doi:[10.3390/metabo10060243](https://doi.org/10.3390/metabo10060243)
12. Nguyen DH, Nguyen CH, Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: A review with emphasis on machine learning approaches. *Brief Bioinform*. 2018;20(6):2028-2043. doi:[10.1093/bib/bby066](https://doi.org/10.1093/bib/bby066)
13. Petrick LM, Shomron N. AI/ML-driven advances in untargeted metabolomics and exposomics for biomedical applications. *Cell Rep Phys Sci*. 2022;3(7):100978. doi:[10.1016/j.xcrp.2022.100978](https://doi.org/10.1016/j.xcrp.2022.100978)
14. Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: Applications of deep learning in metabolomics. *Comput Struct Biotechnol J*. 2020;18:2818-2825. doi:[10.1016/j.csbj.2020.09.033](https://doi.org/10.1016/j.csbj.2020.09.033)
15. Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *J Chem*. 2021;13(1):84. doi:[10.1186/s13321-021-00558-4](https://doi.org/10.1186/s13321-021-00558-4)
16. Ji H, Xu Y, Lu H, Zhang Z. Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification. *Anal Chem*. 2019;91(9):5629-5637. doi:[10.1021/acs.analchem.8b05405](https://doi.org/10.1021/acs.analchem.8b05405)
17. de Jonge NF, Louwen JJR, Chekmeneva E, et al. MS2Query: Reliable and scalable MS2 mass spectra-based analogue search. *Nat Commun*. 2023;14(1):1752. doi:[10.1038/s41467-023-37446-4](https://doi.org/10.1038/s41467-023-37446-4)
18. Steffen A, Kogej T, Tyrchan C, Engkvist O. Comparison of molecular fingerprint methods on the basis of biological profile data. *J Chem Inf Model*. 2009;49(2):338-347. doi:[10.1021/ci800326z](https://doi.org/10.1021/ci800326z)
19. Duhrkop K, Fleischauer M, Ludwig M, et al. SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*. 2019;16(4):299-302. doi:[10.1038/s41592-019-0344-8](https://doi.org/10.1038/s41592-019-0344-8)
20. Duhrkop K, Shen H, Meusel M, Rousu J, Bocker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A*. 2015;112(41):12580-12585. doi:[10.1073/pnas.1509788112](https://doi.org/10.1073/pnas.1509788112)
21. Huber F, Ridder L, Verhoeven S, et al. *PLoS Comput Biol*. 2021;17.
22. Rogers S, Ong CW, Wandy J, Ernst M, Ridder L, van der Hooft JJJ. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. *Faraday Discuss*. 2019;218(0):284-302. doi:[10.1039/C8FD00235E](https://doi.org/10.1039/C8FD00235E)
23. Bocker S, Duhrkop K. *J Chem*. 2016;8:5.
24. Ludwig M, Duhrkop K, Bocker S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics*. 2018;34(13):i333-i340. doi:[10.1093/bioinformatics/bty245](https://doi.org/10.1093/bioinformatics/bty245)
25. Heinonen M, Shen HB, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*. 2012;28(18):2333-2341. doi:[10.1093/bioinformatics/bts437](https://doi.org/10.1093/bioinformatics/bts437)
26. Kondor R, Jebara T. *A Kernel between Sets of Vectors*. Vol. 1; 2003.
27. Wolf S, Schmidt S, Muller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*. 2010;11(1):148. doi:[10.1186/1471-2105-11-148](https://doi.org/10.1186/1471-2105-11-148)
28. Kim S, Chen J, Cheng T, et al. PubChem 2023 update. *Nucleic Acids Res*. 2023;51(D1):D1373-D1380. doi:[10.1093/nar/gkac956](https://doi.org/10.1093/nar/gkac956)
29. Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:[10.1093/nar/28.1.27](https://doi.org/10.1093/nar/28.1.27)
30. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci*. 2019;28(11):1947-1951. doi:[10.1002/pro.3715](https://doi.org/10.1002/pro.3715)
31. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. 2023;51(D1):D587-D592. doi:[10.1093/nar/gkac963](https://doi.org/10.1093/nar/gkac963)
32. Kanehisa M. *Post-Genome Informatics*. 2000. doi:[10.1093/oso/9780198503279.001.0001](https://doi.org/10.1093/oso/9780198503279.001.0001)
33. Brouard C, Shen H, Duhrkop K, d'Alche-Buc F, Bocker S, Rousu J. Fast metabolite identification with input output kernel regression. *Bioinformatics*. 2016;32(12):i28-i36. doi:[10.1093/bioinformatics/btw246](https://doi.org/10.1093/bioinformatics/btw246)
34. Nguyen DH, Nguyen CH, Mamitsuka H. ADAPTIVE: LeArning DAta-dePendenT, concise molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*. 2019;35(14):i164-i172. doi:[10.1093/bioinformatics/btz319](https://doi.org/10.1093/bioinformatics/btz319)
35. Fan Z, Alley A, Ghaffari K, Ressom HW. MetFID: Artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics*. 2020;16(10):104. doi:[10.1007/s11306-020-01726-7](https://doi.org/10.1007/s11306-020-01726-7)
36. Laponogov I, Sadawi N, Galea D, Mirnezami R, Veselkov KA. ChemDistiller: An engine for metabolite annotation in mass spectrometry. *Bioinformatics*. 2018;34(12):2096-2102. doi:[10.1093/bioinformatics/bty080](https://doi.org/10.1093/bioinformatics/bty080)
37. Schymanski EL, Ruttkies C, Krauss M, et al. Critical assessment of small molecule identification 2016: Automated methods. *J Chem*. 2017;9(1):22. doi:[10.1186/s13321-017-0207-1](https://doi.org/10.1186/s13321-017-0207-1)
38. Baygi SF, Barupal DK. IDSL_MINT: A deep learning framework to predict molecular fingerprints from mass spectra. *J Chem*. 2024;16(1):8. doi:[10.1186/s13321-024-00804-5](https://doi.org/10.1186/s13321-024-00804-5)
39. Böcker S, Dürkop K. Fragmentation trees reloaded. *J Chem*. 2016;8(1):8. doi:[10.1186/s13321-016-0116-8](https://doi.org/10.1186/s13321-016-0116-8)
40. Wang M, Carver JJ, Phelan VV, et al. *Nat Biotechnol*. 2016;34:828-837.
41. Goldman S, Wohlwend J, Strazar M, Haroush G, Xavier RJ, Coley CW. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Machine Intell*. 2023;5(9):965-979. doi:[10.1038/s42256-023-00708-3](https://doi.org/10.1038/s42256-023-00708-3)
42. Ljioncheva M, Stepisnik T, Kosjek T, Dzeroski S. Machine learning for identification of silylated derivatives from mass spectra. *J Chem*. 2022;14(1):62. doi:[10.1186/s13321-022-00636-1](https://doi.org/10.1186/s13321-022-00636-1)
43. Qiu F, Lei Z, Sumner LW. MetExpert: An expert system to enhance gas chromatography-mass spectrometry-based metabolite identifications. *Anal Chim Acta*. 2018;1037:316-326. doi:[10.1016/j.aca.2018.03.052](https://doi.org/10.1016/j.aca.2018.03.052)
44. Wei JN, Belanger D, Adams RP, Sculley D. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Cent Sci*. 2019;5(4):700-708. doi:[10.1021/acscentsci.9b00085](https://doi.org/10.1021/acscentsci.9b00085)
45. Ji H, Deng H, Lu H, Zhang Z. Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Anal Chem*. 2020;92(13):8649-8653. doi:[10.1021/acs.analchem.0c01450](https://doi.org/10.1021/acs.analchem.0c01450)
46. van der Hooft JJ, Wandy J, Barrett MP, Burgess KE, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics.

Proc Natl Acad Sci U S A. 2016;113(48):13738-13743. doi:[10.1073/pnas.1608041113](https://doi.org/10.1073/pnas.1608041113)

47. Liu Y, Mrzic A, Meysman P, et al. MESSAR: Automated recommendation of metabolite substructures from tandem mass spectra. *PLoS ONE*. 2020;15(1):e0226770. doi:[10.1371/journal.pone.0226770](https://doi.org/10.1371/journal.pone.0226770)

48. Scheubert K, Hufsky F, Petras D, et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun*. 2017;8(1):1494. doi:[10.1038/s41467-017-01318-5](https://doi.org/10.1038/s41467-017-01318-5)

49. Allen F, Pon A, Greiner R, Wishart D. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Anal Chem*. 2016;88(15):7689-7697. doi:[10.1021/acs.analchem.6b01622](https://doi.org/10.1021/acs.analchem.6b01622)

50. Allen F, Pon A, Wilson M, Greiner R, Wishart D. CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res*. 2014;42(W1):W94-W99. doi:[10.1093/nar/gku436](https://doi.org/10.1093/nar/gku436)

51. Djoumbou-Feunang Y, Pon A, Karu N, et al. CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. *Metabolites*. 2019;9(4):9. doi:[10.3390/metabo9040072](https://doi.org/10.3390/metabo9040072)

52. Wang F, Allen D, Tian S, et al. CFM-ID 4.0 – a web server for accurate MS-based metabolite identification. *Nucleic Acids Res*. 2022;50(W1):W165-W174. doi:[10.1093/nar/gkac383](https://doi.org/10.1093/nar/gkac383)

53. Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart DS. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem*. 2021;93(34):11692-11700. doi:[10.1021/acs.analchem.1c01465](https://doi.org/10.1021/acs.analchem.1c01465)

54. Kangas LJ, Metz TO, Isaac G, et al. *In silico* identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*. 2012;28(13):1705-1713. doi:[10.1093/bioinformatics/bts194](https://doi.org/10.1093/bioinformatics/bts194)

55. Sud M, Fahy E, Cotter D, et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res*. 2007;35:D527-D532. doi:[10.1093/nar/gkl838](https://doi.org/10.1093/nar/gkl838)

56. Fahy E, Sud M, Cotter D, Subramaniam S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res*. 2007;35(Web Server):W606-W612. doi:[10.1093/nar/gkm324](https://doi.org/10.1093/nar/gkm324)

57. Conroy MJ, Andrews RM, Andrews S, et al. LIPID MAPS: Update to databases and tools for the lipidomics community. *Nucleic Acids Res*. 2024;52(D1):D1677-D1682. doi:[10.1093/nar/gkad896](https://doi.org/10.1093/nar/gkad896)

58. Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*. 2014;11(1):98-110. doi:[10.1007/s11306-014-0676-4](https://doi.org/10.1007/s11306-014-0676-4)

59. Zhang BJ, Zhang J, Xia Y, Chen P, Wang B. Prediction of electron ionization mass spectra based on graph convolutional networks. *Int J Mass Spectrom*. 2022;475:116817. doi:[10.1016/j.ijms.2022.116817](https://doi.org/10.1016/j.ijms.2022.116817)

60. Goldman S, Li J, Coley CW. Generating molecular fragmentation graphs with autoregressive neural networks. *Anal Chem*. 2024;96(8):3419-3428. doi:[10.1021/acs.analchem.3c04654](https://doi.org/10.1021/acs.analchem.3c04654)

61. Stefan Verhoeven LR. *marijnsanders*; 2017.

62. Duhrkop K, Nothias LF, Fleischauer M, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*. 2021;39(4):462-471. doi:[10.1038/s41587-020-0740-8](https://doi.org/10.1038/s41587-020-0740-8)

63. Nowatzky Y, Russo F, Liseck J, Kister A, Reinert K, Muth T, Benner P. 2024.

64. Goldman S, Bradshaw J, Xin J, Coley C. *Advances in Neural Information Processing Systems*. Vol. 36; 2023:48548-48572.

65. Zhu RL, Jonas E. Rapid approximate subset-based spectra prediction for electron ionization-mass spectrometry. *Anal Chem*. 2023;95(5):2653-2663. doi:[10.1021/acs.analchem.2c02093](https://doi.org/10.1021/acs.analchem.2c02093)

66. Skinnider MA, Stacey RG, Wishart DS, Foster LJ. Chemical language models enable navigation in sparsely populated chemical space. *Nat Machine Intell*. 2021;3(9):759-+. doi:[10.1038/s42256-021-00368-1](https://doi.org/10.1038/s42256-021-00368-1)

67. Skinnider MA, Wang F, Pasin D, et al. A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nat Machine Intell*. 2021;3(11):973-+. doi:[10.1038/s42256-021-00407-x](https://doi.org/10.1038/s42256-021-00407-x)

68. Mardal M, Andreasen MF, Mollerup CB, et al. HighResNPS.Com: An online crowd-sourced HR-MS database for suspect and non-targeted screening of new psychoactive substances. *J Anal Toxicol*. 2019;43(7):520-527. doi:[10.1093/jat/bkz030](https://doi.org/10.1093/jat/bkz030)

69. Stravs MA, Duhrkop K, Bocker S, Zamboni N. MSNovelist: De novo structure generation from mass spectra. *Nat Methods*. 2022;19(7):865-870. doi:[10.1038/s41592-022-01486-3](https://doi.org/10.1038/s41592-022-01486-3)

70. Wishart DS, Jewison T, Guo AC, et al. HMDB 3.0--the human metabolome database in 2013. *Nucleic Acids Res*. 2013;41(Database issue):D801-D807. doi:[10.1093/nar/gks1065](https://doi.org/10.1093/nar/gks1065)

71. Wishart DS, Tzur D, Knox C, et al. HMDB: The human metabolome database. *Nucleic Acids Res*. 2007;35(Database):D521-D526. doi:[10.1093/nar/gkl923](https://doi.org/10.1093/nar/gkl923)

72. 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. doi:[10.1093/nar/gkab1062](https://doi.org/10.1093/nar/gkab1062)

73. Wishart DS, Knox C, Guo AC, et al. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Res*. 2009;37(Database):D603-D610. doi:[10.1093/nar/gkn810](https://doi.org/10.1093/nar/gkn810)

74. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res*. 2018;46(D1):D608-D617. doi:[10.1093/nar/gkx1089](https://doi.org/10.1093/nar/gkx1089)

75. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of open natural products database. *J Chem*. 2021;13(1):13. doi:[10.1186/s13321-020-00478-9](https://doi.org/10.1186/s13321-020-00478-9)

76. Gao S, Chau HYK, Wang K, Ao H, Varghese RS, Ressom HW. Convolutional neural network-based compound fingerprint prediction for metabolite annotation. *Metabolites*. 2022;12(7):12. doi:[10.3390/metabo12070605](https://doi.org/10.3390/metabo12070605)

77. de Jonge NF, Mildau K, Meijer D, et al. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics*. 2022;18(12):103. doi:[10.1007/s11306-022-01963-y](https://doi.org/10.1007/s11306-022-01963-y)

78. Hoffmann MA, Kretschmer F, Ludwig M, Bocker S. MAD HATTER correctly annotates 98% of small molecule tandem mass spectra searching in PubChem. *Metabolites*. 2023;13(3):13. doi:[10.3390/metabo13030314](https://doi.org/10.3390/metabo13030314)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Russo FF, Nowatzky Y, Jaeger C, et al. Machine learning methods for compound annotation in non-targeted mass spectrometry—A brief overview of fingerprinting, *in silico* fragmentation and de novo methods. *Rapid Commun Mass Spectrom*. 2024;38(20):e9876. doi:[10.1002/rcm.9876](https://doi.org/10.1002/rcm.9876)