Machine Learning for Metabolic Identification

Dai Hai Nguyen[1], Canh Hao Nguyen[2], Hiroshi Mamitsuka[3]

[1]Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwa-no-ha, Kashiwa, Chiba 277-8561, Japan, Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
E-mail: hai@k.u-tokyo.ac.jp

[2]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan
E-mail: canhhao@kuicr.kyoto-u.ac.jp

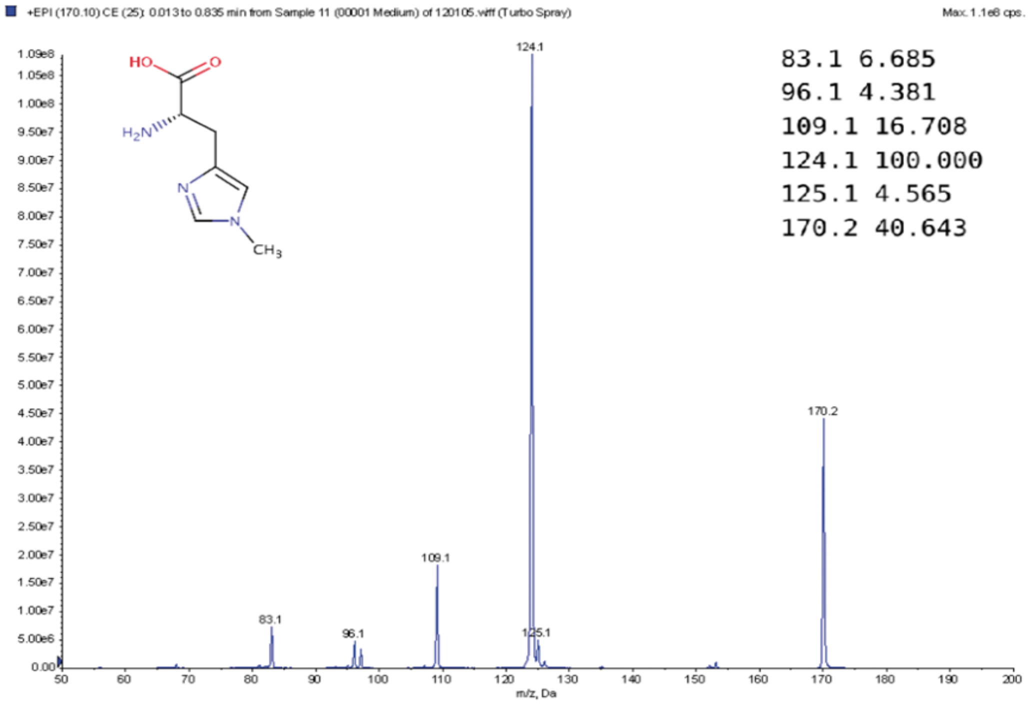
[3]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji 611-0011, Japan and Department of Computer Science, Aalto University, Espoo 02150, Finland
E-mail: mami@kuicr.kyoto-u.ac.jp

Abstract Metabolite identification is an essential part of metabolomics to understand biochemical characteristics of metabolites, which are small molecules and play important functions in biological systems. However, it remains challenging with many unknown metabolites in reality. Mass Spectrometry (MS) is a common technology in to deal with such small molecules. Over decades, many methods have been proposed for MS based metabolite identification task, especially, machine learning, being the key to recent progress in metabolite identification. This article provides a survey on computational methods for metabolite identification with the focus on *machine learning*, with a discussion on potential improvements for the task.

1. Introduction

Metabolomics involves studies of a plenty number of metabolites, which are small molecules present in biological systems. They play a lot of important functions such as energy transport, building blocks of cells and so on (Wishart, 2007). Identification of metabolites or understanding their biochemical characteristics is an essential and significant part of metabolomics to enlarge the knowledge of biological systems. It is also the key to the development of many application domains such as biomedicine, biotechnology or pharmaceutical. However, metabolite identification still remains a challenging task in metabolomics with a huge amount of potentially interesting but unknown metabolites in reality.

Mass spectrometry (MS) is one of common techniques in analytical chemistry (De Hoffmann *et al.*, 1997; Gross 2006; McLafferty *et al.*, 1993) for measuring the mass-to-charge ratio (*m/z*) of one or more molecules in a chemical sample. The output of a mass spectrometer, given a sample, is a mass spectrum, which is simply represented by a graph with *m/z* on the x-axis and the relative abundance of ions with *m/z* values on the y-axis. Another way to represent a mass spectrum is as a list of peaks, each of which is defined by its *m/z* and its relative abundance. An illustration of a mass spectrum is shown in Figure 1. A mass spectrometer consists of at least three components: ionization source, mass analyzer and a detector (De Hoffmann *et al.*, 1997). The ionization source is the component by which input molecules become charged ions. Two common forms of ionization are Electron Ionization (EI) and Electrospray Ionization (ESI). The mass analyzer is the component to physically separate ions by their *m/z*. The common types are quadrupole, time-of-flight and orbitrap devices. Once the ions have been separated, the detector is responsible for detecting and quantifying these ions. The mass spectrum contains peaks corresponding to the masses and relative abundance of the charged fragments and the precursor ions as well. Since these values provide masses of some of substructures, they can be used to elucidate the structure of the measured molecule. In practice, MS/MS, also known as tandem MS, is often preferred, and has been versatile and powerful for many applications. It consists of two mass analyzers coupled with Collision Induced Dissociation (CID). Ions are separated in the first mass analyzer (MS1), then enter a collision or fragmentation cell and fragmented, leading to generation of ions, called product ions, which are separated in the second mass analyzer (MS2) and detected, eventually resulting in MS/MS or tandem mass spectra. Multi-stage MS allows to further fragment the product ions, providing ways to link these product ions to their precursor ions, thus, offering more information about the fragmentation process (Nguyen *et al*., 2019).



83.1	6.685
96.1	4.381
109.1	16.708
124.1	100.000
125.1	4.565
170.2	40.643

Figure 1. Example MS from Human Metabolome Database (Wishart, 2007) for 1-Methylhistidine (HMBD00001), with its corresponding chemical structure (top-left) and list of peaks (top-right)

Metabolite identification from (tandem) mass spectra is an important step for further chem-biological interpretation of metabolomics samples. In practice, this process is presumed to be challenging and also time-consuming task in metabolomics experiments. Different from peptides and proteins where the fragmentation is generally simple due to the repetition of their structures, that of metabolites under varying fragmentation energies is a more complicated stochastic process. Thus, the interpretation of mass spectra is cumbersome and require expert knowledge. MS based metabolite identification can be regarded as a retrieval task, that is, given a query spectrum of an unknown molecule, the goal is to find molecules (usually from a given reference database), which have similar spectra. It is straight-forward to directly compare the query against reference spectra in the reference spectra database (also known as spectra library). The candidate molecules from the spectra library are ranked based on the similarities between their reference spectra and query, and the best matched molecules are returned. However, the spectra libraries often contain spectra of a small fraction of molecules in reality, leading to unreliable results, if the molecule corresponding to the query spectrum is not in the spectra libraries. Therefore, in order to mitigate the insufficiency issue of such database, alternative computational approaches are devised (Nguyen *et al*., 2019).

A number of computational methods and software tools have been developed to deal with the task of metabolite identification. We systematically organize them into four groups based on their methodologies in (Nguyen *et al*., 2019): (1) mass spectra library; (2) *in silico* fragmentation; (3) fragmentation trees and (4) machine learning, see Figure 2 for the details. We briefly describe the approaches as follows: (1) with a given query spectrum of an unknown molecule, mass spectra library is to compare the query against spectra in the library. (2) *In silico* fragmentation based methods attempt to generate simulated spectra from chemical structures of reference compounds in the structural databases, which leverage a huge number of chemicals structures of known molecules, and then compare the query against the simulated ones. (3) Fragmentation trees based approaches take advantages of relations of peaks in spectra, represented by fragmentation trees, which can be directly predicted from spectra by some combinatorial optimizations and used to cluster molecules into categories. (4) Machine learning (ML) approaches are to learn and predict intermediate representations between spectra and molecular structures and then use such representations for matching and retrieval. In this article, we focus on the above (2) and (4), which are *in silico* fragmentation and ML, respectively, for metabolite identification.

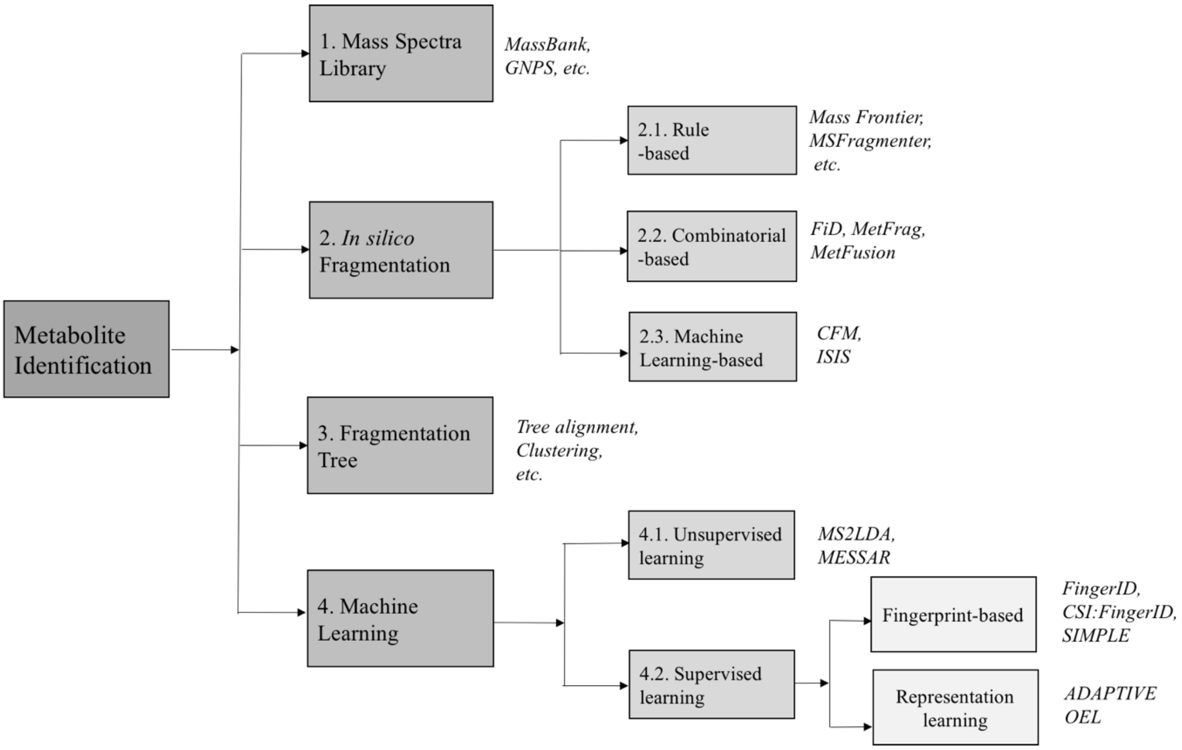

Metabolite Identification
1. Mass Spectra Library
MassBank, GNPS, etc.
2. In silico Fragmentation
2.1. Rule -based
Mass Frontier, MSFragmenter, etc.
2.2. Combinatorial -based
FiD, MetFrag, MetFusion
2.3. Machine Learning-based
CFM, ISIS
3. Fragmentation Tree
Tree alignment, Clustering, etc.
4. Machine Learning
4.1. Unsupervised learning
MS2LDA, MESSAR
4.2. Supervised learning
Fingerprint-based
FingerID, CSI:FingerID, SIMPLE
Representation learning
ADAPTIVE OEL

Figure 2. The overview of approaches for metabolite identification. The figure is adapted from (Nguyen *et al*., 2019).

2. *In silico* fragmentation tools to aid metabolite identification

Due to the lack of MS/MS data of compounds in mass spectra libraries, the capability of identifying unknown molecules through search in mass spectra libraries is limited as mentioned in the previous subsection. Therefore, the advent of software tools for predicting fragments and their abundance from the molecular structures of compounds can fill the gap between spectra and structural databases. This strategy has been successfully applied in protein studies to construct database containing data on trypsin-associated cleavage and MS/MS of peptides, such as MASCOT (Perkins *et al*., 1999) and SEQUEST (Eng *et al*., 1994). However, in comparison with the prediction of fragmentation mechanisms for peptides and protein, which are simple due to the repetition in their structures, the fragmentation of product ions of metabolites in the tandem mass spectrometer is a much more complicated stochastic process, and depends on a variety of factors, including: the detailed three-dimensional structure of metabolites, the amount of energy to break several certain bonds to obtain the product ion, the probabilities of different dissociation reactions, which can be considered as a function of the applied collision energy and the pressure in the collision chamber and so on. Nowadays, many *in silico* fragmentation software tools have been developed and are widely used to identify MS/MS when the sizes of spectra libraries are limited. In this section, we survey different computational tools and methods using various algorithms for *in silico* fragmentation. The algorithms differ in the way they deploy different strategies to generate *in silico* fragments from the molecular *structure/graph* of the candidate compounds. We divide them into three subgroups (Nguyen *et al*., 2019): rule-, combinatorial- and ML- based fragmentation tools (see Figure 2).

2.1. Rule based methods

The rule-based *in silico* fragmentation tools are used to predict/generate theoretical spectra from molecular structures/graphs of compounds in the database using a set of chemical rules. This set of rules is a collection of general and heuristic rules of fragmentation processes extracted from data sets of elucidated MS/MS in literature. The predicted spectra of candidate compounds from the database will be compared with the query (Hill *et al*., 2008; Kumari *et al*., 2011).

A typical commercial software tool, Mass Frontier (Mistrik, 2004), developed by HighChem, can generate fragments according to known general rules, or to specific rule libraries. The libraries can be defined by users or provided by HighChem or combination of both. ACD/MS Fragmenter (referred at: http://www.acdlabs.com), another commercial tool, also uses a comparable set of chemical rules to generate fragments. MOLGEN-MSF (Schymanski *et al*., 2009), developed by the University of Bayreuth, uses general fragmentation rules and also is able to accept additional rules as an optional input file when calculating fragments. Additionally, non-commercial rule based software tools, like MASSIS (Chen *et al*., 2003) and MASSIMO (Gasteiger *et al*., 1992), use different ways. Particularly, structure-specific cleavage rules contained in MASSIS are divided into 26 different molecular classes. An input molecule is classified into one or some of these classes and the corresponding fragmentation rules are applied to obtain a set of fragments. MSSIMO uses a small set of general fragmentation reactions parameterized with reaction probabilities drawn from a collection of determined fragmentations.

In practice, these rule-based methods are not widely used due to several drawbacks: 1) the fragmentation process can significantly vary under small changes in structure of a molecule. Thus, a fragmentation rule collected from a known fragmentation of a molecule may not be properly applied to another, despite that they have highly similar molecular structures; 2) it has been experimentally demonstrated that a single set of general rules is insufficient to identify some observed fragments with a reasonably high accuracy. Even though specific rules are constantly added to rule databases, it is not necessary to apply them to a new undiscovered compound in many cases, and 3) the product ions of generated spectra have the same intensities because the bond cleavage rates are ignored. In reality, different molecules can generate the same product ions and the relative intensities can play an important role in discriminating these molecules.

2.2. Combinatorial based methods

Different from the above software tools, which are mainly based on fragmentation rule databases, combinatorial-based methods are to generate a graph of substructures from the chemical structure of a candidate compound in the database (see Figure 3), then find the most likely subset of the substructures or so-called fragmentation trees that best matches the query by solving optimization problems. An advantage offered by this approach is in situations where MS/MS of compounds with less known fragmentation rules are queried. Some typical methods are reviewed in this subsection. In general, methods belonging to this subsection differ in the way of how they find the fragmentation tree best matches to the query spectra to produce similarity scores.

FiD (Fragment iDentificator, Heinonen *et al.*, 2008) performs a search over all possible fragmentation paths and outputs a ranked list of alternative structures. In particular, given a graph structure of a precursor ion and its MS/MS, FiD first generates all possible connected subgraphs by a depth-first graph traversal, then computing the masses of product ions corresponding to the generated subgraphs to match with observed peak masses in the spectrum. After that, a list of candidate fragments is obtained, then each of which is assigned a cost, namely, the standard bond energy required to cleave bonds from the precursor ion. Obviously, the candidate fragments with smaller costs are preferred. Finally, a combinatorial optimization method, such as mix integer linear programming (MILP) is used to assign candidate fragments to measured peaks with minimal cost. Their experimental results showed that the product ions predicted by FiD are more consistent with the manual identification produced by domain experts than those of the rule-based fragment identification tools mentioned in the previous section. However, the main drawback of FiD is the computational expensiveness because of the following reasons: 1) rapid increase in the number of connected subgraphs; 2) the computational complexity of MILP to explain peaks with most likely candidate fragments. Thus, FiD can be applied to only small sized molecules.

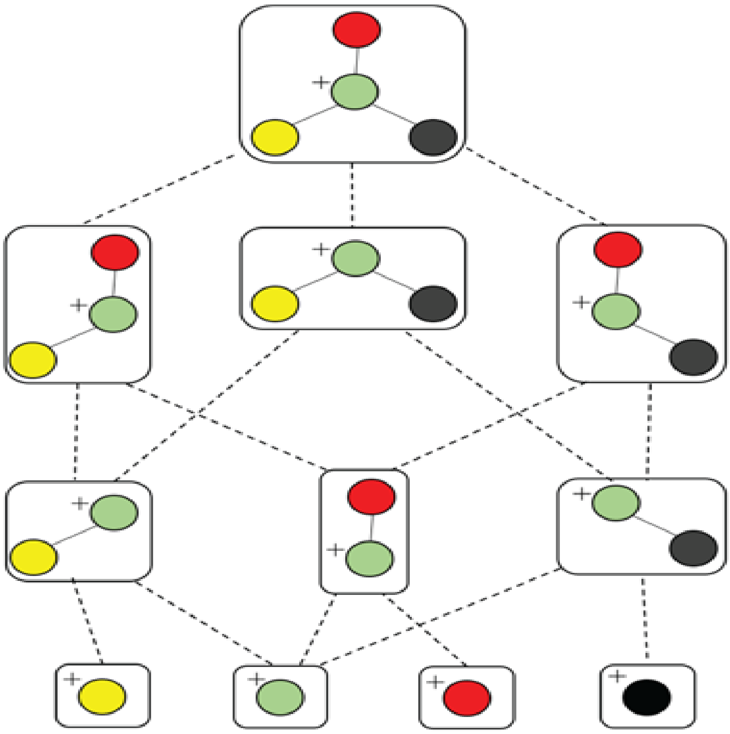


Figure 3. An illustration of generating all connected subgraphs of the precursor graph. The figure is adapted from (Nguyen *et al*., 2019).

Another combinatorial based method is MetFrag (Wolf *et al.,* 2010), which uses heuristic strategies, such as a breadth-first search algorithm with a maximum tree depth parameter or removing duplicated subgraphs, to narrow down the search space of candidate fragments, overcoming the computational issue of FiD, which employs depth-first graph traversal to generate subgraphs. Hence, it is much faster than FiD and can be applied to a full structure database to search for the compound that explains best the spectrum. MetFrag uses bond dissociation energies for the cost of cleaving bonds. The candidate fragments are then used to rank the candidate molecules in the database without finding the most likely fragments corresponding to the spectrum. In a similar fashion, Ridder *et al*., (2012) introduced MAGMA, an extended version to multistage spectral trees MS^n. Different from MetFrag, when a substructure is considered to explain an MS^2 product ion, which is the precursor ion of MS^3, in addition to its substructure score, the resulting MS^3 is also taken into account. This spectrum is temporarily annotated with only sub- set of the substructure, similarly to MS^2 level fragmentation spectrum. Then, the substructure scores obtained at the third level are added to the sore at the second level, and this total core is for ranking substructure candidates for MS/MS peak and its fragmentation spectrum. This procedure is applied recursively to handle MS^n with any level.

Gerlich *et al*., (2013) presented a system, namely MetFusion, to combine the results from MassBank (search in spectra database) and MetFrag (*in silico* fragmentation). The combination aims at taking advantage of complementary approaches to improve the compound identification, that is, the vast coverage of the structural databases queried by MetFrag and reliable matching results achieved by search in spectra libraries if similar spectra are available. The experimental results showed that a combination of an *in silico* fragmentation based method with curated reference measurements can improve compound identification and achieve the best of two approaches. More details about this method and results can be found in (Gerlich *et al*., 2013).

A drawback of this approach is that the above methods are mainly based on a bond disconnection based approach to generate fragments from molecules, e.g. standard bond energy and bond dissociation energy used by FiD and MetFrag, respectively. However, these are solely approximate estimates and bond dissociation energies are much more complicated in reality. These limitations have been tackled with some methods based on learning models, which are presented the following subsections.

2.3. Machine learning based methods

Besides the above approaches to generate *in silico* fragmentation from graph structure of compounds, there are a few work proposed to use machine learning models to learn the fragmentation process from the training data and have shown great promise in generating *in silico* spectra for the structural identification purpose. To avoid the confusion with the content in section 3, we clarify here that machine learning methods are used to learn and predict the presence of certain fragments (e.g. whether a bond between two atoms is broken or not) to generate *in silico* spectra from molecular structures. In a different sense, methods in section 3 are to learn and perform classification or clustering from spectra (see Figure 4 for illustration).

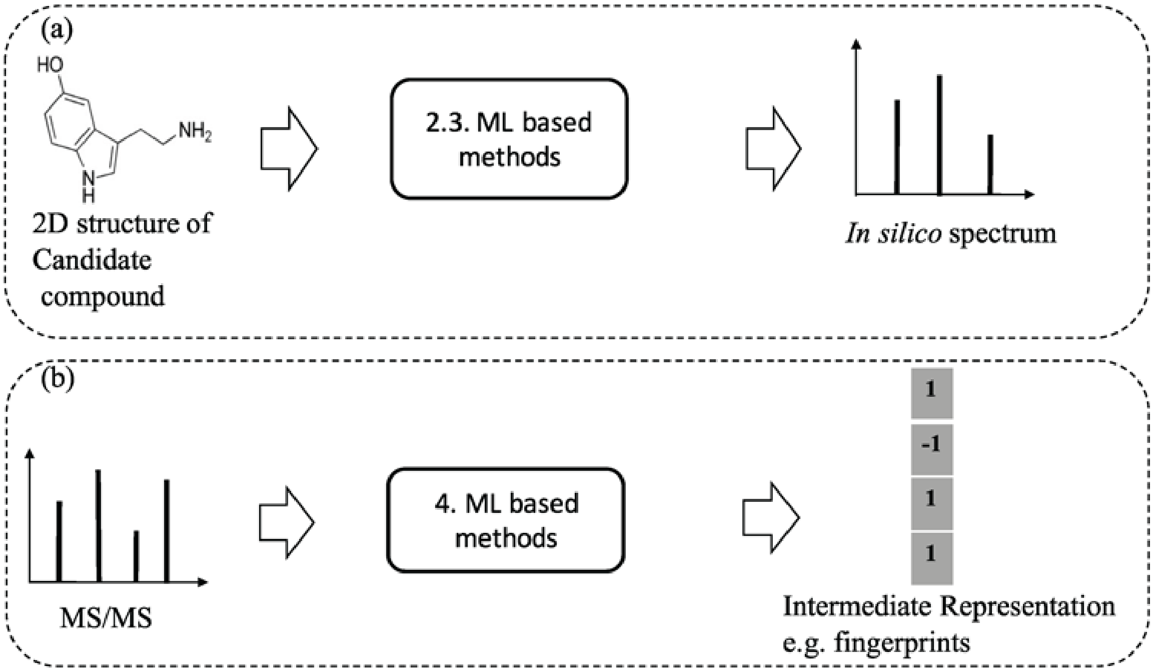


Figure 4. An illustration to clarify the difference between ML-based methods for learning and predicting *in silico* spectra from 2D structures of compounds (a) and ML based methods for learning and predicting substructures or chemical properties from MS/MS (b). The figure is adapted from (Nguyen *et al.*, 2019).

The previously mentioned methods to generate *in silico* fragments from molecular structure of compounds are based on either chemical reaction equations or approximate bond strength. None of them have shown sufficient accuracy in generating *in silico* spectra for enable automated and correct identification of metabolites. To overcome the difficulty, Kangas *et al.*, (2012) presented a method, named ISIS, using machine learning to generate *in silico* MS/MS spectra for lipids solely from chemical structure of compounds without fragmentation rules and no need to define bond dissociation energy. The main idea is that, for every bond in the molecular structure, one artificial neural network (ANN) is designed to predict bond cleavage energy from which bond cleavage rates can be calculated to determine the relative intensities;

another is to predict which side of the bond is charged and captured by the detector in the mass spectrometer. These ANNs are iterated over all bonds in a molecule to find bond cleavage energies and charged ions. For the leaning process, the weights of the former ANN are trained by genetic algorithm (GA) to better predict the bond cleavage energies that produce ions and their corresponding intensities in the *in silico* spectra. The objective of GA is to have the *in silico* spectra match those in the experimental spectra using a Pearson R^2 correlation. The latter ANN is trained by backpropagation algorithm in which the labels can be found by comparing the fragment masses to the experimental spectra.

Allen *et al*., (2015) proposed a probabilistic generative model, namely Competitive fragmentation mode, for the fragmentation process. They assume that each peak in the spectrum is generated by a fixed length sequence of random fragment states. It consists of two models: transition model to define the probability of each fragment leads to another at one step in the process and an observation model to map the final intermediate fragment state to the give peak. The parameter estimation for the transition and observation models is performed by an Expectation Maximization-like algorithm. The trained CFM can be used to predict peaks in the spectrum and for metabolite identification. The results showed that, CFM obtained significantly better ranking for the correct candidate than bot of MetFrag and FingerID (Heinonen *et al*., 2012). However, like above methods, this method is limited to small molecules due to the combinatorial enumeration of fragmentation possibilities. It is also worthy noting that, while ISIS is based on supervised machine learning, CFM is based on unsupervised learning to predict spectra.

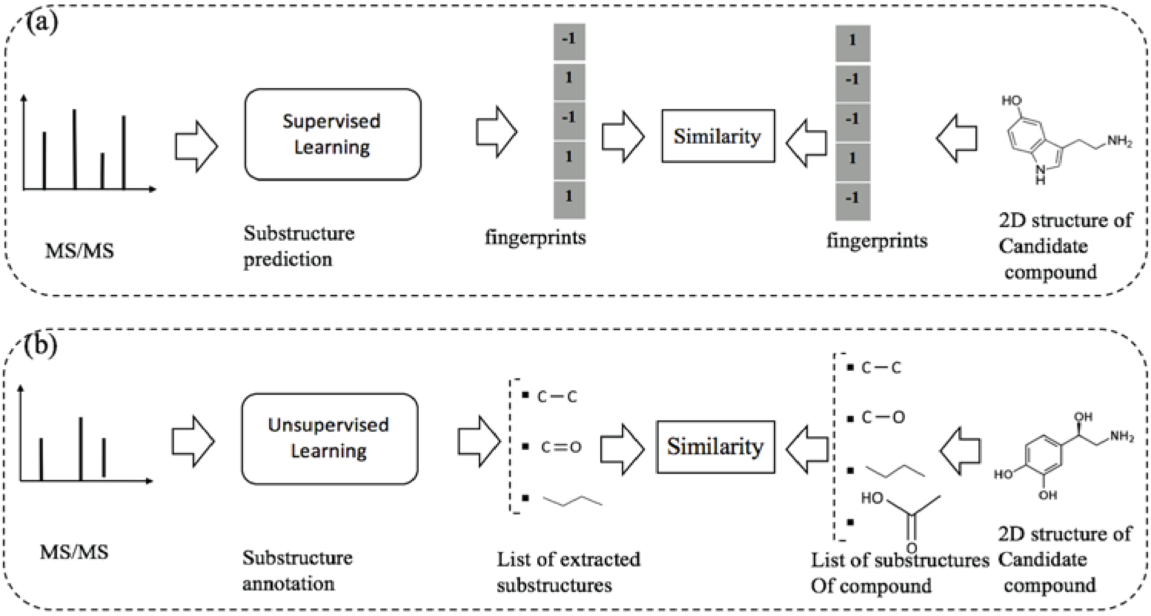


Figure 5. An illustration to clarify the difference between supervised and unsupervised learning for metabolite identification: (a) substructure prediction using supervised learning to map a given MS/MS to an intermediate representation (e.g. fingerprints), which is subsequently used to retrieve candidate metabolites in the database. (b) substructure annotation using unsupervised learning to extract biochemically relevant substructures with certain confidence from the given spectrum. Then, the similarity between the MS/MS and a chemical structure of a metabolite is estimated according to their common substructures. Note that the output of supervised learning (e.g. fingerprints) may indicate the presence/absence of all ‘predefined’ substructures whereas that of unsupervised learning may be a list of substructures frequently occurring in the database. The figure is adapted from (Nguyen *et al*., 2019).

3. Machine learning for metabolite identification

A number of computational methods or tools have been introduced to deal with the task of metabolic identification. Remarkably, ML is the key to recent development of the task. Besides identifying molecular compounds by searching in spectra and structural database as presented in the previous sections, there are a number methods proposed to predict predefined substructures or more generally chemical properties such as (Heinonen *et al.,* 2012; Dührkop *et al*., 2015; Brouard *et al.,* 2016; Nguyen *et al*., 2018; Nguyen *et al*., 2019), to name a few. Another direction is to automatically discover substructures directly from a set of MS/MS, from which we can identify the candidate compounds from the database based on their substructures, such as (Mrzic *et al*., 2017; Van Der Hooft *et al*., 2016). This section is devoted to covering ML frameworks for these purposes, which can be divided into two groups: supervised learning for predicting substructures and unsupervised learning for annotating substructures. Furthermore, the difference between the two subgroups can be intuitively illustrated as in Figure 5.

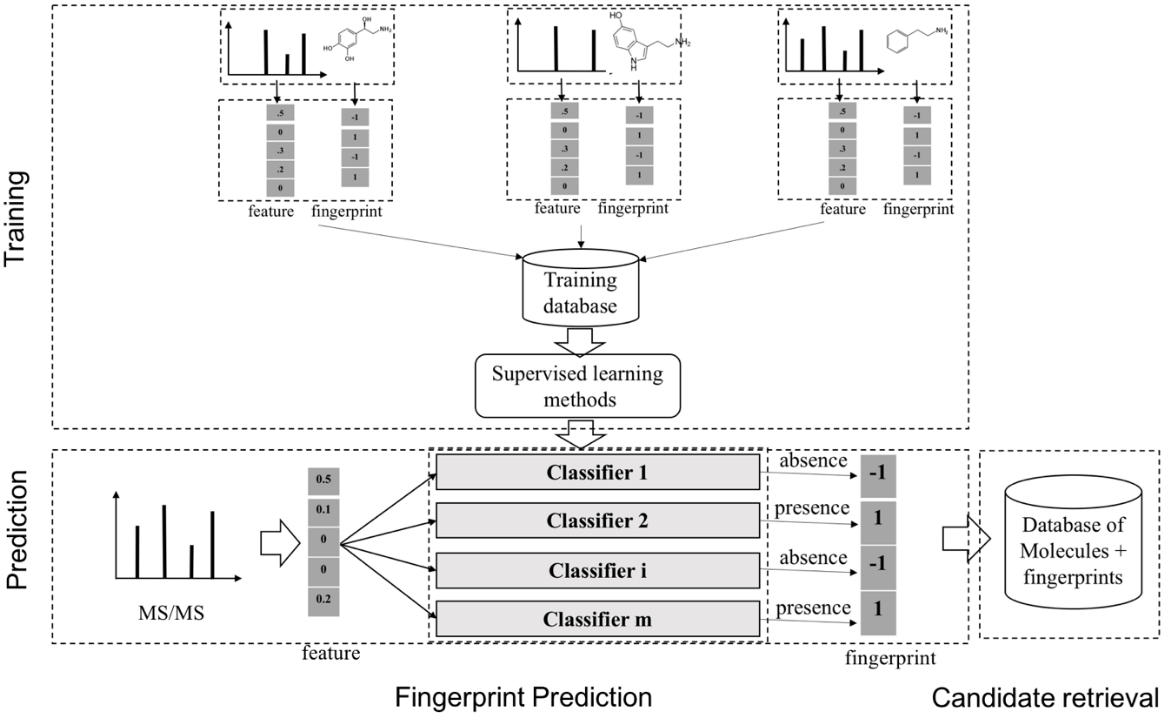


Figure 6: A general scheme to identify unknown metabolites based on molecular fingerprint vectors. There are two main stages: 1) fingerprint prediction: learning a mapping from a molecule to the corresponding binary molecular fingerprint vector by classification methods, given a set of MS/MS spectra and fingerprints; 2) candidate retrieval: using the predicted fingerprints to retrieve candidate molecules from the databases of known metabolites. The figure is adapted from (Nguyen *et al*., 2019).

3.1. Supervised learning for predicting substructures.

The task of supervised learning for metabolite identification is that, given a set of MS/MS and corresponding molecular structures of known molecules, supervised learning methods are to learn a mapping function from a MS/MS to a molecular structure of molecule. However, this task is technically challenging because both input and output spaces (spectra and molecular structures, respectively) are highly structured objects. Instead of directly learning a mapping from spectra to molecules, the approach of learning through intermediate representations between spectra and molecules has been used in many systems. The intermediate representations can be either manually designed fingerprints of molecules or representations directly learned from the data. We go into details of such kinds of intermediate representations in the following subsections.

3.1.1. Fingerprint based methods

A molecular fingerprint is a feature vector, which is used to encode the structure of a molecule. In general, the values of this vector are often binary, indicating the presence or absence of a certain substructure or chemical property. The methods based on fingerprints are often carried out in two steps as illustrated in Figure 6. The first step is to predict a fingerprint with supervised ML, which is regarded as a collection of binary classification task, each corresponds to a bit in the fingerprint. The second step then uses the predicted fingerprint to query the database with techniques in ranking/information retrieval.

Kernel methods have been shown effective in dealing with the first step of predicting fingerprints from MS/MS. A notable method is FingerID (Heinonen *et al.*, 2012), which used support vector machine (SVM, Burges *et al.*, 1998) with kernels defined on MS/MS to predict fingerprints. The kernels for pairs of MS/MS were defined, including integral mass kernel and probability product kernel (PPK, Jebara *et al.*, 2004). It is worth noting that the above kernels are mainly based on the information from individual peaks present in the spectra while ignoring their interactions. In fact, such information is proven to be useful in predicting fingerprints.

CSI:FingerID (Dührkop *et al.*, 2015), an extended version of FingerID, jointly takes MS/MS and corresponding fragmentation trees (FTs, Böcker *et al.*, 2008) as input to improve the predictive performance. FTs play an important role in interpreting the structure of molecule since it is usually assumed that only MS/MS is insufficient to describe the fragmentation process. It is noteworthy that FTs are constructed from spectra and can be used to provide prior knowledge about the structure of compounds e.g. dependencies between peaks in the spectra, which was ignored in the FingerID. In order to incorporate MS/MS and FTs, kernels for FTs have to be defined, which range from simple ones for vertices, including node binary (NB), node intensity (NI); for edges, including loss binary (LB), loss count (LC), loss intensity (LI) to complicated ones like common path counting (CPC), common subtree counting (CSC) and so on (see more details in Dührkop *et al.*, 2015). Subsequently, multiple kernel learning (MKL, Gönen *et al.,* 2011) is used to combine these kernels with ones defined for MS/MS using one of the following methods: centered alignment (ALIGNF), quadratic combination and l_p-norm regularized combination. The combined kernel is then used in learning the final model for fingerprint prediction. CSI:FingerID presented improved scores against other benchmark tools but has the current limitation of processing MS/MS one at a time because of the need of computationally heavy conversion of spectra into FTs. For a similar purpose of incorporating peak dependencies into the learning for fingerprint prediction, Nguyen *et al.*, (2018) designed a kernel for peak interactions and combine this kernel with other kernels defined for individual peaks through MKL. The combined kernels were then combined with SVM to predict fingerprints. The experimental results showed that the micro-average accuracy and F1 score of the combined kernels were higher than PPK, while being comparable with the state-of-the-art kernels for CSI:FingerID. More importantly, this kernel achieved faster computation than CSI:FingerID because it does not use FTs in the training and testing phases.

Input output kernel regression (IOKR, Brouard *et al.,* 2016) is another kernel based method and has been shown to outperform the previous methods, in terms of both predictive performance and computational speed. It learns a function mapping from MS/MS to molecular fingerprints (or molecular structures directly). For this purpose, IOKR defines two kernels to encode similarities in the input space (e.g. MS/MS) and output space (fingerprint or molecular structures). The information about input and output is implicitly encoded in these two kernels. Then, the advantage of IOKR over previous kernel methods stem from the two following points: (i) unlike previous kernel based methods, IOKR can handle the structured output space (e.g. feature interactions in molecular fingerprints) by the kernel defined for the output, which improves the predictive performance; (ii) IOKR can simultaneously predict fingerprints rather than considering fingerprint prediction as a set of separate tasks, leading to an efficient computation in the prediction stage. In fact, one can take structures of molecules into account by using graph kernels such as path, shortest-path or graphlet kernels. It is shown that kernels defined on molecular fingerprints obtained the best performance (Brouard *et al*., 2016).

Kernel-based methods are difficult to deal with sparse data and lack of interpretation. That is, each bit in a fingerprint represents a predetermined chemical property or substructure and its presence is often decided by a few number of peaks in MS/MS. Also the number of training data is small, while sparse learning models have not been considered yet. In addition, sparse learning models are advantageous in that their results are easily interpretable. Nguyen *et al*., (2018) proposed a sparse, interpretable model, called SIMPLE, to incorporate peak interactions explicitly and have a high interpretability. The model is as follows: given a MS/MS, represented by a feature $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_1}, \dots, \boldsymbol{x_d}]$, for one particular bit in the fingerprint, they formulate the model for individual peaks and interactions as follows:

$$\begin{aligned} f(x; w, W) &= b + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} x_i x_j \\ &= b + w^T + x^T W x \end{aligned}$$

where $b \in \mathbb{R}$, $w \in \mathbb{R}^d$ and $W \in \mathbb{R}^{d \times d}$. The prediction function consists of a bias b and two terms: main effect term parameterized by the weight vector w and interaction term parameterized by the weight matrix W. Their roles are different, as illustrated in Figure 7. While the former captures the information about the individual peaks, the latter captures the information about peak interactions. For the purpose of interpreting the weights learned by the model, they impose $l1$- norm (Tibshirani *et al*., 1996) and nuclear norm (Srebo *et al*., 2005) regularizations on main effect and interaction terms to induce sparsity in w and lowrankness in W after training. The training stage is performed by minimizing a convex objective function, guaranteeing that the obtained solution is globally optimal. The incorporation of peak interactions in the prediction model $f(x; w, W)$ were found to significantly improve the prediction accuracy of not using interaction, resulting in comparable performance with the current top methods. Furthermore, SIMPLE could show the interpretability of results (see Nguyen *et al*., 2018 for more details).

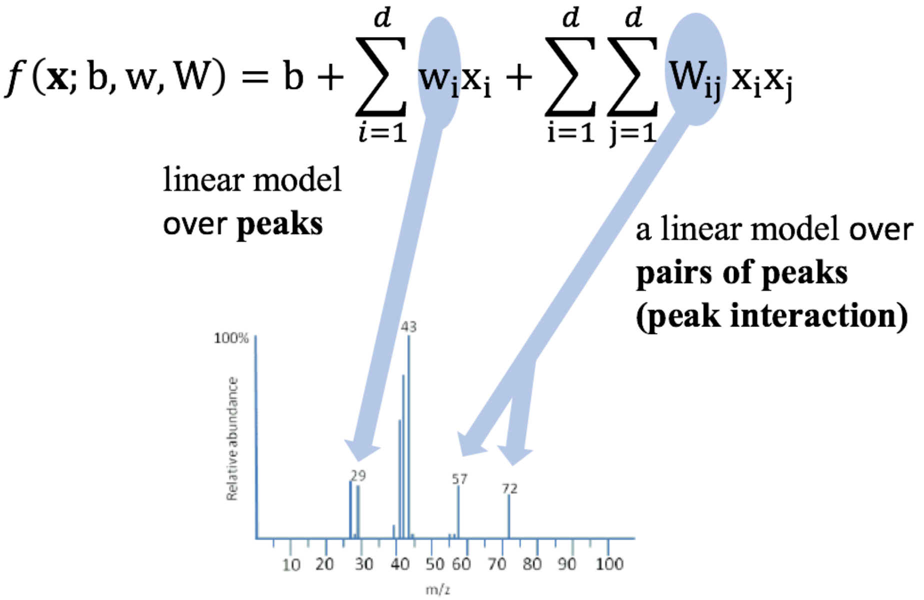
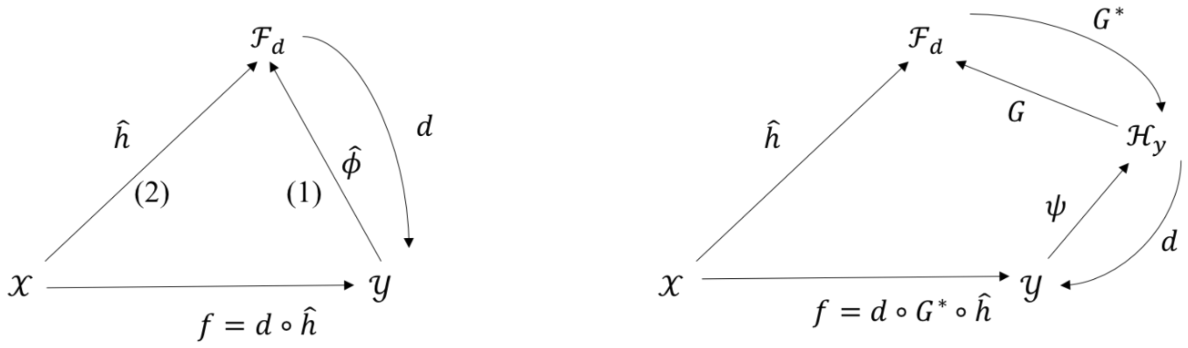

$f(\mathbf{x}; \mathrm{b}, \mathrm{w}, \mathrm{W}) = \mathrm{b} + \sum_{i=1}^{d} \mathrm{w_i x_i} + \sum_{i=1}^{d} \sum_{j=1}^{d} \mathrm{W_{ij}\, x_i x_j}$
linear model over **peaks**
a linear model over **pairs of peaks (peak interaction)**
100%
Relative abundance
43
29
57
72
10 20 30 40 50 60 70 80 90 100
m/z

Figure 7: Illustration of the predictive model of SIMPLE: the weight vector w of the main effect term captures information about the individual peaks, while interaction weight matrix W of the interaction term captures information about the peak interactions.

3.1.2. Learning intermediate representations for molecules from spectra

Using molecular fingerprints as representations for molecules have been shown effective in metabolite identification task. However, they have a couple of drawbacks: (i) molecular fingerprints should be large in size to encode all possible substructures and chemical properties related to molecules, causing slow prediction in the candidate retrieval step; (ii) molecular fingerprints are not necessarily specific to any task nor data, and therefore redundant in the sense that they might contain much information irrelevant to the task and data, resulting in limited predictive performance.



(a) ADAPTIVE

(b) OEL

Figure 8: Schematic illustration of ADAPTIVE and OEL for learning representations for molecules from the given set of spectra and molecular structures. (a) ADAPTIVE has two functions to learn: (1) function $\hat{\phi}$ (parameterized by a graph neural network) mapping from molecular structures to representations; (2) function $\hat{h}$ regressing representations from given input spectra. (b) OEL jointly learns two functions: (1) G (and G^*) mapping from molecular

structures in the Hilbert space to representations; (2) function $\hat{h}$ regressing representations given input spectra.

Motivated by these drawbacks of molecular fingerprints, Nguyen *et al.*, (2019) proposed ADAPTIVE, which allows to generate representations for molecules using their molecular structures. The representations are learned from pairs of spectra and corresponding molecular structures, and thus specific to both data and task of metabolite identification. In a technical detail, ADAPTIVE has two subtasks in the learning step: (1) learning a mapping $\hat{\phi}$ from molecular structure space $\mathcal{Y}$ to representation space $\mathcal{F}_d$ and (2) learning another mapping $\hat{h}$ from spectra space $\mathcal{X}$ to the representations obtained from step (1), as illustrated in Figure 8a. In subtask (1), the mapping function $\hat{\phi}$ is parameterized by a model, named a message passing neural network (MPNN, Gilmer *et al.,* 2017) for mapping a molecular structure of molecule to a representation vector. MPNN is a framework, which can take graphs of arbitrary sizes and structures as inputs, to generate their representations at different levels (i.e. nodes, subgraphs and the whole graph) with regard to a number of learning tasks including supervised, unsupervised and semi-supervised learning (see Gilmer *et al.*, 2017; Nguyen *et al.*, 2017; Duvenaud *et al.*, 2015 for more details). A key advantage is that MPNN allows to learn representations specific to the given task from the given data. The parameters of the MPNN are trained so that the correlation between the spectra and vectors mapped from molecular structures is maximized. The correlation is estimated by Hilbert-Schmidt Independence Criterion (HSIC, Gretton *et al.*, 2005). For subtask (2), IOKR is used to learn the mapping $\hat{h}$ from spectra to representations. The empirical validation of ADAPTIVE with a benchmark data set showed the advantages of ADAPTIVE over existing methods, including the state-of-the-art IOKR, both in predictive performance and computation time for prediction step.

In a somewhat similar fashion, Output Embedding Learning (OEL, Brogat-Motte *et al.*, 2020) is a kernel based framework, which jointly learns a finite embedding (representation) of the outputs (chemical structures of molecules) and the regression of the representations given inputs (spectra), using the prior information about the structure of outputs and unlabeled output data. Formally, given a collection of pairs of spectra and structures $\{(x_i, y_i), i = 1, \dots, n\}$ and a collection of unlabeled molecular structures $\{y_j, j = 1, \dots, m\}$, the optimization problem is as follows:

$$\min_{h,G} \frac{\gamma}{n} \sum_{i=1}^{n} \|h(x_i) - G\psi(y_i)\|_{\mathcal{F}_d}^2 + \frac{1-\gamma}{n} \sum_{j=1}^{m} \|\psi(y_i) - G^* G\psi(y_i)\|_{\mathcal{F}_d}^2$$

where $\psi\colon \mathcal{Y} \rightarrow \mathcal{H}_y$ is a function that maps structured objects (molecular structures) into a Hilbert space $\mathcal{H}_y$; G is an operator from the Hilbert space $\mathcal{H}_y$ to the space of representations $\mathcal{F}_d$ and orthogonal ($GG^* = Id_d$); γ is a hyperparameter controlling the tradeoff between two terms in the objective function. It is noted that the first term corresponds to the regression of the embedding of chemical structures given spectra. The second term corresponds to learning a one-layer linear auto-encoder (parameterized by G^*G) whose inputs and outputs belong to the Hilbert space $\mathcal{H}_y$. However, the hidden layer is trained in the supervised manner by having a constraint with the regression part. The learning scheme of OEL is illustrated in Figure 8b.

While both of the methods above have the same purpose of learning the representations for molecular structures of molecules through the available data, the key difference lies on the ways they learn representations. While the former relies on the graph neural network to parameterize the model to generate representations for chemical structures, the latter adopts kernels to encode the information about structures in an implicit way. Both methods have been shown to outperform the fingerprint-based methods for the metabolite identification tasks in terms of both the predictive performance and speed. The improvements suggest that learning representations for molecular structures with the available information of spectra is beneficial for the metabolite identification task in overcoming the limitations of manually-designed feature vectors for molecules such as molecular fingerprints. However, a drawback of ADAPTIVE and OEL would be interpretability, because structural information of molecular structures is implicitly encoded by complicated nonlinear neural networks and kernels in ADAPTIVE and OEL, respectively, and cannot be made explicit easily. In metabolite identification, it would be desirable to connect the set of peaks to the corresponding substructures or chemical properties of molecules (see SIMPLE, Nguyen *et al.*, 2018). Developing a model with the advantages of interpretability and high predictive performance would be interesting future work.

3. 2. Unsupervised Learning for Substructure Annotation

Unlike supervised learning methods which aim to learn the mapping function from input to output, the purpose of unsupervised learning methods is to learn the underlying structures from a set of inputs without outputs specified. In the context of MS/MS of metabolites, the inputs are only a collection of MS/MS. Metabolites may have common substructures, yielding similar product ions in their MS/MS. Many substructures among them contain information pertaining to the biochemical processes present. Therefore, extraction of such biochemically relevant substructures allows metabolites to be grouped based on their shared substructures regardless of classical spectral similarity. Also, this can be used to improve the accuracy of metabolite identification (Nguyen *et al.*, 2019).

One of the typical software tools for chemical substructure exploration is MS2Analyzer (Ma *et al.*, 2014), which is a library-independent tool. It allows to exploit the potential structure information contained in MS/MS. It was developed to elucidate substructures of small molecules from accurate MS/MS. The main function of this tool is to search mass spectral features including neutral loss, precursor, fragment ions mass and mass differences in a large number of spectra. Through the combination of searching results and substructures/compound class relationship knowledge, compounds can be identified. However, MS2Analyzer can find all molecules sharing a specific set of mass spectral features provided by users, and sample-specific features are likely to be ignored. Another technique, namely molecular networking (Wang *et al.*, 2016; Watrous *et al.*, 2012; Yang *et al.*, 2013), groups parent ions i.e. MS1 peaks, based on their MS2 spectral similarity, e.g. cosine score, such that metabolites which are structurally annotated in a cluster can be used to annotate their neighbors. However, a drawback of molecular networks is that only MS1 peaks with high similarity are grouped and spectral features specifying the clusters have to be manually extracted. Thus, it may fail to cluster molecules sharing small substructures with low MS2 spectral similarity.

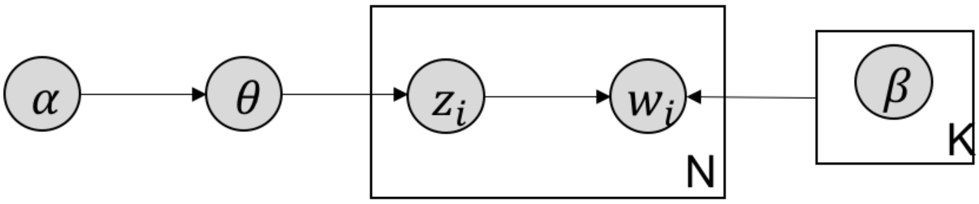


Figure 9: Simplified graphical representation of LDA. The figure is adapted from (Nguyen *et al.*, 2019).

MS2LDA, presented in (Van Der Hooft *et al.*, 2016), is a software tool offering benefits of both above methods, while overcoming their disadvantages. It can automatically extract relevant substructures in molecules based on their co-occurrence of mass fragments and neutral losses, and cluster the molecules accordingly. Based on the assumption that each observed MS/MS is composed of one or more substructures, MS2LDA adopts Latent Dirichlet Allocation (LDA, Blei *et al.*, 2003), initially developed for text mining for extracting such substructures. LDA is a Bayesian version of probabilistic latent semantic analysis. In standard setting for text mining, each of D documents is modeled as a discrete distribution over T latent topics, each of which corresponds to a discrete distribution over a vocabulary of V words. For a document d, the distribution over topics, denoted by θ_d, is drawn from a Dirichlet distribution $Dir(\alpha)$, and for each topic t, the distribution over words, denoted by ϕ_t, is drawn from a Dirichlet distribution $Dir(\beta)$. A generative process in LDA is defined on document d as follows (note that the index d for document d is omitted for simplification):

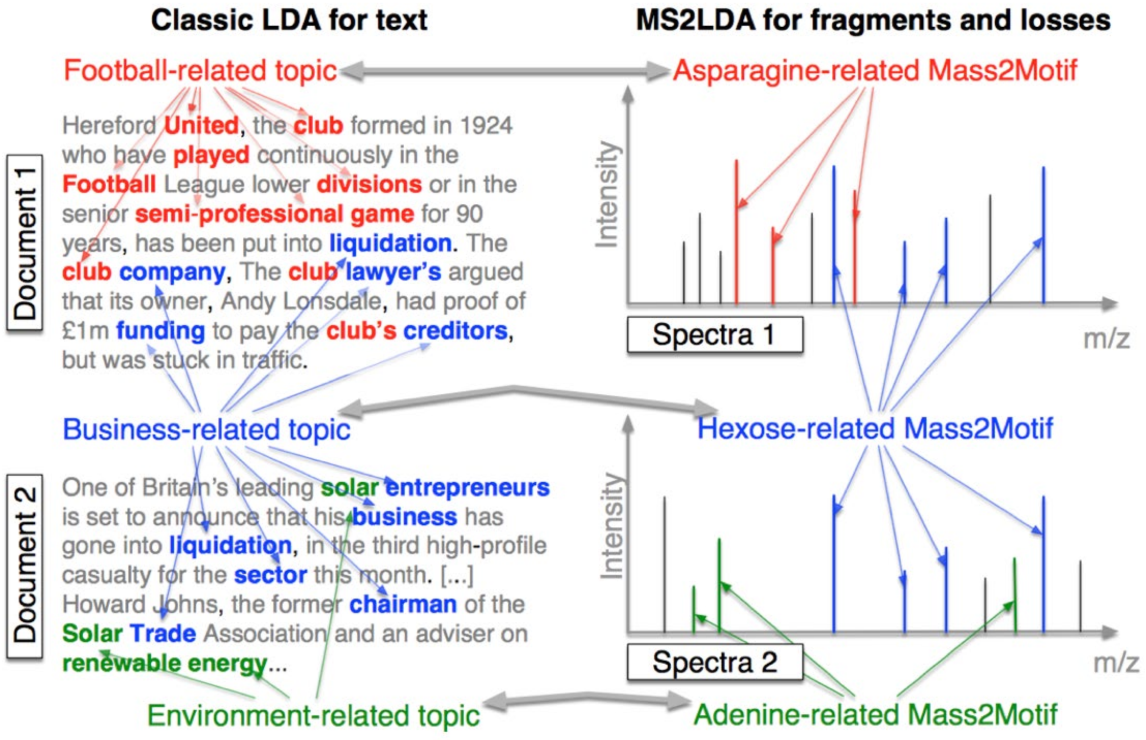
i) Choose $\theta \sim Dir(\alpha)$.

ii) For each word w_i in document d:

   a. Choose a topic $z_i \sim Multinomial(\theta)$.

   b. Choose a word $w_i \sim Multinomial(\phi_{z_i})$.

Where latent variable z_i is a topic assignment for i^{th} word w_i in the document d. The parameters to be learned include α and β. The graphical representation of this process is illustrated in Figure 9.


Classic LDA for text
MS2LDA for fragments and losses
Football-related topic
Asparagine-related Mass2Motif
Document 1
Intensity
Spectra 1
m/z
Business-related topic
Hexose-related Mass2Motif
Document 2
Intensity
Spectra 2
m/z
Environment-related topic
Adenine-related Mass2Motif

Hereford United, the club formed in 1924 who have played continuously in the Football League lower divisions or in the senior semi-professional game for 90 years, has been put into liquidation. The club company, The club lawyer's argued that its owner, Andy Lonsdale, had proof of £1m funding to pay the club's creditors, but was stuck in traffic.

One of Britain's leading solar entrepreneurs is set to announce that his business has gone into liquidation, in the third high-profile casualty for the sector this month. [...] Howard Johns, the former chairman of the Solar Trade Association and an adviser on renewable energy...

Figure 10: The correspondence between LDA for text and MS2LDA for mass spectra: LDA finds topics based on the co-occurrence of words while MS2LDA finds substructures based

on the co-occurrence of mass fragments and neutral losses. This figure is adapted from (Van Der Hooft *et al*., 2016).

The correspondence between text documents and fragmentation spectra can be obviously observed from machine learning perspective. LDA decomposes a document into topics based on the co-occurring words, while MS2LDA decomposes MS/MS into patterns of co-occurring fragments and losses. Learning LDA (MS2LDA) is to extract these topics (patterns or so-called (Mass2) Motifs) as illustrated in Figure 10. For reference, either collapsed Gibb sampling (Griffiths *et al*., 2004) or Variational Bayes (Blei *et al*., 2003) can be used to assign topics (Mass2Motifs) to words (peaks). This step applied to MS/MS is called substructure annotation. By MS2LDA, each metabolite can be explained by one or more Mass2Motifs by which we can partly identify unknown metabolites via their spectra. Also, it can be used to quickly classify metabolites into functional classes without knowing the complete structures.

Table 1: Comparison of main representative methods for supervised and unsupervised learning approaches. The performance of supervised methods is evaluated by the accuracy of the returned list of candidates, whereas that of unsupervised methods is evaluated by their capability of substructure annotation.

Approaches	**Methods**	**Info. Type for learning**	**Performance**	**Training cost**	**Prediction cost**
Supervised	FingerID (Heinonen *et al*., 2012)	spectra	high	low	low
	CSI:FingerID (Dührkop *et al*., 2015)	spectra+trees	high	high	high
	SIMPLE (Nguyen *et al*., 2018)	spectra	high	low	low
	IOKR (Brouard *et al*., 2016)	spectra+trees	high	medium	medium
	OEL (Brogat-Motte *et al*., 2020)	spectra+trees	high	medium	medium
	ADAPTIVE (Nguyen *et al*., 2019)	spectra+trees	high	high	low
Unsupervised	MS2Analysis (Ma *et al*., 2014)	user-specific features	low	N/A	N/A
	MolecularNetwork (Wang *et al*., 2016)	spectra	Low	N/A	N/A
	MS2LDA (Van Der Hooft *et al*., 2016)	spectra	High (expert-driven)	N/A	N/A
	MESSAR (Mrzic *et al*., 2017)	spectra+molecular graph	High (automation)	N/A	N/A

A drawback of the aforementioned MS2LDA is that, the extracted motifs still need to be structurally annotated based on expert knowledge, which is a complex and time-consuming process. To overcome this issue, Mrzic *et al.,* (2017) introduced a method, named MESSAR, for automated substructure recommendation from mass spectra, motivated by frequent set mining (Goethals *et al*., 2005). Similar to MS2LDA, this method is also capable of extracting recurring patterns from a given set of spectra. Briefly, molecular substructures are first generated from molecular structures/graphs of metabolites in a given database, which consists of both MS/MS and corresponding molecular structures of known metabolites. Then, they are associated with fragment ions (i.e. peaks) and mass differences between peaks to construct a single data set in the transactional format. Subsequently, frequent set mining techniques are applied to this set to extract rules of the following format: peaks p (or mass difference md) can be associated with substructure s with support f and confidence c. Such rules can be used to annotate substructures with calculated scores of support and confidence for mass spectra, in which the given peaks and mass differences are observed. Moreover, the recommended substructures can also be used for ranking candidate metabolites retrieved from a database by the similarity between recommended substructures and candidate molecular structures.

Metabolites with a high number of substructures with high confidence are assigned higher ranks.

It is noteworthy that the aim of the aforementioned methods is similar, i.e. substructure annotation. While MS2LDA needs a set of unlabeled MS/MS for learning without prior information about the molecular structures, MESSAR utilizes both experimental spectra and the corresponding structures, thus, providing an automated substructure recommendation as opposed to expert-driven substructure annotation by MS2LDA. To end this section, we give a brief comparison of methods in both supervised and unsupervised approaches for substructure prediction and substructure annotation, respectively, in Table 1.

4. Conclusion

Metabolite identification is an essential part in metabolomics to enlarge the knowledge of biological systems. However, this remains a challenging task with a huge number of potentially interesting but unknown metabolites in reality. The aim of this article is to review computational techniques and software tools to deal with the task of metabolite identification with the focus on machine learning based methods used in *in silico* fragmentation and machine learning approaches, which are the key to recent progress in metabolite identification.

It is suggested in this article that statistical machine learning-based methods should be a reasonable choice for the task of metabolite identification. Especially, when the amount of spectra and molecular structure data is increasing over time, the ability of machine learning algorithms to learn and predict relationships inherent in the data will be more enhanced. For example, IOKR, ADAPTIVE and OEL are currently best ML-based tools/methods for learning useful representations for molecules and achieving the best performance in the metabolite identification task. Additionally, we also emphasize that the combination of different approaches should be also taken into account, by which we can take advantages of them for significant improvement. For example, IOKR and CSI:FingerID are using *machine learning* and *fragmentation trees to incorporate structure information from trees and ML for prediction*. Another is MetFusion, mentioned in subsection 2.2, combines the results from MassBank (mass spectra library) and MetFrag (*in silico* fragmentation) to take advantages of complementary approaches.

Acknowledgements

D. H. N. has been supported in part by Otsuka Toshimi scholarship and JSPS Research Fellowship for Young Scientists (DC2) with KAKENHI [grant number 19J14714]. C. H. N. has been supported in part by MEXT KAKENHI [grant number 18K11434]. H. M. has been supported in part by JST ACCEL [grant number JPMJAC1503], MEXT KAKENHI [grant numbers 16H02868, 19H04169], FiDiPro by Tekes (currently Business Finland) and AIPSE program by Academy of Finland.

References

[1] Wishart, D. S. (2007). Current progress in computational metabolomics. *Briefings in bioinformatics*, *8*(5), 279-293.
[2] De Hoffmann, E., Charette, J., & Stroobant, V. (1997). Mass Spectrometry: Principles and Applications.
[3] Gross, J. H. (2006). *Mass spectrometry: a textbook*. Springer Science & Business Media.
[4] McLafferty, F. W., Tureček, F., & Turecek, F. (1993). *Interpretation of mass spectra*. University science books.
[5] 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*(6), 2028-2043.
[6] Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS: An International Journal*, *20*(18), 3551-3567.
[7] Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the american society for mass spectrometry*, *5*(11), 976-989.
[8] Hill, D. W., Kertesz, T. M., Fontaine, D., Friedman, R., & Grant, D. F. (2008). Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Analytical chemistry*, *80*(14), 5574-5582.
[9] Kumari, S., Stevens, D., Kind, T., Denkert, C., & Fiehn, O. (2011). Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Analytical chemistry*, *83*(15), 5895-5902.
[10] Mistrik, R. (2004). A new concept for the interpretation of mass spectra based on a combination of a fragmentation mechanism database and a computer expert system. *Advances in Mass Spectrometry, Elsevier, Amsterdam*, *16*, 821.
[11] Schymanski, E. L., Meringer, M., & Brack, W. (2009). Matching structures to mass spectra using fragmentation patterns: are the results as good as they look?. *Analytical chemistry*, *81*(9), 3608-3617.
[12] Chen, H., Fan, B., Xia, H., Petitjean, M., Yuan, S., Panaye, A., & Doucet, J. P. (2003). MASSIS: a mass spectrum simulation system. 1. Principle and method. *European Journal of Mass Spectrometry*, *9*(3), 175-186.
[13] Gasteiger, J., Hanebeck, W., & Schulz, K. P. (1992). Prediction of mass spectra from structural information. *Journal of Chemical Information and Computer Sciences*, *32*(4), 264-271.
[14] Heinonen, M., Rantanen, A., Mielikäinen, T., Kokkonen, J., Kiuru, J., Ketola, R. A., & Rousu, J. (2008). FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, *22*(19), 3043-3052.
[15] Wolf, S., Schmidt, S., Müller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, *11*(1), 148.
[16] Ridder, L., van der Hooft, J. J., Verhoeven, S., de Vos, R. C., van Schaik, R., & Vervoort, J. (2012). Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid Communications in Mass Spectrometry*, *26*(20), 2461-2471.

[17] Gerlich, M., & Neumann, S. (2013). MetFusion: integration of compound identification strategies. *Journal of Mass Spectrometry*, *48*(3), 291-298.

[18] Kangas, L. J., Metz, T. O., Isaac, G., Schrom, B. T., Ginovska-Pangovska, B., Wang, L., ... & Miller, J. H. (2012). In silico identification software (ISIS): a machine learning approach to tandem mass spectral identification of lipids. *Bioinformatics*, *28*(13), 1705-1713.

[19] Allen, F., Greiner, R., & Wishart, D. (2015). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, *11*(1), 98-110.

[20] Heinonen, M., Shen, H., Zamboni, N., & Rousu, J. (2012). Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, *28*(18), 2333-2341.

[21] Dührkop, K., Shen, H., Meusel, M., Rousu, J., & Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences*, *112*(41), 12580-12585.

[22] Brouard, C., Shen, H., Dührkop, K., d'Alché-Buc, F., Böcker, S., & Rousu, J. (2016). Fast metabolite identification with input output kernel regression. *Bioinformatics*, *32*(12), i28-i36.

[23] 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*(13), i323-i332.

[24] Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. (2019). ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*, *35*(14), i164-i172.

[25] Mrzic, A., Meysman, P., Bittremieux, W., & Laukens, K. (2017). Automated recommendation of metabolite substructures from mass spectra using frequent pattern mining. *bioRxiv*, 134189.

[26] Van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E., & Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, *113*(48), 13738-13743.

[27] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

[28] Jebara, T., Kondor, R., & Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, *5*(Jul), 819-844.

[29] Böcker, S., & Rasche, F. (2008). Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics*, *24*(16), i49-i55.

[30] Gönen, M., & Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, *12*, 2211-2268.

[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

[32] Srebro, N., & Shraibman, A. (2005, June). Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory* (pp. 545-560). Springer, Berlin, Heidelberg.

[33] Gilmer,J. et al. (2017). Neural message passing for quantum chemistry. In: Precup,D. and Teh,Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, Volume 70 of Proceedings of Machine Learning Research. International Convention Centre, PMLR, Sydney, Australia, pp. 1263–1272.

[34] Nguyen, H., Maeda, S. I., & Oono, K. (2017). Semi-supervised learning of hierarchical representations of molecules using neural message passing. *arXiv preprint arXiv:1711.10168*.

[35] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems* (pp. 2224-2232).

[36] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005, October). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* (pp. 63-77). Springer, Berlin, Heidelberg.

[37] Brogat-Motte, L., Rudi, A., Brouard, C., Rousu, J., & d'Alché-Buc, F. (2020). Learning Output Embeddings in Structured Prediction. *arXiv preprint arXiv:2007.14703*.

[38] Ma, Y., Kind, T., Yang, D., Leon, C., & Fiehn, O. (2014). MS2Analyzer: a software for small molecule substructure annotations from accurate tandem mass spectra. *Analytical chemistry*, *86*(21), 10724-10731.

[39] Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., ... & Porto, C. (2016). Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature biotechnology*, *34*(8), 828-837.

[40] Watrous, J., Roach, P., Alexandrov, T., Heath, B. S., Yang, J. Y., Kersten, R. D., ... & Moore, B. S. (2012). Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences*, *109*(26), E1743-E1752.

[41] Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N., ... & Wong, W. R. (2013). Molecular networking as a dereplication strategy. *Journal of natural products*, *76*(9), 1686-1699.

[42] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[43] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*(suppl 1), 5228-5235.

[44] Goethals, B. (2005). Frequent set mining. In *Data mining and knowledge discovery handbook* (pp. 377-397). Springer, Boston, MA.

Dai Hai Nguyen received his B.S. degree in Computer Science from Hanoi University of Science and Technology, Hanoi, Vietnam, and Ph.D. degree from Kyoto University, Japan. He has been working on machine learning and its applications. His current research interests are machine learning for biological data.

Canh Hao Nguyen received his B.S. degree in Computer Science from the University of New South Wales, Australia, M.S. and Ph.D. degrees from JAIST, Japan. He has been working in machine learning and bioinformatics. His current interests are machine learning for graph data, sparse modeling with applications in biological network analysis.

Hiroshi Mamitsuka received the B.S. degree in biophysics and biochemistry, the M.E. degree in information engineering, and the Ph.D. degree in information sciences from the University of Tokyo, Tokyo, Japan, in 1988, 1991, and 1999, respectively. He has been working on research in machine learning, data mining, and bioinformatics. His current research interests include mining from graphs and networks in biology and chemistry.