Metabolomics (2020) 16:104
https://doi.org/10.1007/s11306-020-01726-7

ORIGINAL ARTICLE

MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation

Ziling Fan[1] · Amber Alley[2] · Kian Ghaffari[2] · Habtom W. Ressom[2]

Received: 8 May 2020 / Accepted: 19 September 2020

Abstract

Introduction Metabolite annotation is a critical and challenging step in mass spectrometry-based metabolomic profiling. In a typical untargeted MS/MS-based metabolomic study, experimental MS/MS spectra are matched against those in spectral libraries for metabolite annotation. Yet, existing spectral libraries comprise merely a marginal percentage of known compounds.

Objective The objective is to develop a method that helps rank putative metabolite IDs for analytes whose reference MS/MS spectra are not present in spectral libraries.

Methods We introduce MetFID, which uses an artificial neural network (ANN) trained for predicting molecular fingerprints based on experimental MS/MS data. To narrow the search space, MetFID retrieves candidates from metabolite databases using molecular formula or m/z value of the precursor ions of the analytes. The candidate whose fingerprint is most analogous to the predicted fingerprint is used for metabolite annotation. A comprehensive evaluation was performed by training MetFID using MS/MS spectra from the MoNA repository and NIST library and by testing with structure-disjoint MS/MS spectra from the NIST library, the CASMI 2016 dataset, and in-house MS/MS data from a cancer biomarker discovery study.

Results We observed that training separate models for distinct ranges of collision energies enhanced model performance compared to a single model that covers a wide range of collision energies. Using MetaboQuest to retrieve candidates, MetFID prioritized the correct putative ID in the first place rank for about 50% of the testing cases. Through the independent testing dataset, we demonstrated that MetFID has the potential to improve the accuracy of ranking putative metabolite IDs by more than 5% compared to other tools such as ChemDistiller, CSI:FingerID, and MetFrag.

Conclusion MetFID offers a promising opportunity to enhance the accuracy of metabolite annotation by using ANN for molecular fingerprint prediction.

Keywords Metabolite identification · Artificial neural network · Molecular fingerprint · Metabolomics

Abbreviations

ANN	Artificial neural network
CASMI	Critical Assessment of Small Molecule Identification
GC	Gas chromatography
GNPS	Global Natural Products Social Molecular Networking
HCD	Higher-energy C-trap dissociation
HCC	Hepatocellular carcinoma
HMDB	Human Metabolite Database
IT	Ion trap
LC	Liquid chromatography
MS	Mass spectrometry
MS/MS	Tandem MS
Q-TOF	Quad-time-of-flight
METLIN	Metabolite and Chemical Entity Database

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11306-020-01726-7) contains supplementary material, which is available to authorized users.

✉ Habtom W. Ressom
hwr@georgetown.edu

[1] Department of Biochemistry and Molecular & Cellular Biology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC, USA

[2] Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Suite 173, Building D, 4000 Reservoir Road NW, Washington, DC 20057, USA

Published online: 30 September 2020

MoNA MassBank of North America
NIST National Institute of Standards and Technology

1 Introduction

Metabolites are positioned farthest downstream in the hierarchical regulatory processes of a biological system and are, therefore, generally recognized to be the most correlated with biological phenotype (Guijas et al. 2018; Johnson and Gonzalez 2012; Zhang et al. 2017). Thus, metabolomics has become an effective strategy in exploring biomarkers for diagnosis, treatment, and mechanistic studies (Guijas et al. 2018). A commonly employed analytical platform for metabolomic studies includes liquid chromatography coupled with mass spectrometry (LC–MS), a technology that has risen in prominence in the field of metabolomics due to its ability to analyze a sizable number of metabolites with a limited amount of biological material compared to other platforms. Nonetheless, metabolite annotation of analytes detected by LC–MS still remains an immense challenge (Patti et al. 2012; Scheubert et al. 2013).

Various properties of an analyte detected by LC–MS can be utilized to accurately identify a metabolite. These properties include m/z, retention time, spectral fragmentation, etc (2013). In particular, the MS/MS spectrum of an analyte is an invaluable resource for metabolite annotation and structural elucidation as it unveils a considerable amount of information about the substructure of the compound (Li et al. 2018). Different types of mass spectrometers, such as collision cell or ion trap, and different experimental setups (e.g. different collision energies) are used to generate MS/MS spectra. The most common approach to utilize MS/MS spectra for metabolite annotation is through spectral matching against those in libraries, such as HMDB, METLIN, GNPS, etc (Wang et al. 2016; Guijas et al.2018; Wishart et al. 2018). However, only a minute fraction of known compounds has its curated reference MS/MS spectra in these spectral libraries. Thus, the ability to annotate 'known unknowns' through MS/MS spectral matching is largely limited (Matsuda 2016). Such a limitation necessitates developing novel methods to close the gap between existing experimental spectra and spectra absent from libraries.

Recently, multiple computational methods, such as rule-based, combinatorial-based, and machine learning-based, have been developed to improve MS/MS-based metabolite annotation (Nguyen et al. 2019; Kim et al. 2019; Horai et al. 2010; Meringer and Schymanski 2013; Scheubert et al. 2013; Gerlich and Neumann 2013). We applied a machine-learning method to improve the accuracy of spectral matching of experimental MS/MS spectra to those in spectral libraries (Zhou et al. 2010). Since 2012, other papers have reported using machine learning for compound fingerprint prediction as an intermediate step for metabolite annotation (Nguyen et al. 2018; Li et al. 2020; Brouard et al. 2016; Duhrkop et al. 2015; Heinonen et al. 2012). A compound's fingerprint is represented by a vector with each element of the vector being a binary value indicating the presence or absence of a singular property or substructure (e.g. aromatic ring, aldehyde, etc.) of the compound, which can be determined by a range of tools (O'Boyle et al. 2008). A trained machine learning model receives an MS/MS spectrum of an unknown compound as input and predicts its fingerprint as output. The fingerprint similarity between the unknown compound and its candidates retrieved from a compound database is calculated via a scoring function. Lastly, fingerprint similarity scores are ranked and the candidate with the highest rank is used for annotation. Previously, a variety of machine learning methods were pursued to predict molecular fingerprints of analytes. For example, SIMPLE is a tool that predicts the fingerprint of an unknown compound based on its MS/MS spectrum via a regression model. An advantageous feature of SIMPLE is its specialized regression equation which can consider pair-wise peak interactions (Nguyen et al. 2018). Brouard et al. adopted the Input Output Kernel Regression method to learn the mapping between MS/MS and molecular structure (Li et al. 2020). Li et al. designed SF-matching using Random Forest to predict if a certain spectrum can be generated from a given structure (Brouard et al. 2016). Another tool, FingerID, predicts the fingerprint based on different classes of features extracted for different MS kernels. FingerID's successor, CSI:FingerID, applied fragmentation trees with a multiple kernel learning method to enhance classifier performance (Duhrkop et al. 2015; Heinonen et al. 2012). However, a significant disadvantage that all of these methods share is that they involve multiple models where a single model can only predict one digit or one block of the fingerprint. This approach neglects the latent hidden relationship between the digits within a fingerprint. To address this, we recently proposed a multi-class and multi-label task formulation using an artificial neural network (ANN) to be able to predict a vector representing the whole compound fingerprint at one time (Fan et al. 2019).

In this paper, we introduce MetFID, which builds on our previously proposed method to perform compound fingerprint prediction and rank candidate compounds retrieved from compound databases for metabolite annotation. Furthermore, we performed a comprehensive evaluation of MetFID against other tools, such as MetFrag,

ChemDistiller, and CSI:FingerID, using MS/MS spectra from the NIST library, the CASMI 2016 dataset, and in-house MS/MS data from a cancer biomarker discovery study.

2 Method

2.1 Workflow overview

The MetFID workflow consists of two phases: training and prediction, as depicted in Fig. 1. In the training phase, an ANN is trained using MS/MS data from spectral libraries to map the relationship between MS/MS spectra (input to ANN) and compound fingerprints (output of ANN). In the prediction phase, an MS/MS spectrum of an analyte is fed into the trained ANN to predict the molecular fingerprint of the compound. Databases (e.g. MetaboQuest, HMDB, PubChem, etc.) are searched by the m/z value of the precursor ion or by the molecular formula of the compound to retrieve potential candidate compounds. Each candidate compound is then given a similarity score, which is calculated by a similarity scoring function. In the subsequent sections, dataset preparation and the steps of the workflow are discussed.

2.2 Dataset

The core task in MetFID is training an ANN using MS/MS spectra of known compounds (input) and their pre-calculated molecular fingerprints (output). The trained ANN is then used to predict an unknown compound's fingerprint by using its experimental MS/MS. Although fragmentation patterns are largely determined by molecular structures, different fragmentation techniques and experimental setups employed

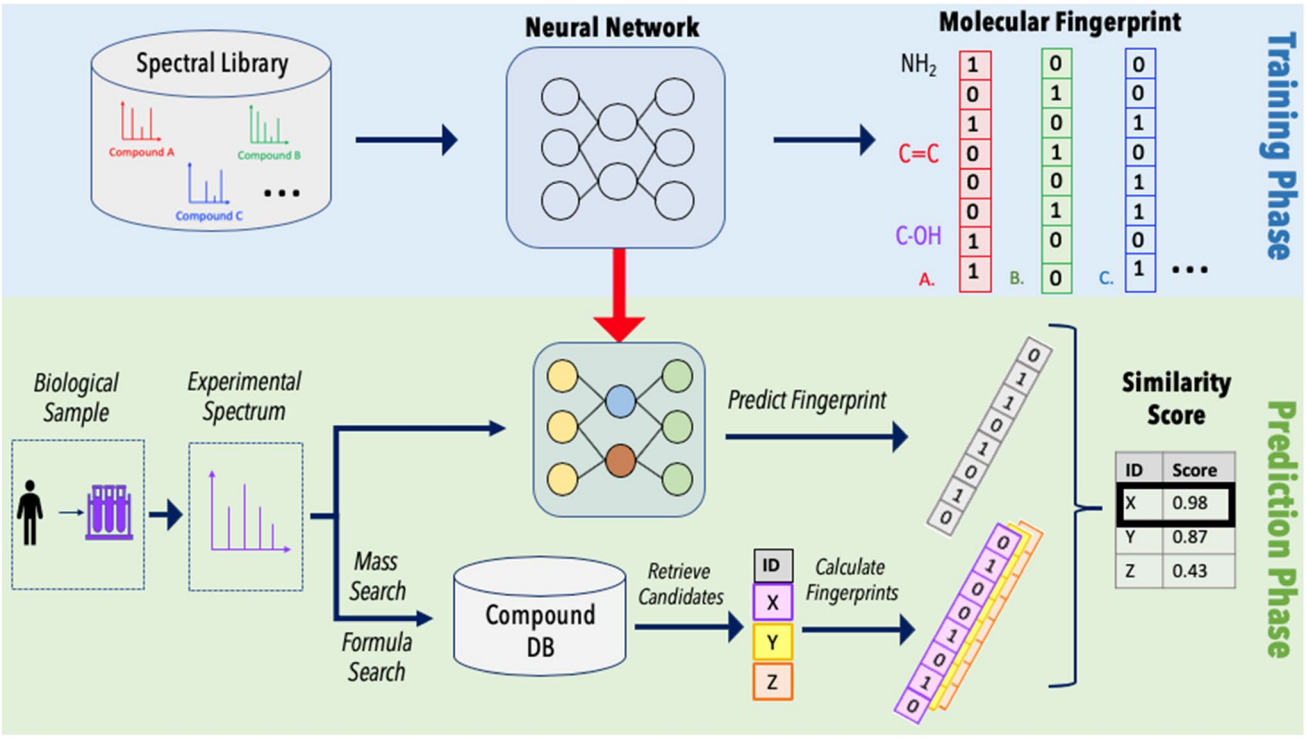

Fig. 1 Workflow of MetFID

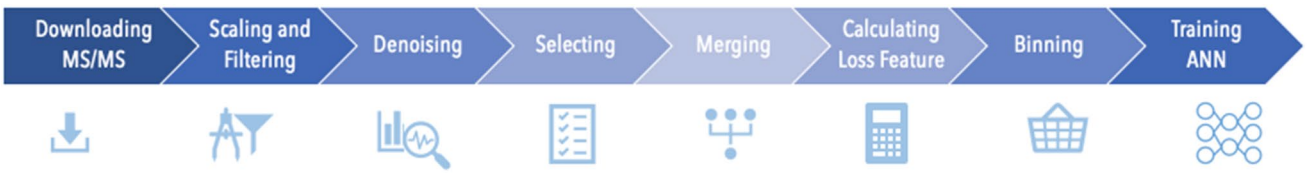

Fig. 2 Steps involved in processing spectra prior to training an ANN

by mass spectrometers may generate different fragmentation patterns for the same compound. Therefore, we developed a pipeline, shown in Fig. 2, to obtain a relatively homogeneous dataset for model training.

2.2.1 Downloading training MS/MS spectra

We downloaded MS/MS spectra from the following libraries available in the MoNA repository (https://mona.fiehnlab.ucdavis.edu/): Vaniya/Fiehn Natural Products Library, GNPS, RIKEN PlaSMA, MassBank, and Fiehn HILIC. In addition, we obtained the NIST 17 library from one of NIST's MS/MS library distributors. For the remainder of this paper, we will refer to MS/MS data downloaded from MoNA as MoNA spectra and those downloaded from NIST 17 as NIST 17 spectra.

2.2.2 Scaling and filtering

We scaled the peak intensities such that all intensity values are between 0 and 100. Spectra that consisted of fewer than five peaks with relative intensity above 2% were removed (Duhrkop et al. 2015).

2.2.3 Denoising

We recognize that spectra with substantial noise and peaks very close to the baseline may affect the training of ANNs. To address this, we first identified peaks with the highest intensities that also had greater mass than the precursor mass. Then, we removed peaks whose intensity values were smaller than that of the highest intensity, subsequently removing all peaks with intensity lower than 10.

2.2.4 Selecting spectra

To make the training MS/MS data as homogeneous as possible, we selected those that meet the following criteria: (a) acquired in the positive mode; (b) acquired via instruments that use collision cell (HCD, QqQ, and Q-TOF) or ion trap (IT); (c) fall within a mass range of 100 to 1010 Da; and (d) consist of only one of two forms of adducts (H+ or NH4+). After this step, we obtained 11,784 MoNA spectra representing 5667 compounds and 122,481 NIST 17 spectra representing 10,730 compounds. Thus, a total of 15,228 unique compounds were considered, combining the spectra from MoNA and NIST 17. The IDs of all training and testing spectra used in this study are provided in Supplementary Table S1.

2.2.5 Merging spectra

We merged multiple MS/MS spectra with different collision energies acquired from the same compound using InChIKey as a compound identifier. Peaks from these various MS/MS spectra were merged into one peak if their m/z difference was 0.1 or less and the mean of these peaks' intensities became the new merged peak's intensity. Note that all testing cases were single collision energy spectra, without merging those acquired at different collision energies (Duhrkop et al. 2015).

2.2.6 Calculating loss features

After fragmentation by MS/MS, only charged product ions can be detected and the non-charged (neutral) segments of the resulting fragment will not appear in the MS/MS spectrum. We refer to the latter as loss features. The m/z values of the loss features for each peak were determined by calculating the difference in m/z between precursor ion's m/z and m/z value of each peak, thus, we used the peak intensity as a proxy for the loss feature (Heinonen et al. 2012).

2.2.7 Binning

We binned the m/z range of each MS/MS spectrum into prespecified bins, which indicate continuous integer m/z values, and calculated the accumulated intensities within each bin as feature values. We removed bins that consisted of all 0's across all the spectra in the training set. This binning method has been applied previously (Nguyen et al. 2019; Wang et al. 2019). The distribution and statistics of the features in the training dataset are provided in the Supplementary Figure S1 and Table S2.

2.3 Determining fingerprints

The molecular fingerprints of all compounds in the training set were determined using OpenBabel (O'Boyle et al. 2008). Specifically, MACCS, FP3, and FP4 fingerprints were determined and assembled into a vector comprising of 528 binary entries.

2.3.1 Training ANN

ANN was trained to learn the relationship between spectral pattern and compound fingerprint. A review on the use of machine learning methods for compound fingerprint prediction has been previously reported (Nguyen et al. 2019).

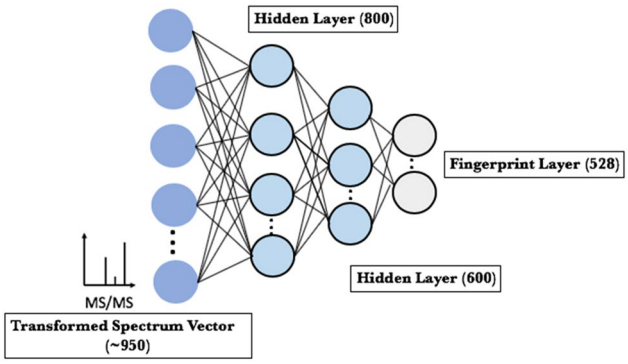

Fig. 3 Architecture of the ANN

2.4 Architecture of ANN

Figure 3 is a depiction of the architecture of the ANN we developed using the Keras python package on top of the TensorFlow backend. First, we converted the MS/MS spectra using binning into vectors to be used as input for the ANN. Two hidden layers of the ANN were created, following the input layer. Rectified linear unit (RLU) was used as an activation function. We chose this activation function to alleviate gradient vanishing problems. The benefit of ANN is that a single model, with multiple nodes in the output layer, can be used to predict all digits of a compound fingerprint without the need to build individual models for each digit. Therefore, prior to the fingerprint layer, sigmoid activation function and binary cross-entropy loss function were adopted instead of the combination of softmax activation and categorical cross-entropy loss function typically used for tasks involving single labels. The proposed ANN architecture allows the prediction of the entire fingerprint at once in contrast to traditional digit-by-digit approaches (Nguyen et al. 2018; Li et al. 2020; Brouard et al. 2016; Duhrkop et al. 2015; Heinonen et al. 2012). For ANN training, the 'Adam' optimizer was used. Epochs and batch size were 30 and 100, respectively.

2.5 Candidate retrieval and ranking

Metabolite databases and tools, such as METLIN, PubChem, MetaboQuest, and HMDB, offer different options for searching compounds, the most common of which are mass-based and formula-based search methods. In this study, we chose MetaboQuest (https://omicscraft.com/MetaboQuest) to retrieve a candidate list based on the m/z values of the precursors (10 PPM, H^+ adduct) corresponding to the MS/MS spectra. Furthermore, we used the molecular formulae, provided in our downloaded MS/MS spectra, to search candidates in PubChem, which consists of the largest number of compounds among currently available public compound databases (Kim et al. 2019). We used PubChem to retrieve candidates for the purposes of comparing the performance of our method to other tools. As formula is required to query PubChem, we assumed that the formula for each spectrum in the testing set was known (formula is provided in the library we downloaded). In practical cases, where the formulae are unknown for experimental MS/MS spectra, putative compound formulae can be calculated based on m/z values of precursor ions using tools such as Sirius (Duhrkop et al. 2019).

We used the python package PubChemPy to query the PubChem database based on molecular formula. Once the candidate list was obtained from PubChem, OpenBabel was used to calculate a 528-digit fingerprint for each candidate compound. The similarity scores between the predicted fingerprint of an unknown compound and the fingerprints of the candidates were computed using the formula in Eq. (1). The similarity scores range from 0 to 1. The candidate with the fingerprint that possesses the highest similarity to the predicted fingerprint was used to determine the predicted annotation of the unknown compound.

$$\delta\left(\boldsymbol{f}',\boldsymbol{f}\right) = 1 - \frac{\sum |\boldsymbol{f}' - \boldsymbol{f}|}{n} \tag{1}$$

where $\boldsymbol{f}'$ and $\boldsymbol{f}$ are the predicted and known fingerprint vectors, respectively, and n denotes the length of these vectors.

2.6 Evaluation

For evaluation, we used previously unseen compounds for testing in a structure-disjoint setup (i.e. no training compound which has the same first part of the InChIKey is included in the testing set). The MS/MS spectra from these compounds were processed in the same way as the training dataset, with the exception of merging. Thus, the performance of the ANN is evaluated using individual MS/MS spectra (without merging those acquired by different collision energies) in terms of its ability to perform both fingerprint prediction and metabolite annotation.

To keep a balance between precision and recall, we adopted accuracy and F1 scores to measure fingerprint prediction performance for MetFID. Let TN, TP, FN, and FP denote the numbers of true negatives, true positives, false negatives, and false positives, respectively. The precision equation is TP/TP + FP and the recall equation is TP/TP + FN. Equations (2) and (3) were used to compute the macro-average accuracy and F1 score, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

To assess the metabolite annotation accuracy, MetaboQuest or PubChem was used to retrieve the candidate list. The candidates were then ranked according to the similarity score calculated in Eq. (1). To compute the metabolite annotation accuracy for the top k predicted candidates, we searched for the true compound ID in the top k predicted list (Nguyen et al. 2018; Li et al. 2020; Brouard et al. 2016; Duhrkop et al. 2015; Heinonen et al. 2012).

3 Results

3.1 Training dataset and model separation based on collision energy

To accomplish the task of predicting fingerprints of unknown compounds through their MS/MS spectra, an ANN needs to be trained via a training set to map the hidden relationship between MS/MS fragmentation patterns and molecular substructures/properties. In our prepared ANN training dataset, most of the compounds had more than one spectrum generated by different instrument types or different collision energies, which produced some spectra with different fragmentation patterns for the same compound. We adopted a commonly used method that merges all spectra generated from one compound into one spectrum (Duhrkop et al. 2015; Heinonen et al. 2012). In addition, we explored whether or not training separate models using different collision energy ranges and instrument types would enhance model performance.

To illustrate the difference and similarity in the MS/MS patterns of a compound acquired by different instruments and settings, we extracted as examples three MS/MS spectra from the NIST 17 library for cycloheptylamine (CID:2899) representing different experimental setups, as shown in Fig. 4. The spectrum in Fig. 4a has similar pattern with the spectrum in Fig. 4b. These were all generated using collision cell MS (QqQ and HCD) with similar collision energies (14 eV and 17 eV). Although the spectrum in Fig. 4a was generated using the same instrument type as the spectrum in Fig. 4c, their fragmentation patterns are very different due to significant differences in the collision energies used for fragmenting the precursor ions (14 eV vs. 38 eV). Higher collision energies tend to generate more peaks in total as well as more peaks with smaller m/z values.

For further investigation, we performed cluster analysis of spectra corresponding to 916 compounds (more than 30 spectra per compound) in the NIST 17 library to characterize the training dataset we prepared. Principle component

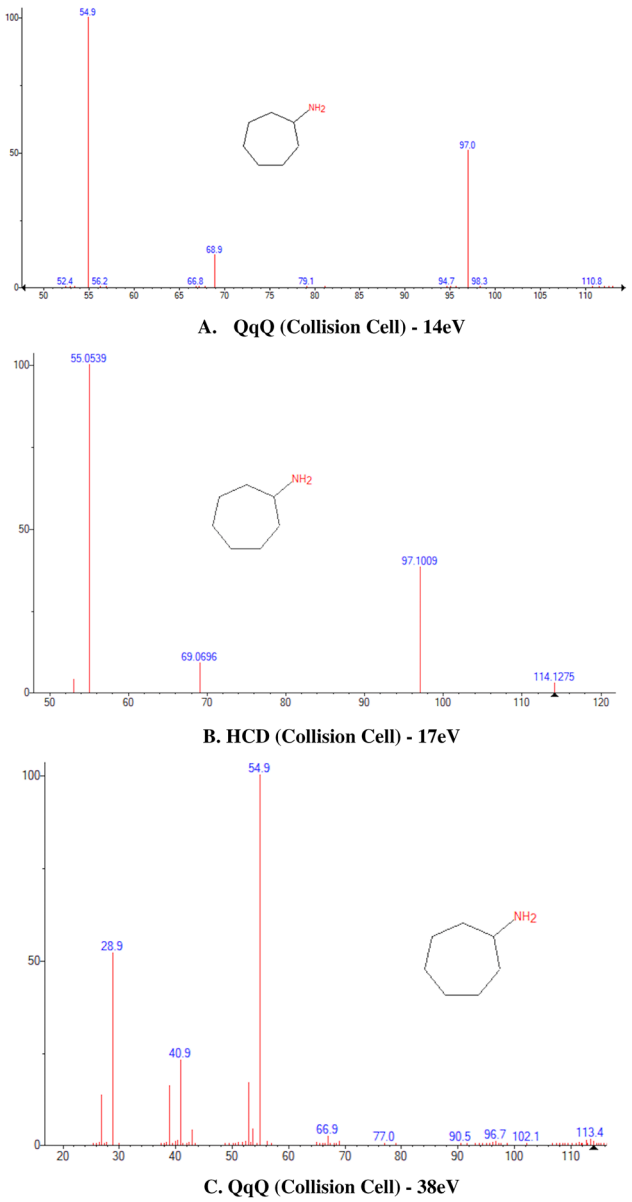

Fig. 4 Examples of MS/MS for Cycloheptylamine acquired by different machine types and collision energy. *HCD* higher-energy C-trap dissociation, *QqQ* triple quadrupole

analysis (PCA) was used for dimensional reduction and K-means was used to group the spectra into three clusters. Figure 5 shows two examples of our K-means results where InChIKey of the compound is shown on the top of each figure. In the figures, each dot represents a spectrum, dot shape denotes instrument type, and dot labels indicate the collision energy used to generate that spectrum. We learned from these results the following: (1) as expected, spectra generated using similar collision energies tend to cluster together while, in contrast, spectra with vastly different collision energies are located far away from each other; this phenomenon is consistent with our observation

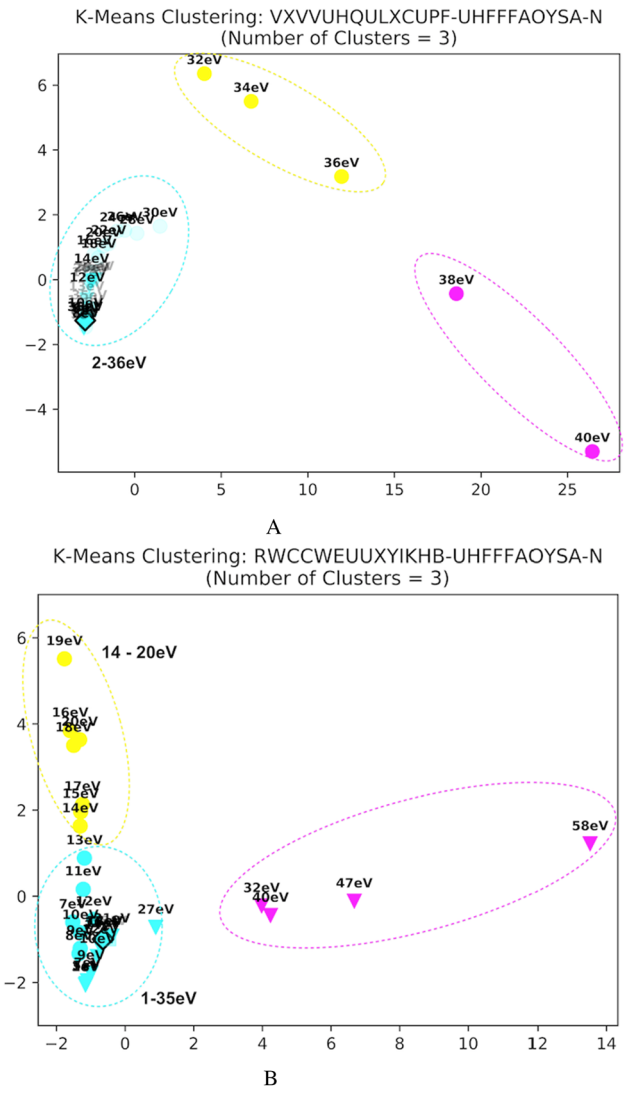

Fig. 5 Examples of clusters of MS/MS spectra for two compounds. Each dot represents an MS/MS spectrum and its collision energy is labeled (filled circle = ESI QqQ; filled down triangel = HCD; filled diamond = Ion Trap)

from individual spectra (discussed above), (2) through manually examining all cluster analysis results, we found that 30 eV was a reasonable cutoff for high and low collision energies, and (3) in terms of instrument type, spectra acquired by HCD and QqQ, which belong to the collision cell type, clustered together. Interestingly, despite the fact that IT adopts a completely different fragmentation mechanism compared to collision cell instruments, spectra acquired by IT also clustered with spectra acquired by collision cell.

Based on the clustering results, we conducted two sets of testing to determine if it was necessary to train separate ANNs for different ranges of collision energy and for different instrument types. Separate ANNs were trained for low energy ($\leq$30 eV), high energy ($\geq$30 eV), and high & low energies combined. Specifically, we selected 6677 compounds generated using HCD MS from the NIST 17 library that had at least one corresponding spectrum with $\leq$30 eV collision energy and one corresponding spectrum with $\geq$30 eV collision energy. To train a low energy model, all spectra of a compound generated using $\leq$30 eV were merged to represent each of the 6677 compounds with one merged MS/MS spectrum. The same procedure was done for high energy and combined energy models. We trained all ANNs of different collision energies using approximately the same number of training MS/MS data and tested through structure-disjoint five-fold cross validation. As shown in Table 1a, the accuracy of the combined model is slightly lower than the accuracy of the low collision energy model when testing it with low energy testing spectra. Moreover, F1 scores of the combined model are lower compared to the F1 scores of the low and high enrergy collision energy models. The result supports the idea that the selection of training dataset helps enhance the ANN's prediction performance.

Furthermore, we trained three separate models using spectra generated by IT, HCD, and IT and HCD combined. Consistent with the cluster analysis result where IT spectra

Table 1 Evaluation of MetFID's compound fingerprint prediction performance via five-fold cross-validation of ANNs trained for different collision energy ranges (A) and instrument types (B) based on MS/MS spectra acquired from 6677 compounds

A				
Training spectra	$\leq$ 30 eV	$\geq$30 eV	All collision energy	
Testing spectra	$\leq$ 30 eV	$\geq$30 eV	$\leq$ 30 eV	$\geq$30 eV
Accuracy	94%	94%	92%	94%
F1	69%	71%	58%	69%
B				
Training spectra	IT	HCD	Combined HCD&IT	
Testing spectra	IT	HCD	IT	HCD
Accuracy	94%	93%	94%	94%
F1	71%	68%	74%	68%

clustered with HCD spectra, the result of this analysis shows that combing IT and HCD does not affect the prediction performance compared to separate models (Table 1b).

3.2 Evaluation of MetFID's performance compared to other tools

The final step in MetFID is to perform metabolite annotation by using a similarity score computed by comparing the fingerprint of each candidate with the predicted fingerprint. The candidate with the highest similarity score is used for metabolite annotation. In this section, we evaluate different models for ranking metabolite candidates obtained by mass-based search using MetaboQuest. In this set of testing, we did not use spectra from MoNA for ANN training, because a large portion of spectra from MoNA do not provide collision energy information. Among the spectra described in Sect. 2.2, we randomly selected 1500 MS/MS spectra representing 1500 compounds for testing. The remaining spectra, which exclude the 1500 testing (structure-disjoint) compounds, were used for ANN training.

Table 2 shows the evaluation results from different ANNs trained under different scenarios. We used a similar number of training spectra for model training to avoid the effect brought on by training dataset size (the number of training spectra for low, high, and combined models is 9671, 9108 and 9560, respectively). The accuracy in the table refers to the percentage of testing cases in which the correct metabolite appears in the top k of the ranked candidate list. The results show that MetFID successfully ranked the correct identification in more than 50% of cases. Consistent with our previous testing, separate models outperform the combined model by a small margin. We also evaluated MetFID assuming that the formulae of the unknown compounds are known. For these datasets, we observed that the use of formula information helps shrink the average length of candidate lists from 18 to 9 candidates, thereby enhancing the annotation accuracy, as illustrated in Table 2.

To compare MetFID with MetFrag (Ruttkies et al. 2016), ChemDistiller (Laponogov et al. 2018), and CSI:FingerID (Duhrkop et al. 2019), we trained MetFID with nearly all training spectra obtained from NIST 17 and MoNA excluding testing spectra that represent 482 compounds and guarantee structure-disjoint testing set. We tested MetFrag and ChemDistiller using these testing spectra whose candidates were retrieved from the PubChem database through the PubChemPy python package and assuming that the formulae of testing compounds are known (formula for each compound is provided in NIST17 library). MetFrag applies the combinatorial method to predict an MS/MS spectrum for each candidate and ranks candidates based on their similarity to the spectrum of the unknown compound. ChemDistiller combines the idea of MetFrag and CSI:FingerID by calculating a composite score for each candidate (Laponogov et al. 2018). As shown in Table 3, MetFID outperforms these tools by a good margin. We did not include CSI:FingerID in this evaluation, because NIST 17 spectra were involved in the training of the current version of CSI:FingerID, thus, they cannot be used as a structure-disjoint testing set. Instead, we used CASMI 2016 testing set to compare MetFID vs.

Table 2 Evaluation of MetFID in ranking metabolite candidates from MetaboQuest by different models based on testing MS/MS spectra acquired from 1500 compounds

	Low energy model		High energy model		Combined model	
Accuracy	93%		92%		92%	
F1	64%		64%		59%	
Ranking		Formula known		Formula known		Formula known
Top 1	50%	57%	52%	56%	46%	51%
Top 3	73%	76%	76%	77%	70%	71%
Top 5	80%	84%	84%	84%	77%	83%

Table 3 Comparison of MetFID with MetFrag, ChemDistiller, and CSI:FingerID in ranking metabolite candidates using NIST 17 or CASMI 2016 as testing sets

Rank	NIST 17 testing			CASMI 2016 testing			
	MetFrag	ChemDistiller	MetFID	MetFrag*	CSI:FingerID* (2016)	CSI:FingerID# (2019)	MetFID
Top 1	15%	21%	27%	12%	28%	39%	38%
Top 3	17%	32%	40%	–	55%	–	67%
Top 10	25%	44%	69%	–	70%	75%	72%

*The CASMI 2016 testing results for CSI:FingerID (2016) and MetFrag were obtained from a previous report (Schymanski et al. 2017). The accuracy for the top ranked candidate only was reported in testing MetFrag via CASMI 2016

#The CASMI 2016 testing results for CSI:FingerID (2019) were obtained from a previous report (Duhrkop et al. 2019)

CSI:FingerID (Schymanski et al. 2017). Note that the dataset used for training ChemDistiller includes NIST 14 spectra and, thus, a subset of our testing dataset may have been used for training. Detailed information on ChemDistiller and MetFrag can be found in Supplementary Table S6.

Finally, we selected from the CASMI 2016 benchmark dataset MS/MS spectra acquired in the positive mode from 127 compounds (Schymanski et al. 2017). To have a fair comparison with other participants, we made structure-disjoint testing and adopted the candidate lists for each testing compound previously provided (Schymanski et al. 2017). As shown in Table 3, MetFID correctly identified 38% (48 compounds) of testing cases in CASMI 2016 dataset and outperformed MetFrag and CSI:FingerID (2016). Duhrkop et al. did a re-evaluation for CASMI 2016 and reported an improved performance that CSI:FingerID (2019) correctly identified 50 compounds out of 127 compounds using SIRIUS 4, in which CSI:FingerID is integrated (Duhrkop et al. 2019). Also, Li et al. reported similar performance for SF-matching and CSI:FingerID using the CASMI 2016 testing dataset (Li et al. 2020). We obtained very close performance compared to CSI:FingerID and SF-Matching when MetFID was evaluated on the CASMI 2016 testing dataset.

3.3 Evaluation of MetFID using MS/MS spectra from a cancer biomarker discovery study

We previously reported 18 metabolites that were significantly altered in hepatocellular carcinoma (HCC) tumors vs. adjacent normal liver tissues (Ferrarini et al. 2019). These metabolites were identified by combining gas chromatography coupled with mass spectrometry (GC–MS) and LC–MS data acquired by analysis of liver tissues from 40 HCC patients. For 98 analytes, MS/MS spectra were acquired using the ACQUITY UPLC system coupled to a Synapt G2-Si QTOF-MS (Waters Corporation, Milford, MA, USA) with collision energy ranging from 10 to 30 eV. To predict the metabolite annotation of these analytes, we considered the low energy MetFID model that was trained using about 9000 NIST 17 spectra and about 1000 MoNA spectra with collision energy ≤ 30 eV. The trained MetFID was then used to rank candidate analytes retrieved by MetaboQuest based on the 98 MS/MS spectra (Supplementary Table S3). Because the training set did not include two compounds that were previously reported (Ferrarini et al. 2019), MetFID's prediction of the metabolite IDs for these two analytes can be considered structure-disjoint evaluation. The annotation for these two metabolites was previously performed by spectral matching using MetaboQuest, METLIN, and CEU Mediator (Ferrarini et al. 2019). Supplementary Table S4 presents putative metabolite IDs for these two analytes. The putative IDs were obtained by searching the m/z values of the two analytes against multiple compound databases using MetaboQuest. The putative IDs that were reported previously based on spectral matching are highlighted in the table and are considered as ground-truth information for our evaluation (Ferrarini et al. 2019). Through mass-based search of m/z values of precursor ions via MetaboQuest, we retrieved 12 candidates for m/z = 758.5697 and 6 candidates for m/z = 258.1112 with ≤ 10 ppm of mass tolerance. Interestingly, all candidates retrieved for each m/z had the same PPM, thus, it was impossible to further rank them by PPM. The results in Supplementary Table S4 show that MetFID successfully ranked the two metabolite IDs that were reported previously at the top of the candidate lists (Ferrarini et al. 2019). However, for the analyte with m/z = 758.5697, four out of the 12 candidates had the same highest score. This is due to high similarity in structure among the four candidates as shown in Supplementary Table S5. This renders a challenging task for our method to distinguish structurally similar candidates, thus, additional ranking criteria are needed to enhance the performance of MetFID.

4 Discussion

The proposed MetFID trains an ANN to capture the hidden mapping relationships between MS/MS fragmentation patterns and molecular substructures/properties of analytes to accomplish the task of predicting fingerprints of unknown compounds based on measured MS/MS spectra. There are different tools for fingerprint calculation and we adopted a combination of FP3, FP4 and MACCS, which were calculated using OpenBabel (Fan et al. 2019). These sets of fingerprints are all created based on SMARTS pattern, which is a language for describing molecular functional groups and molecular patterns. This type of fingerprint is also commonly used for substructure searching among chemical compounds. Our results, along with other machine learning-based method results, which use a similar fingerprint type (Nguyen et al. 2018; Li et al. 2020; Brouard et al. 2016), demonstrate that the mapping relationship between compound fragmentation pattern and SMARTS pattern fingerprint can be captured. Duhrkop et al. pointed out that a selected subset of extended-connectivity fingerprints (ECFPs) can be sufficiently learned from training data (Duhrkop et al. 2019). We evaluated the performance of MetFID using topology-based fingerprints calculated from RDKit (G. RDKit: Open-Source Cheminformatics Software. 2016). However, the ANN achieved much lower fingerprint prediction performance using topology-based fingerprints compared to the fingerprints adopted by MetFID. Thus, further investigation is needed to thoroughly evaluate different sets of fingerprints to determine the set of fingerprints that

leads to the highest performance on fingerprint prediction and metabolite annotation.

Similarity scores calculated by MetFID to rank candidates based on predicted fingerprints can have very similar values, as shown in Supplementary Table S3. Such scores pose a challenge in selecting candidates with confidence for further investigation. To address this, methods such as the target-decoy approach (Gupta et al. 2011) can be used to help estimate the false discovery rates based on the similarity scores calculated by MetFID. Future work will focus on considering variations of ANNs to enhance the fingerprint prediction performance. For example, convolutional neural networks (CNNs) are able to exploit the locality of spectra and can be considered as an alternative to the fully connected ANN.

5 Conclusion

This paper introduces a new method, MetFID, for metabolite annotation using experimental MS/MS data. A multi-class, multi-label approach is used by MetFID to predict the entire molecular fingerprint at one time instead of the traditional digit-by-digit approach. We showed that the selection of training spectra is critical to enhance the model's prediction performance. In addition, we demonstrated that MetFID has the potential to achieve better accuracy in terms of metabolite annotation compared to currently used tools, such as MetFrag, ChemDistiller, and CSI:FingerID. Through experimental data from an in-house liver cancer study, we showed the ability of our method to rank putative metabolite IDs.

Author contributions ZF designed and implemented the algorithms, conducted evaluation experiments and wrote the manuscript; AA and GK performed data parsing, preprocessing, tool comparison and edited the draft; HWR directed the project and finalized the manuscript.

Funding The work presented in this paper is supported by NIH Grants U01CA185188 and R01GM123766 awarded to H.W.R.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethical approval In this paper, MS/MS data from a previous study (approved by Institutional Review Board) were used for evaluation.

References

Brouard, C., Shen, H., Duhrkop, K., d'Alche-Buc, F., Bocker, S., & Rousu, J. (2016). Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics, 32,* i28–i36.

Duhrkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., et al. (2019). SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods, 16,* 299–302.

Duhrkop, K., Shen, H., Meusel, M., Rousu, J., & Bocker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 12580–12585.

Fan, Z., Ghaffari, K., Alley, A., & Ressom, H. W. (2019). Metabolite identification using artificial neural network. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, November 18–21, 2019* (pp. 244–248).

Ferrarini, A., Di Poto, C., He, S., Tu, C., Varghese, R. S., Kara Balla, A., et al. (2019). Metabolomic analysis of liver tissues for characterization of hepatocellular Carcinoma. *Journal of Proteome Research, 18,* 3067–3076.

Gerlich, M., & Neumann, S. (2013). MetFusion: Integration of compound identification strategies. *Journal of Mass Spectrometry, 48,* 291–298.

Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., et al. (2018a). METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry, 90,* 3156–3164.

Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E., & Siuzdak, G. (2018b). Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nature Biotechnology, 36,* 316–320.

Gupta, N., Bandeira, N., Keich, U., & Pevzner, P. A. (2011). Target-decoy approach and false discovery rate: when things may go wrong. *Journal of the American Society for Mass Spectrometry, 22,* 1111–1120.

Heinonen, M., Shen, H., Zamboni, N., & Rousu, J. (2012). Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics, 28,* 2333–2341.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: A public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry, 45,* 703–714.

Johnson, C. H., & Gonzalez, F. J. (2012). Challenges and opportunities of metabolomics. *Journal of Cellular Physiology, 227,* 2975–2981.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2019). PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Research, 47,* D1102–D1109.

Landrum, G. (2016). RDKit: Open-Source Cheminformatics Software.

Laponogov, I., Sadawi, N., Galea, D., Mirnezami, R., & Veselkov, K. A. (2018). ChemDistiller: An engine for metabolite annotation in mass spectrometry. *Bioinformatics, 34,* 2096–2102.

Li, Y., Kuhn, M., Gavin, A. C., & Bork, P. (2020). Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. *Bioinformatics, 36,* 1213–1218.

Li, W., Yang, H., Buckley, B., Wang, L., & Kong, A. N. (2018). A Novel Triple Stage Ion Trap MS method validated for curcumin pharmacokinetics application: A comparison summary of the latest validated curcumin LC/MS methods. *Journal of Pharmaceutical and Biomedical Analysis, 156,* 116–124.

Matsuda, F. (2016). *Technical challenges in mass spectrometry-based metabolomics*. Tokyo: Mass Spectrometry.

Meringer, M., & Schymanski, E. L. (2013). Small molecule identification with MOLGEN and mass spectrometry. *Metabolites, 3,* 440–462.

Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. (2018). SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics, 34,* i323–i332.

Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. (2019). Recent advances and prospects of computational methods for metabolite identification: A review with emphasis on machine learning approaches. *Briefings in Bioinformatics, 20,* 2028–2043.

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal, 2,* 5.

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology, 13,* 263–269.

Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics, 8,* 3–016.

Scheubert, K., Hufsky, F., & Bocker, S. (2013a). Computational mass spectrometry for small molecules. *J. Cheminform, 5,* 12–2946.

Scheubert, K., Hufsky, F., & Bocker, S. (2013b). Computational mass spectrometry for small molecules. *Journal of Cheminformatics, 5,* 12–2946.

Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Duhrkop, K., et al. (2017). Critical Assessment of small molecule identification 2016: Automated methods. *Journal of Cheminformatics, 9,* 22.

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology, 34,* 828–837.

Wang, L., Li, S., & Tang, H. (2019). msCRUSH: Fast tandem mass spectral clustering using locality sensitive hashing. *Journal of Proteome Research, 18,* 147–158.

Watson, D. G. (2013). A rough guide to metabolite identification using high resolution liquid chromatography mass spectrometry in metabolomic profiling in metazoans. *Computational and Structural Biotechnology Journal*, *4*, e201301005.

Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vazquez-Fresno, R., et al. (2018). HMDB 40: the human metabolome database for 2018. *Nucleic Acids Research, 46,* D608–D617.

Zhang, F., Zhang, Y., Zhao, W., Deng, K., Wang, Z., Yang, C., et al. (2017). Metabolomics for biomarker discovery in the diagnosis, prognosis, survival and recurrence of colorectal cancer: a systematic review. *Oncotarget, 8,* 35460–35472.

Zhou, B., Cheema, A. K., & Ressom, H. W. (2010). SVM-based spectral matching for metabolite identification. In *Proceeedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, August 31–September 4, 2010* (pp. 756–759).