Universiteit Antwerpen

This item is the archived peer-reviewed author-version of:

Current and future deep learning algorithms for MS/MS-based small molecule structure elucidation

Reference:
Liu Youzhong, de Vijlder Thomas, Bittremieux Wout, Laukens Kris, Heyndrickx Wouter.- Current and future deep learning algorithms for MS/MS-based small molecule structure elucidation
Rapid communications in mass spectrometry - ISSN 0951-4198 - (2021), e9120
Full text (Publisher's DOI): https://doi.org/10.1002/RCM.9120
To cite this reference: https://hdl.handle.net/10067/1780250151162165141

uantwerpen.be
Institutional repository IRUA

Liu Youzhong (Orcid ID: 0000-0002-1449-3841)

De Vijlder Thomas (Orcid ID: 0000-0002-8692-425X)

Laukens Kris (Orcid ID: 0000-0002-8217-2564)

Current and future deep learning algorithms for MS/MS-based small molecule structure elucidation

Youzhong Liu[1*], Thomas De Vijlder[1], Wout Bittremieux[2,3,4], Kris Laukens[2,3], Wouter Heyndrickx[1]

1. Janssen Research & Development, Beerse, Belgium, 2. University of Antwerp, Antwerp, Belgium, 3. Biomedical Informatics Network Antwerpen (biomina), University of Antwerp, Antwerp, Belgium, 4. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, San Diego, CA, United States of America

*Correspondence to: Y. Liu, Janssen Research & Development, Turnhoutseweg 30, 2340 Beerse, Belgium.
E-mail : YLiu186@its.jnj.com

Abstract

RATIONALE: Structure elucidation of small molecules has been one of the cornerstone applications of mass spectrometry for decades. Despite the increasing availability of software tools, structure elucidation from MS/MS data remains a challenging task, leaving many spectra unidentified. However, as an increasing number of reference MS/MS spectra are being curated at a repository scale and shared on public servers, there is an exciting opportunity to develop powerful new deep learning (DL) models for automated structure elucidation.

ARCHITECTURES: Recent early-stage DL frameworks mostly follow a "two-step approach" that translates MS/MS spectra to database structures after first predicting molecular descriptors. The related architectures could suffer from: 1) computational complexity because of the separate training of descriptor-specific classifiers, 2) the high dimensional nature of mass spectral data and information loss due data preprocessing, 3) low substructure coverage and class imbalance problem of predefined molecular fingerprints. Inspired by successful DL frameworks employed in drug discovery fields, we have conceptualized and designed hypothetical DL architectures to tackle the above issues. For 1), we recommend multitask learning to achieve better performance with fewer classifiers by grouping structurally related descriptors. For 2) and 3), we introduce feature engineering to extract condensed and higher-order information from spectra and structure data. For instance, encoding spectra with subtrees and pre-calculated spectral patterns add peak interactions to the model input. Encoding structures with graph convolutional networks incorporates connectivity within a molecule. The joint embedding of spectra and structures can enable simultaneous spectral library and molecular database search.

CONCLUSIONS: We believe that in principle, given enough training data, adapted DL architectures, optimal hyperparameters and computing power, DL frameworks can predict small molecule structures, completely or at least partially, from MS/MS spectra. However, their

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/rcm.9120

This article is protected by copyright. All rights reserved.

performance and general applicability should be fairly evaluated against classical machine learning frameworks.

Keywords

Small molecule, Structure elucidation, Spectral library, Deep learning, Feature engineering

Abbreviations

CNN, Convolutional neural network;

DDA, Data dependent acquisition;

DIA, Data independent acquisition;

DL, Deep learning;

DNN, Deep neural network;

ECFP, Extended-connectivity fingerprint;

FT, Fragmentation tree;

GAN, Generative adversarial network;

GPU, Graphics processing unit;

MFP, Molecular fingerprint;

ML, Machine learning;

MLP, Multi-layer perceptron;

MPNN, Message-passing neural network;

MTL, Multitask learning;

PLS-DA, Partial least square discriminative analysis;

QSAR, Quantitative structure activity relationship;

ReLU, Rectified linear activation function;

RNN, Recurrent neural network;

SMILES, Simplified molecular input line entry specification;

SVM, Support vector machine;

VAE, Variational autoencoder;

Thesaurus

Machine Learning Machine Learning (ML) algorithms build a statistical model based on sample data (or training data) to make predictions or decisions without explicit instructions. ML technology powers many aspects of society: from web searches to e-commerce to business decision-making. In many research projects, ML assists complex data analysis and stimulates fundamental biological discoveries. ML can provide automated solutions for MS/MS-based small molecule structure elucidation to assist or partially replace manual spectral interpretation. Community-scale spectral libraries contain a large

This article is protected by copyright. All rights reserved.

amount of sample data, including both spectra and corresponding structure labels. Ideally, we can translate the spectral library into a statistical model that captures and records the complex relationships between spectra and structures or structure properties in a robust and reliable fashion. Without explicit instructions from experts, the ML model created can automatically predict structures or structure descriptors for an unseen spectrum. ML models can be further developed into a user-friendly structure elucidation software.

Deep Learning Deep Learning (DL) is a subfield of ML technology. DL algorithms, due to their complex architectures, are considered as an evolution of statistical models in standard ML. In fact, DL applications use a multiple layer structure of algorithms called a Deep Neural Network (DNN). The concept of a DNN is inspired by the biological neural network of the human brain, leading to a process of learning that is more sophisticated and flexible than that of traditional ML models. DL methods have considerably improved the state-of-the-art in speech recognition, computer vision and many other domains such as drug discovery and genomics. Likewise, DL may improve computational frameworks for MS/MS-based small molecule structure elucidation. An advantage of DL over standard ML is that artificial neural networks might better recognize complex spectral patterns observed from library spectra, such as peak interactions, and map these patterns to small molecule structures or structure properties.

Supervised and Unsupervised Learning Within the field of ML (DL), there are two major learning tasks: supervised or unsupervised. The main difference between these is that supervised learning requires ground truth, or in other words, the training data contains expected output values for each sample. The goal of supervised learning is to learn a statistical model that best approximates the relationship between the input (e.g. library spectra) and output (e.g. molecular descriptors) from the training data. Unsupervised learning, on the other hand, does not require labeled outputs, and its goal is to infer the natural structure or condensed information present within a set of data inputs (e.g. unlabeled library spectra). The natural structure can be in the form of patterns (e.g. MS2LDA motifs and MESSAR rules) or vectors in an embedded space (as for Spec2Vec).

Kernel To develop either supervised or unsupervised ML(DL) algorithms for MS-based structure elucidation, the common challenge is to put similar training spectra or structures close to each other. MS/MS spectra, after centroiding and *m*/*z* discretization, can be represented as vectors of numbers. Therefore, the dot product between two vectors can be a simple similarity measure between two spectra. However, chemical structures in the spectral library are not easily convertible into numeric vectors, and neither are fragmentation trees built from MS/MS spectra. Kernel methods are a class of algorithms that extend the similarity measure to virtually any type of samples including these non-vectorizable ones. Kernel methods require only a user-specified kernel (a similarity function) over pairs of data points (e.g. pairs of fragmentation trees), which enable them to perform prediction or dimension reduction tasks in an implicit feature space.

Parameters and Hyperparameters Model parameters are the parameters internal to the ML(DL) model whose values are learned from training data. In fact, the values of the parameters determine how accurately the model fits the training data and how good it performs the prediction task. During the learning phase, parameters are estimated automatically by minimizing a loss function. The parameters of DNNs are typically the weights of the connections between neurons. Hyperparameters are configurations of the model whose values cannot be estimated from data. Before the learning phase, they are optimized for a given predictive modeling task and set by the practitioner. They usually remain unchanged throughout the learning and prediction phase. The aim of hyperparameter tuning is to find the combination that strikes the balance between underfitting and overfitting on the validation set in the hope that these results generalize to the testing set. The hyperparameters of

This article is protected by copyright. All rights reserved.

DNNs are related to both network structure (e.g. layers, units, and dropout rate) and training algorithm (e.g. learning rate, momentum, number of epochs, and batch size).

Multi-Layer Perceptron An Artificial Neural Network (ANN) in DL technology is a network connecting units or nodes called artificial neurons, which loosely model the neurons in human brains. The input of each neuron is a real number (e.g. intensity of a product ion), and the output of each neuron is computed by non-linear transformations of the sum of its inputs (e.g. non-linear functions on intensities of product ions). The basic example of an ANN is a Multi-Layer Perceptron (MLP), which consists of at least three layers of neurons: the input layer, one or more hidden layers, and the output layer. The input layer receives the numeric input data to be processed, for example, the intensities of all *m/z* bins. The result of the prediction task, such as the presence or absence of molecular descriptors, is given by the output layer. A pre-defined number of hidden layers that are placed between the input and output layers are the true computational engine of the MLP. In an MLP, the data flows in the forward direction from the input layer to the output layer, while the weights of the neurons are trained via back-propagation. MLPs are designed to approximate any continuous function to recognize complex spectral patterns observed from library spectra, such as peak interactions, and map these patterns to molecular descriptors in the output layer. An ANN, including a basic MLP, can be called a deep neural network (DNN), if there are multiple hidden layers between the input and output layer.

Graph Neural Network Graph Neural Networks (GNN) are a type of ANN that directly operate on a graph structure. In computer science, a graph is a data structure consisting of two components: nodes and edges. Molecular graphs are the most natural representation of small molecules by considering atoms as nodes and bonds as edges. Among diverse GNN architectures, the Message Passing Neural Network (MPNN) is a promising and recent approach in QSAR modelling for engineering molecular descriptors from molecular graphs. They have outperformed predefined molecular descriptors, such as ECFP, for building accurate models for the prediction of various properties and activities of small molecules. A MPNN framework usually contains four steps: (1) feature initialization step, where chemical information of each atom and bond is coded as feature vectors, (2) message passing step, where, for each atom, atom or bond features from its neighbors are propagated, based on the graph structure, into a so-called message vector; (3) update step, where atom features are updated by the message vector; (4) aggregation step, where atom features in the molecule are aggregated into a molecule feature vector. The output of an MPNN (i.e. molecule feature vector) can be obtained by maximizing its correlation with the spectral data. A good example is the “molecular vector” in the ADAPTIVE framework.

CNN and RNN Besides the graph-based representation, different DNN architectures can use other types of molecular representations as input. Among such DNN architectures, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have both gained remarkable success in QSAR and drug discovery. For instance, we can use CNNs on images of 2D drawings of molecules. SMILES strings can be directly used as the input to RNNs. RNNs are usually applied in combination with attention mechanisms to better process long SMILES strings. Like GNNs, CNNs and RNNs can both learn features from structures directly without using any predefined structure descriptors. In fact, the architectural flexibility is a prominent benefit of DL compared to other ML methods.

This article is protected by copyright. All rights reserved.

1. Introduction

Mass spectrometry (MS) is being used across diverse fields that require structure elucidation of small molecules, such as metabolomics, pharmaceutical discovery and development, forensic toxicology, environmental science, and clinical research[1]. As examples, metabolite identification is an essential step towards the understanding of complex biological systems and biomarker discovery[2,3]. The pharmaceutical industry must investigate and identify all process impurities and degradation products to meet the requirements of regulatory authorities[4]. Drug discovery from natural products and toxicity analysis cannot proceed without correct structure proposals of isolated compounds[5].

One of the standard methods for identifying small molecules (below 1,500 Da) is liquid chromatography (LC) techniques coupled with tandem mass spectrometry (MS/MS). Data acquired from LC-MS/MS platforms capture each analyte in several layers of information, including exact mass, retention time, isotopic profile, and fragmentation patterns (both MS2 and MSn). Modern tandem MS technologies provide data dependent acquisition (DDA) and data-independent acquisition (DIA) functionalities to acquire multiple stages of mass analysis including MS1, MS2 (MS/MS), and MS^n spectra for small molecules present in a complex mixture[6]. Such experiments start with measuring precursor ions of intact small molecules (MS1 stage) in a sample. In the MS2 stage, some precursor ions are selected, isolated and fragmented, and the product ions detected are used to generate the MS2 spectrum of a selected precursor ion. With some MS instruments, if the intensity of product ions allows, this process may be extended by selectively retaining product ions in the mass analyzer and repeating the fragmentation process to generate the MS3 or even MS^n ($n>3$) spectra.

Hundreds of small molecules can be screened from a complex sample in minutes on an LC-MS/MS system. Meanwhile, mass spectra of authentic standards and real-world samples are being stored in spectral libraries after manual curation and annotation by experts[7]. Most spectral libraries record selected MS1/MS2 scans (two-column matrices: mass, intensity), analytical techniques (e.g. instrument, gradient), measurement in other dimensions (e.g. retention time, drift time) along with structure metadata provided by experts (e.g. formula, adduct, InChIKey). MS3 and MS^n scans are also recorded in some spectral libraries such as mzCloud[8]. Repositories exist to gather spectral libraries from individual laboratories into larger resources. Community-scale spectral repositories, such as GNPS[7], MassBank[9], METLIN[10] and MetaboLights[11], collect spectra of up to 1,000,000 small molecule structures acquired on diverse instruments.

Despite tremendous efforts of the MS community to contribute manually curated spectra, only a marginal fraction of compounds present in biological systems, chemical processes, and the environment are covered by these community-scale repositories. According to experts, about 2% of spectra acquired in LC-MS/MS experiments can be annotated via direct reference spectral matching, although for well-studied biological matrices, such as *Escherichia coli*, human cell lines, plasma or urine this may recently become as high as 10%[12–14]. Therefore, alternative algorithms have been developed to deal with the unavailability of measured reference spectra in small molecule structure elucidation.

This article is protected by copyright. All rights reserved.

Among them, machine learning (ML) approaches, including deep learning (DL)-based frameworks are rapidly emerging in the field[15]. These frameworks can be distinguished from spectral library search and *in silico* fragmentation in a few aspects: 1) ML and DL models are initially created based on existing spectral libraries, with the MS/MS spectra and structure metadata used for model training, 2) subsequent compound identification does not directly rely on spectral libraries with a limited coverage; 3) unlike when applying *in silico* fragmentation, these frameworks are purely data-driven, in other words, there is no attempt to predict the fragmentation behavior through bond breaking, gas-phase chemistry, or fragmentation rules.

ML models have shown clear advantages in diverse applications that involve small molecule structure elucidation (Box 1). First, compared to classical spectral library searching, they enable an extension of the search space from smaller spectral libraries to comprehensive molecular databases (e.g. PubChem[16]) through correct prediction of substructures or other chemical properties (Figure 1). Next, since data-driven approaches do not rely on predefined rules, they might be able to capture rare, hard-to-explain fragmentation rules and unexpected patterns from the spectral library. In addition, unlike mechanistic or combinatorial models, once ML models are trained, the prediction phase typically does not require heavy computing. As a result, the compound identification can be performed in a high-throughput manner.

Despite the success of ML-based software that predict molecular substructures such as CSI:FingerID[17], key drawbacks have been raised, for example, by Nguyen and colleagues[15]. They discuss how the complexity of existing spectral library data often leads to limitations related to dimensionality, higher-order interaction, interpretability, and sparsity, which are partially overlooked by current ML models (see details in Section 4). Meanwhile, spurred by the increasing availability of training data (expansion of available spectral libraries) and improved computational resources, deep neural networks (DNN) are currently being introduced in the field as an alternative to classical ML models such as linear regression, discriminant analysis, and support vector machine (SVM). The consensus is that, with enough training data, DL models can outperform related ML models for the same application. However, in our opinion, the conceptual drawbacks of ML for structure elucidation must be carefully addressed before seeing a clear improvement with DL-based methods.

In this perspective paper, we will first revisit typical ML-based frameworks by analyzing their underlying reasoning, architectures, and drawbacks. Potential DL implementations are then proposed with inspiration from early-stage developments in the field as well as popular architectures used in quantitative structure activity relationship modelling (QSAR)[18–20]. These architectures are designed to not only improve prediction accuracy and computational performance, but also to address, to a certain extent, limitations of current ML-based frameworks. Finally, we will discuss potential risks, challenges, and future directions for implementing DL and ML technologies in general.

2. Revisiting machine learning strategies for structure elucidation

The first computational efforts to extend identified small molecules beyond the spectral libraries are *in silic*o fragmentation tools. These tools attempt to predict theoretical MS/MS spectra of compounds outside spectral libraries using expert knowledge fragmentation rules and on calculations of standard bond energy. Since the fragmentation behavior of small molecules on LC-MS/MS systems is complex and sometimes hard to predict, experimental MS/MS spectra were used in some tools, such as ISIS[21] and CFM-ID[22], to parameterize and

This article is protected by copyright. All rights reserved.

improve physical models that mimic fragmentation processes. Although these computational frameworks are sometimes considered as ML-based methods and haven shown great potential in structure elucidation, they are not included in the scope of the paper. Instead, we focus on a class of "black-box" algorithms called "supervised learning for substructure prediction". The underlying models are often built and trained using large spectral datasets without explicitly encoding physicochemical knowledge of the fragmentation process. Instead, each small molecule is transformed into a molecular descriptor according to predefined rules. A molecular descriptor in these algorithms is a binary vector indicating the presence or absence of chemical properties (in CSI:FingerID[17]) and substructures (CSI:FingerID & MESSAR[23]), or the membership to certain chemical classes (ClassyFirePredict for MS2LDA[24]).

Related computational frameworks follow similar learning and prediction procedures (Figure 1). The modeling or learning phase uses MS/MS spectra of known molecules, sometimes transformed, for instance, to fragmentation trees (FTs), or combined with data from other analytical dimensions. The presence/absence of molecular descriptors is learnt in a supervised way. Most frameworks use binary classifiers such as partial least square discriminative analysis (PLS-DA)[25], decision tree[26] or SVM. During the prediction phase, ML models predict the probability of presence of each molecular descriptor for a query spectrum. Meanwhile, by searching a massive molecular database with precursor *m/z* or formula, a limited set of structure candidates are retrieved. In CSI:FingerID, these structures can be further scored by comparing their corresponding molecular descriptors to the predicted probabilities. As the output, candidates whose molecular descriptors agree with the predicted probabilities are provided as potential annotations of the query spectrum.

Accepted Article

This article is protected by copyright. All rights reserved.

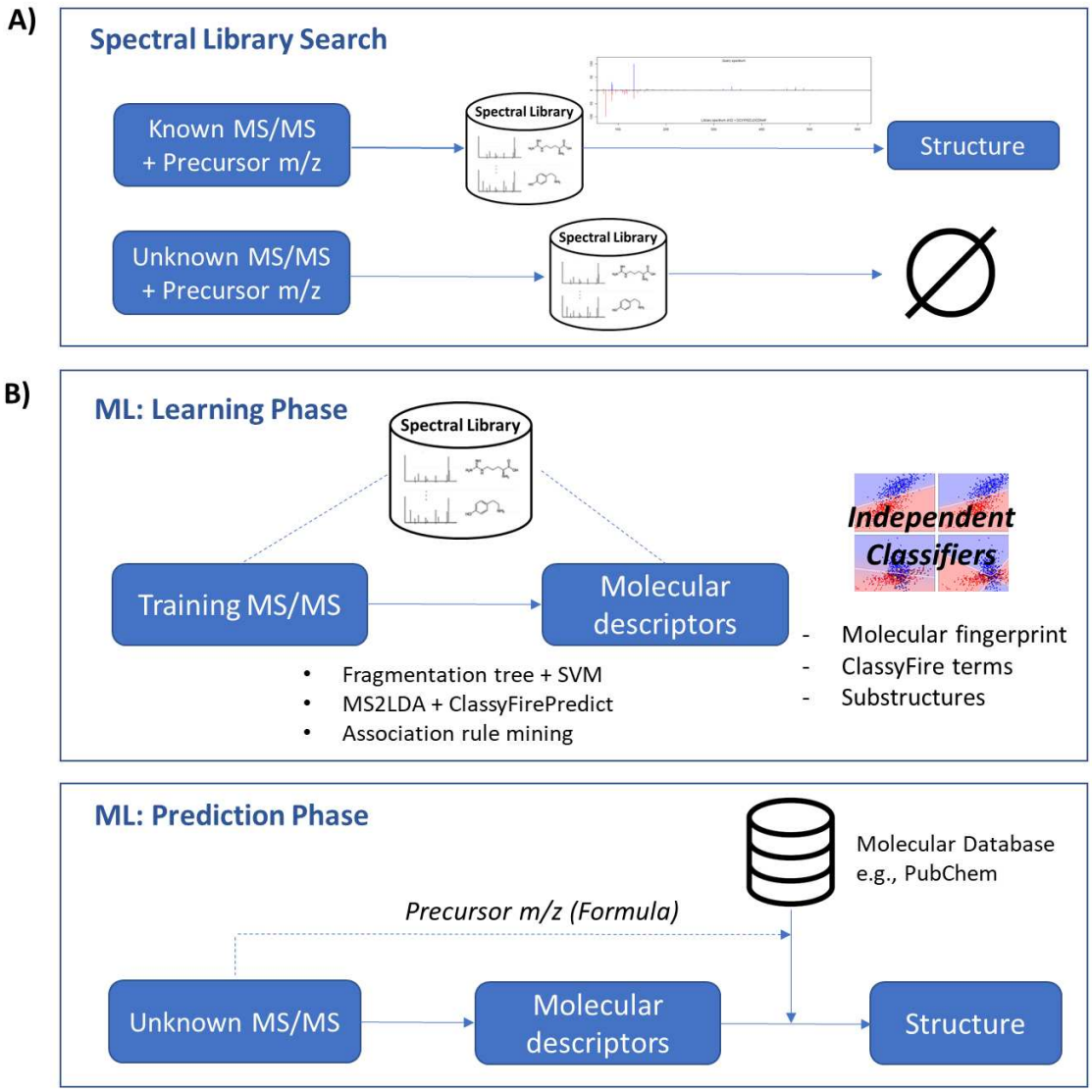

Figure 1. Current spectral library search and machine learning concepts for MS-based small molecule structure elucidation. A) Spectral library search compares the query spectrum to reference spectra of structure candidates retrieved from a spectral library (e.g. via precursor m/z). Candidates are scored and ranked based on the spectral similarity with the query spectrum. Therefore, if the molecule to be identified is not yet included by the spectral library, no accurate structure proposal can be made via spectral library search. B) ML can extend the structure candidate retrieval from a spectral library to an entire molecular database (with considerably more structures available). Current ML frameworks use a two-step supervised learning approach that consists of a learning and a prediction phase. First, training spectra in a spectral library are pre-processed using FTs[27] (CSI:FingerID), text mining[28] (MS2LDA), or frequent pattern mining[23] (MESSAR). Meanwhile, the corresponding training molecules are encoded into molecular fingerprint bits, ClassyFire terms[29] (chemical classes) or broken into substructures. Next, independent binary classifiers are trained and validated using transformed spectra as input and the presence/absence of each molecular fingerprint bit, ClassyFire term, or substructure as output. CSI:FingerID used SVM on kernel similarities of FTs to predict molecular fingerprint bits. MS2LDA was combined with a multilayer neural network called ClassyFirePredict to predict ClassyFire terms[24]. Patterns of spectral features and substructures were mined by MESSAR and converted to association rules (binary classifiers). During inference with these ML frameworks, the probability of each molecular descriptor is computed using the trained models. To make structure proposals, the precursor *m/z* or the exact formula of an unknown MS/MS spectrum is searched in a molecular database, after which the retrieved candidates are scored and ranked based on the likelihood of agreement between the inferred and retrieved molecular descriptor. Only the top candidates are proposed to users as putative annotations. Currently, the last step (structure proposal) is only available in CSI:FingerID (through molecular fingerprint prediction). However, searching molecular databases and scoring candidates can be implemented for MS2LDA and MESSAR as well, via ClassyFire term and substructure prediction, respectively.

This article is protected by copyright. All rights reserved.

These ML methods belong to a "two-step approach" since predefined molecular descriptors, including molecular fingerprints (MFPs), substructures or chemical classes can be used as intermediate without learning the direct mapping between spectra and small molecule structures. In other words, full structure elucidation of unknown spectra is performed in a separate phase (i.e. descriptor prediction and candidate scoring) after training and validating ML models. Interestingly, molecular descriptors are also a popular intermediate step of ML models for QSAR. Since the prediction of chemical properties or biological activity directly from chemical structures is not yet fully explored, the QSAR field uses numerous approaches to first encode a drug candidate into a list of molecular descriptors[30]. One avenue is to look at descriptors related to electronic, topological, spatial, or physicochemical aspects of part of the molecule, or the complete molecule[31]. The possibilities are extensive, and examples include the HOMO-LUMO gap, dipole moment, polarizability, polar surface area, molecular volume, LogP, molecular weight, number of H-bridge donors/acceptors, number of bond types, max positive/negative charge on atom, sum of positive/negative partial charge or graph theoretical connectivity indexes[32]. MFPs are another approach to representing molecular structures. Although originally developed for data retrieval and similarity matching, such MFPs have proven valuable in QSAR modelling as well. Some prominent examples include: 1) MACCS[33], where the substructure encoding is based on a predefined dictionary, 2) ECFP[34], where the atom environments are encoded exhaustively in a combinatorial fingerprint, and 3) enumeration of pharmacophore fragments[35]. For a contemporary overview, including 3D-based MFPs, the reader can refer to[36,37].

Over the past decade, there has been a remarkable growth in experimental and biomedical data related to compound activity. DL has taken advantage of the increased QSAR data availability and improved computational power to outperform previous state-of-the-art ML algorithms such as random forest and SVM[38,39] and more recently gradient boosted decision trees[40]. The main advantages of DL-based methods are: (i) DNNs can automatically learn complex patterns from raw data including potential interaction between thousands of input features; (ii) DL provides much more flexible architecture so it is possible to create a DNN architecture tailor-made for a specific problem[41]; (iii) Many existing algorithms and libraries can take advantage of graphics processing units (GPUs) to accelerate the training of DL models and to optimize their predictive performance[20]. A subtle advantage of DNN over classical ML models is that striking a balance between bias and variance is usually possible by iteratively fine-tuning model hyperparameters (because more hyperparameters are available for optimization than for an average ML model). However, finding optimal hyperparameters to avoid overfitting and underfitting requires expertise and extensive trial and error, making the process time-consuming for some DL applications[41].

3. Enhancing the two-step machine learning strategy by deep learning

The ever-increasing size and quality of publicly available MS spectral libraries has enabled the development of several early-stage DL frameworks for small molecule identification, notably in the metabolomics field. These tools follow the two-step concept centered around the prediction of structure descriptors from MS/MS data (Figure 2, Architecture 1). For example, DeepEI[42] proposed a DL approach to predict the retention index and MFPs of unknown compounds from GC-EI-MS data. Using 184,874 training spectra as input, DeepEI built separate models for each of the 636 fingerprint bits (selected from 6 types of MFPs) calculated from training structures. Each model was a multi-layer perceptron (MLP) containing three fully connected hidden layers (units = 2000, 1000, and 500, respectively) with ReLU activation functions. The output was presented as a fully connected layer with a SoftMax activation function. When compared to alternative approaches under identical conditions (same

This article is protected by copyright. All rights reserved.

training/testing set, same number of classifiers), the DL-based approach clearly outperformed linear regression, discriminant analysis, and gradient boosted decision trees. In addition, DeepEI is effective for identifying the correct structure of unknowns by searching predicted MFPs in a molecular structure database.

Another tool built under a similar framework was MetFID[43]. Trained with merged (LC-)MS/MS spectra of 15,228 known compounds (input), MetFID aimed at predicting 528 bits from three types of pre-calculated MFPs (output). The neural network consisted of two ReLU hidden layers (units = 800 and 600). In contrast to other "two-step" ML methods (including DeepEI), MetFID allows the prediction of all fingerprint bits at once by inserting a 528-dimensional multiple label layer (1 for every fingerprint bit). This layer involves a sigmoid activation function and was trained using the binary cross-entropy loss function. MetFID has demonstrated the potential to improve metabolite putative annotation in combination with other ML-based software.

Like predicting all fingerprint bits, many QSAR studies aim at predicting multiple potentially related tasks at once (e.g. affinity of drugs with similar tumor receptors). This DL approach is called multitask learning (MTL), in which related tasks are learnt in parallel by constructing a common representation over these tasks[39]. Due to the extra information in the training signals, learning related tasks together can enable knowledge transfer and data across tasks, thus improving overall predictive performance[44]. The underlying mechanism by which these benefits occur has been thoroughly discussed[45]. In QSAR modeling, for realistically-sized datasets in drug discovery (100,000s of compounds, several 100s of tasks, millions of data points), network architectures of 1 to 4 layers with thousands of nodes per layer are performing satisfactorily[41].

Although the context is fundamentally different, the multi-label learning in MetFID is comparable to a typical MTL implementation, consisting of classification with attributes limited to binary labels. Besides predicting all bits of MFP simultaneously, MetFID has a clear advantage of learning the latent hidden relationships between elements within a fingerprint. However, when there is no task relatedness, a risk of deteriorated performance of a multitask model as compared to a respective single-task model exists[44,46]. This risk is called "negative transfer" in technical terms.

An aspect of MetFID that should be addressed, according to the authors[43], is the scoring and ranking of structure candidates via MFP prediction. In practice, similar structure candidates could be better discriminated by adding other MFP types into the MetFID framework. However, a decrease of prediction performance was observed when certain types of MFPs were included. This issue can be partially explained by the negative transfer between MFP bits. To avoid negative transfer, we recommend extra shared layers in the MetFID architecture by grouping similar molecular descriptors (Figure 2, Architecture 2). The hypothesis underlying this strategy is that molecular descriptors can have very close chemical meaning or structural representations. For instance, the bit in position 758 of PubChem fingerprint, the ClassyFire term "Indoles and derivatives", and the substructure with SMILES code "c12c([nH]cc1)cccc2" are all related to tryptophan-like structures, and they should be captured by similar patterns of fragments and neutral losses[23]. Therefore, training MTL models by groups of molecular descriptors is a way to maximize task relatedness and thus overall model performance. Practically speaking, the underlying links between different molecular descriptors are not fully understood, thus the expert-driven grouping can be quite labor intensive. We recommend exploring data-driven task grouping, for instance, by clustering molecular descriptors based on their coappearance in training structures.

This article is protected by copyright. All rights reserved.

Accepted Article

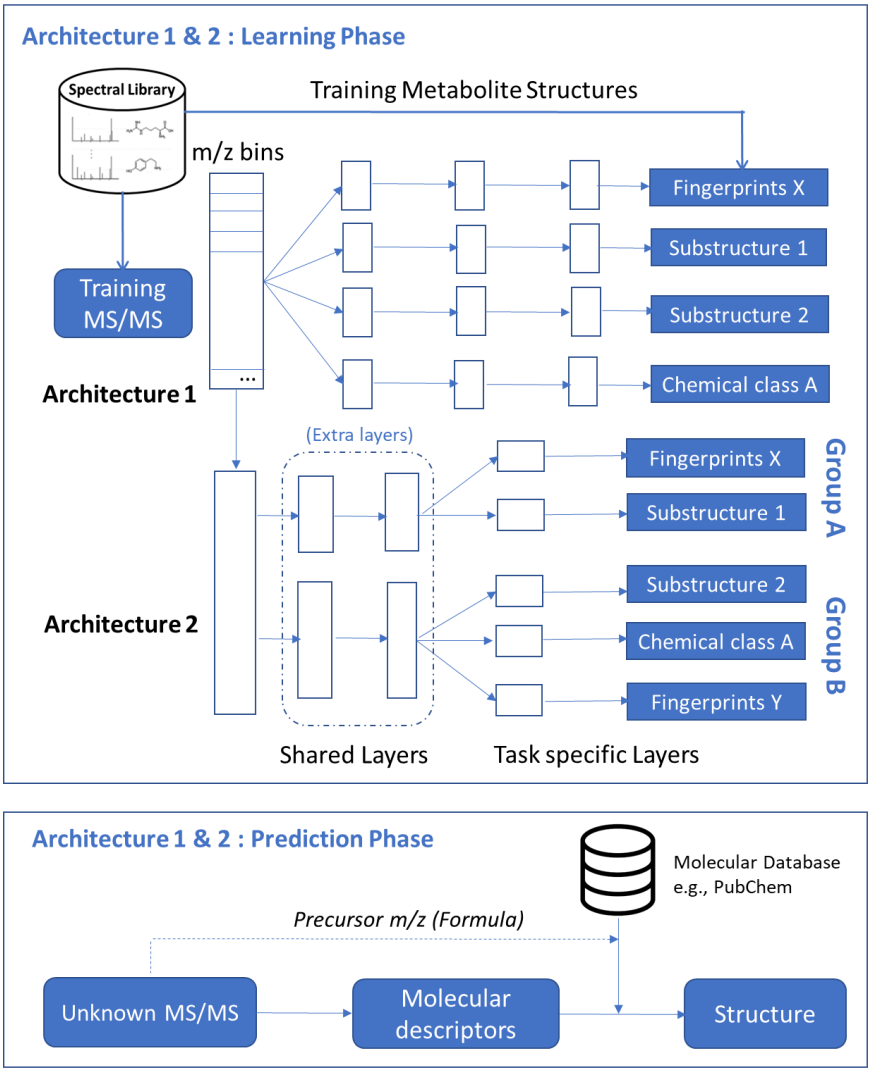

Figure 2. Deep learning architectures to enhance the two-step supervised ML approach for structure elucidation. In the learning phase, training MS/MS spectra are converted into vectors with a fixed dimension by dividing the *m/z* range into bins. Molecular descriptors are determined in the same way as for traditional ML approaches. In Architecture 1 (implemented by DeepEI), individual DNNs (with different hyperparameters) are built to learn the relationship between spectral patterns and each molecular descriptor. In Architecture 2, we suggest a MTL architecture that consists of shared layers and task-specific layers to predict molecular descriptors. Architecture 2 is an augmentation of the current MetFID architecture by adding shared layers in front of groups of chemically related molecular descriptors. The prediction phase of both Architecture 1 and 2 is identical to other two-step supervised ML frameworks.

4. Major limitations of the two-step machine (deep) learning strategy

Despite the clear success of software based on the two-step ML strategies such as CSI:FingerID[17], some key drawbacks hereof have been raised in Nguyen et al. (2018)[15]. In our opinion, the afore-mentioned DL frameworks are facing similar limitations and challenges. We summarize them into two major drawbacks:

1) The early-stage DL framework usually starts by data pre-processing: dividing the m/*z* range into bins to create high-dimensional feature vectors (Figure 2). In practice, the optimal bin width is hard to determine. While wider bins can introduce noise to higher resolution MS data (e.g. loss of specificity of bins to exact masses), too narrow bins result in extremely large vectors that are hard to handle by DL frameworks. Narrow

This article is protected by copyright. All rights reserved.

bins also lead to alignment errors, i.e. the same fragment is assigned to different bins. Moreover, higher-order interactions between fragments and neutral losses are lost during the discretization process. Depending on the type of fragmentation technique applied, due to consecutive fragmentation processes, peaks in MS/MS spectra can have a hierarchical relationship starting from a precursor ion to higher and subsequently lower *m/z* product ions[27,47]. The underlying links between product ions are clearly reflected by their frequent co-occurrence observed in spectral libraries[23]. Introducing higher-order interaction as domain knowledge as encoded features, offers the possibility to improve current DL frameworks without the need for massive training data. Several spectral encoding strategies will be introduced in Section 5.

2) The focus of both ML and early-stage DL frameworks is to predict the presence/absence of predefined molecular descriptors to score compound candidates. Despite the successful use of molecular descriptors in metabolite annotation software for certain classes of molecules, it is known from QSAR and virtual screening that many molecular descriptors struggle to achieve equally high performances on all classes of compounds in various settings[48]. MFPs can be limited in representing higher-mass fragments and neutral losses since they rarely encode larger substructures that confer to higher *m/z* product ions. Alternatively, some ML frameworks use predefined chemical rules to generate information-rich substructures. However, to our knowledge, no single set of rules can cover all relevant substructures in a spectral library[49]. Another challenge of using molecular descriptors is the class imbalance problem. For instance, DeepEI opted to select a subset of fingerprint bits for DNN training since the rest was either present in very few or almost all training structures. Moreover, the use of rare molecular descriptors can hinder model validation due to the lack of their presence in testing samples. In section 6, we will go through advanced structure encoding methods to tackle the afore-mentioned challenges. Also, DL architectures that avoid completely predefined molecular descriptors (descriptor-free architectures) will be proposed in section 5 and 6.

5. MS/MS spectral encoding to improve DL frameworks

DL frameworks nowadays often start with a data pre-processing step to make raw data more amenable to DNN architectures. We mainly focus on MS/MS spectra in this section although input data can include other layers of information such as adduct type, retention time, precursor formula etc. An MS/MS peak list consisting of *m/z* and intensity pairs can be considered as a virtual (sparse) vector, and they can be easily discretized into a *m/z* vector with a fixed dimension by binning (e.g. vector of length 100,000 for the *m/z* range 0-1000 Da with a bin size of 0.01 Da) (Figure 2).

Based on discretized mass values from data pre-processing, we can develop sophisticated spectral encoding methods to extract underlying information from MS/MS spectra and to improve overall performance of DL frameworks. The task is called feature engineering in ML and DL terms, and its goal is to build information-rich features using domain knowledge and data mining techniques. In the context of the perspective paper, “features” indicate all types of new spectral information created from raw MS/MS spectra that can be served as inputs of DNN architectures. “Feature engineering”, an intermediate step between data pre-processing and DNN training, is defined as the transformation from pre-processed MS/MS spectra to spectral encodings. This transformation is predefined and hard coded by the network designer, and not modifiable throughout the learning process.

This article is protected by copyright. All rights reserved.

According to these definitions and requirements, many existing metabolite annotation tools provide means to engineer information-rich features. For instance, both MESSAR[23] and MS2LDA[28] extract patterns or motifs featuring co-occurring mass fragments and neutral losses (Figure 1). In addition, MESSAR extracts spectral patterns potentially related to a substructure, while MS2LDA provides expert interpretation of motifs in terms of fragmentation events, substructures, and compound classes. Elemental formulas of fragments computed by CSI:FingerID through fragmentation tree building can serve as input features to DNNs (Figure 3B-③).

A multi-layer perceptron (MLP) can be imagined with structurally relevant patterns or motifs as input nodes (replacing bins representing individual mass peaks) and pattern overlapping scores as node values (instead of intensities) (Figure 3B-②). The new MLP has clear advantages over the "binned" version (Figure 2): 1) before model training, domain knowledge is efficiently used in different data preparation steps including calculation of neutral losses, determination of theoretical masses of product ions and removal of non-monoisotopic, noisy or structurally-irrelevant mass peaks etc. 2) the dimensionality is reduced from, for example 100,000 input variables to, a few thousand MESSAR patterns or a few hundred MS2LDA motifs; 3) the sparsity of input features decreases since only frequent patterns or motifs are kept; 4) peak interactions (based on their co-occurrence related to a shared substructure or chemical class) are captured and fed as supplementary information to the MLP.

The previously described pattern mining algorithms for feature engineering consider MS/MS spectra as one-dimensional vectors. Alternatively, it is generally accepted that a hierarchical relationship can exist between precursor mass and fragments as well as among product ions due to the fragmentation cascade[50]. Such relationships are revealed through fragmentation trees (FTs) built from MS/MS data[27]. Initially designed for precursor (and fragment) formula determination, FTs can be computed for most high-resolution MS/MS spectra independently from spectral or structural databases. Compared to other data preparation procedures, FT calculation not only annotates precursor formulas that can be used as important input features for DL, but it also creates edges between mass signals (nodes) if the elemental formula of a fragment is a sub-formula of another (Figure 3B). Next to fragmentation trees, commercial software tools, such as Mass Frontier, create MS^n trees to capture the sequential stages and relationships of mass spectral acquisition in MS^n experiments[8,51]. MS^n trees in Mass Frontier also represent precursor and product ions as nodes and mass differences as edges. The difference is that FTs are built based on implicit formula/sub-formula relationships of precursor/product ions and product/product ions within the MS/MS stage, while MS^n trees highlight the hierarchical order from precursor to all product ions using their explicit links observed between different MS^n stages. Therefore, MS^n trees provide means to directly reconstruct the entire fragmentation cascade without elemental formula calculation. From both FTs and MS^n trees, the directed tree (graph) nature of small molecule MS/MS and MS^n spectra opens the door to graph feature engineering. Graph features, including node/edge attributes (e.g. elemental formula of fragments and mass differences), local structure features (e.g. subtrees), and node/edge embeddings, can be easily fed into a MLP[52–55]. Since both FTs and MS^n trees can capture known fragmentation events, mechanisms, and substructures[50,56], we strongly recommend expert-driven feature engineering on these trees. Figure 3B-③ provide examples of information-rich graph features extracted from a FT.

Besides encoding via pattern or graph feature mining, another possibility is indirect encoding through spectral similarity. The reasoning is that MS/MS spectral similarity can reflect the structural similarity of small molecules[7]. A category of algorithms and software are emerging in metabolomics with the goal to optimize spectral similarity measures to better explain the

This article is protected by copyright. All rights reserved.

structural similarity[57]. For instance, CSI:FingerID[17] computes spectral similarity as FT similarity using kernel methods. The kernels range from node (fragment) intensities and chemical elements of edges (mass difference) to graph-related ones such as common paths and subtrees counting. Subsequently, CSI:FingerID applies multiple kernel learning[58] to combine these kernels in order to build binary classifiers for fingerprint prediction. CSI:FingerID captures both individual peak information and peak interaction through FTs. However, it has two limitations: 1) SVMs produce sparse solutions in dual space instead of primal space, leading to the lack of interpretability in kernel-based methods. 2) Query spectra must be converted into FTs before similarity calculation, and the conversion can be computationally demanding.

A few alternative methods have been proposed in the field to deal with the issues of sparsity and to improve the spectral similarity calculation. Sparse Interaction Model over Peaks of moLEcules (SIMPLE)[59] explicitly incorporate peak interactions after *m/z* discretization. Sparsity is introduced through a sparsity-induced regularizer in the training stage[60], and FT as background knowledge to regularize peak interactions, making the model more robust. SIMPLE resolves two limitations of CSI:FingerID: 1) the contribution of individual peaks and their interactions to spectral similarity is made explicit, 2) the computationally-demanding FT conversion is only needed for regularization but not for prediction of unknown spectra. Spec2Vec[61] is an unsupervised ML technique that creates an embedded space to optimize spectral comparison. It adapts the Word2Vec algorithm[62] to the ML task by transforming MS/MS spectra into "sentences" composed of "vocabularies" (i.e. fragments and neutral losses). Using Word2Vec, co-occurring vocabularies are placed close to each other in a new embedded space during model training. For model testing, spectral similarities are measured based on distances between sentences. A strong correlation is observed between Spec2Vec spectral similarity and structural similarity. DeepMASS[63] is the first DL-based framework for spectral similarity calculation. Starting from binned spectrum pairs and their cross-correlation coefficients, it trains a 4-layer DNN to predict spectral similarity. The output of DeepMASS is an optimized similarity measure (known as MASS scores) that shows a high correlation with the structural similarity.

From CSI:FingerID, SIMPLE, Spec2Vec and DeepMASS, we can see that spectral similarity calculation offers a holistic and flexible approach to encode the entire spectrum, creating information-dense features. In practice, when encoding a training spectrum, features extracted could be similarity scores by matching the training spectrum to a list of reference spectra. Alternatively, we can project the training spectrum in the Spec2Vec embedded space, and the scores obtained can be used as input features of DL frameworks.

In summary, we presented several potential feature engineering strategies for encoding high-dimensional small molecule MS/MS spectra. In DL applications, features extracted from different algorithms can capture complementary pieces of information in the raw data. Therefore, they can be combined to achieve improved prediction performance. We propose here DL architectures 3 and 4 by combining blocks of features extracted from three selected spectral encoding methods (Figure 3A).

To begin, a pool of reference compounds can be used to encode input spectra based on spectral similarity (Figure 3B-①). To select reference compounds, we suggest metabolic network hubs to encode metabolite spectra since these compounds are involved in diverse biochemical reactions, which can help position structurally related input spectra. Likewise, unmodified drugs can be used to encode spectra of pharmaceutical process impurities and degradation products.

This article is protected by copyright. All rights reserved.

Second, input spectra can be matched against expert-curated and substructure-related patterns available in MS2LDA and MESSAR (Figure 3B-②). Features generated will be score vectors that represent the presence/absence of spectral patterns in the input spectra or degree of overlapping.

For the last encoding method, we suggest a knowledge-based approach to extract graph features. We will first present to experts a collection of FTs computed for MS/MS spectra of reference compounds. Using the knowledges about fragmentation mechanisms and substructures, alongside with elemental formulas assigned during FT computations, they can identify and record graph features that are useful and important for structure elucidation (Figure 3B-③). Typical features to be identified include nodes (representing fragments), edges (neutral losses and mass differences), and subgraphs (e.g. long paths indicating consecutive losses of functional groups). New algorithms are required to automatically determine the values of these graph features for the entire training spectra after converting them to FTs.

Since the three feature engineering approaches proposed are complementary, a single DNN combining three classes of input features is expected to improve the model performance[64,65]. To fusion heterogeneous multimodal features, we recommend the strategy adopted in DeepMASS[63] of first connecting each feature class to fully-connected layers followed by a activation layer. After that, the outputs can be concatenated to form information-rich representations of the input spectra. In Architecture 3, we link this shared representation to task-specific layers (individual classifiers) for molecular descriptor prediction (Architecture 3). The new architecture is an improved version of Architecture 1 in the category of two-step ML approaches (Figure 2).

Architecture 4 was inspired by the GLEAMS framework[66]. Its goal is to create an output space that directly translates spectral similarity to small molecule structural similarity. Training can happen in a Siamese DNN setup: instead of processing one spectrum at a time, the neural network *NN* is trained on pairs of spectra, such as S_1 and S_2. The distance d_s between encodings, noted as $NN(S_1)$ and $NN(S_2)$, will be computed. If the structural distance between underlying molecules m_1 and m_2 is d_m, the loss function to minimize when training the DNN would be $(d_m\text{-}d_s)^2$. In the output space, structurally similar training molecules will be positioned together while unrelated molecules will be pushed away from each other.

In the prediction phase of Architecture 3 and 4, blocks of features will be extracted from a query spectrum in a similar manner as for training spectra and further encoded by the DNN. For Architecture 3, structure elucidation can be achieved by comparing predicted molecular descriptors to candidates retrieved from a molecular database. Architecture 4 produces distances by feeding the Siamese network the query spectrum and reference spectra from a selected spectral library. The putative annotation will be the label (structure) of the candidate spectrum which has the smallest distance to the query spectrum.

As opposed to "two-step" approaches (i.e. Architectures 1, 2, and 3), Architecture 4 does not require intermediate tasks of molecular descriptor prediction since structure elucidation is achieved via direct library spectra matching in the Siamese output space. However, Architecture 4 requires library spectra in its prediction phase if exact structure annotation is desired. For never-seen compounds, it is expected that structurally similar molecules are retrieved.

This article is protected by copyright. All rights reserved.

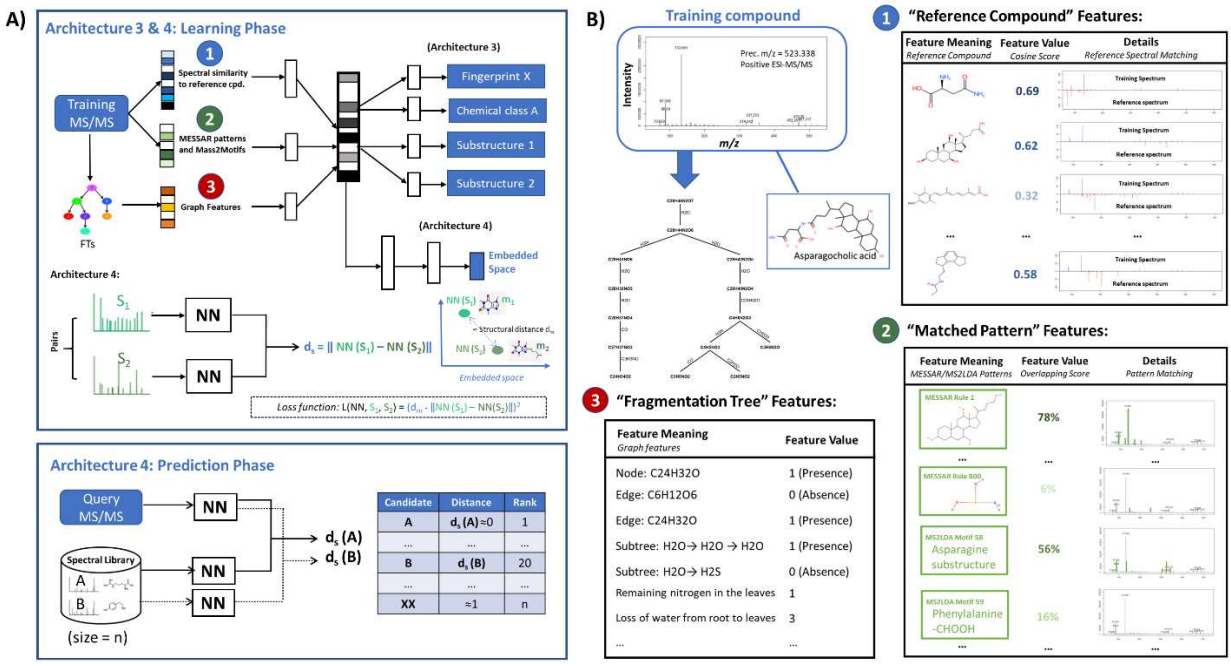

Figure 3. Deep learning architectures based on spectral feature engineering. A) Three blocks of feature values are calculated by mapping input MS/MS spectra to: ① reference spectra of important metabolites ② structurally annotated spectral patterns from MESSAR and MS2LDA ③ graph features designed by experts based on FTs of selected training spectra. Each block of input features is followed by a fully connected layer with activation function before concatenating into a shared layer. In Architecture 3, the shared layer will be connected to task-specific layers for individual molecular descriptor prediction. Architecture 4 follows a Siamese network architecture NN that learns an output encoding that minimizes the difference between spectral distance *ds* and small molecule structural distance d_m. The prediction phase of Architecture 4 can improve spectral library searching by processing pairs of query-library spectra through *NN* and computing the spectral distance in the informative encoding space. Structure candidates can be ranked according to their distances to the query spectrum in the output space. B) Example of asparagocholic acid (positive ion mode MS/MS spectrum with fragmentation tree built by CSI:FingerID) to break down the three feature engineering methods used in Architecture 3 and 4: ① "Reference compound" features are a vector of spectral similarity values (here cosine score) between input spectrum and selected reference spectra (of some important metabolites in this example). ② "Matched pattern" features are a vector of overlapping scores between input spectrum and spectral patterns from MESSAR rules and MSL2DA motifs. Overlapping score is defined as the percentage of fragments and neutral losses explained by each pattern. ③ "Fragmentation tree" features are expert-designed graph features that capture important structure information and fragmentation events from FTs. Elemental formula assigned to nodes (product ions) can also be examples of graph features. Feature values were determined manually in the example.

6. Encoding small molecule structures to improve DL frameworks

Most ML or DL frameworks for automated structure elucidation have used predefined molecular descriptors to encode training molecules to a binary vector. However, MS/MS data, due to its complexity and potential lack of relevant fragmentation patterns, cannot always be related to known molecular properties, substructures, or chemical classes. Interestingly, an active research topic in QSAR is to improve ML or DL model performance by finding optimal molecular representations or structural feature engineering of small molecules[67].

One important attempt is feature engineering from small molecule structures using deep neural networks. Spurred by the increasing number of large molecular databases that are available for pharmaceutical research, several DL methods have recently been proposed to obtain features from 2D molecular information[68].

Drugs, as all small molecules, are composed of atoms and chemical bonds. By treating small molecules as graphs or networks in which atoms are nodes and chemical bonds are edges, task-

This article is protected by copyright. All rights reserved.

relevant features can be extracted using convolution operators adapted for molecular graphs. Message-passing neural networks (MPNNs) generalize several graph-based neural networks, including graph convolutional networks[69,70]. These networks use atom and bond information as input and are capable of learning flexible fingerprints most relevant for the task at hand. Besides graph convolutions, recent developments in QSAR have used types of NNs to generate meaningful chemical representations. Using SMILES as inputs, recurrent neural networks (RNNs), possibly in combination with attention mechanisms[71], and convolutional neural networks (CNNs) which accept one-dimensional input, can directly predict molecular properties or biological activities[72–75].

The above DL methods can be adapted to encode small molecules to improve automated structure elucidation. ADAPTIVE[76] is the first framework in the field, to our knowledge, that adopts such an approach. ADAPTIVE is an improved version of the IOKR algorithm[57], developed for fingerprint-free metabolite identification. IOKR first computes spectral similarity and fingerprint-based structural similarity matrices from the training data. The learning phase of IOKR minimizes the distance between the two similarity matrices. MFPs are only used for structural similarity calculation and are not required in the prediction phase for compound identification. ADAPTIVE further eliminates MFPs in the learning phase by using a "molecular vector" – i.e. descriptors learned from training structures through the graph convolution network MPNN. The MPNN is trained so that the correlation between the spectra and the molecular vector is maximized. Because molecular vectors are compact and implicit representations of small molecules adapted specifically to MS/MS data, ADAPTIVE led to more accurate metabolite identification than IOKR, which was trained using classical molecular descriptors.

Architecture 5, ADAPTIVE, was initially designed for processing multi-modal data (Figure 4). The reasoning is that input MS/MS spectra and small molecule structures are two distinct types of training data that should be encoded with separate DNNs. However, if the two types of data are paired and derived from the same spectral library, it is possible to build a unified learning framework via joint or cross-modal embedding. The common representation (joint embedding space) generated may offer similar benefits as the molecular vectors in ADAPTIVE.

In DL research, cross-modal embedding[77–79] has been widely used to map different modes of media, such as text, image, music, and video, into the joint embedding space in which media can be compared solely based on their captions. Inspired from embedding frameworks of multimedia data, the learning phase of Architecture 5 requires a triplet loss function L[79]. By minimizing the triplet loss, spectral and structural encoding can be learned so that an input spectrum (e.g. S_1) will be closer to its corresponding structure (m_1) in the joint embedding space than any other structure (e.g. the compound m_2). In the prediction phase, a query spectrum can be projected onto the joint embedding space. For direct candidate match, we can evaluate the distances between the query spectrum and pre-calculated coordinates of database structures or library spectra. In contrast to Architecture 4, Architecture 5 is not limited to training compounds which have a spectrum, and the candidate retrieval can be extended to a combined search space of molecular databases and spectral libraries.

This article is protected by copyright. All rights reserved.

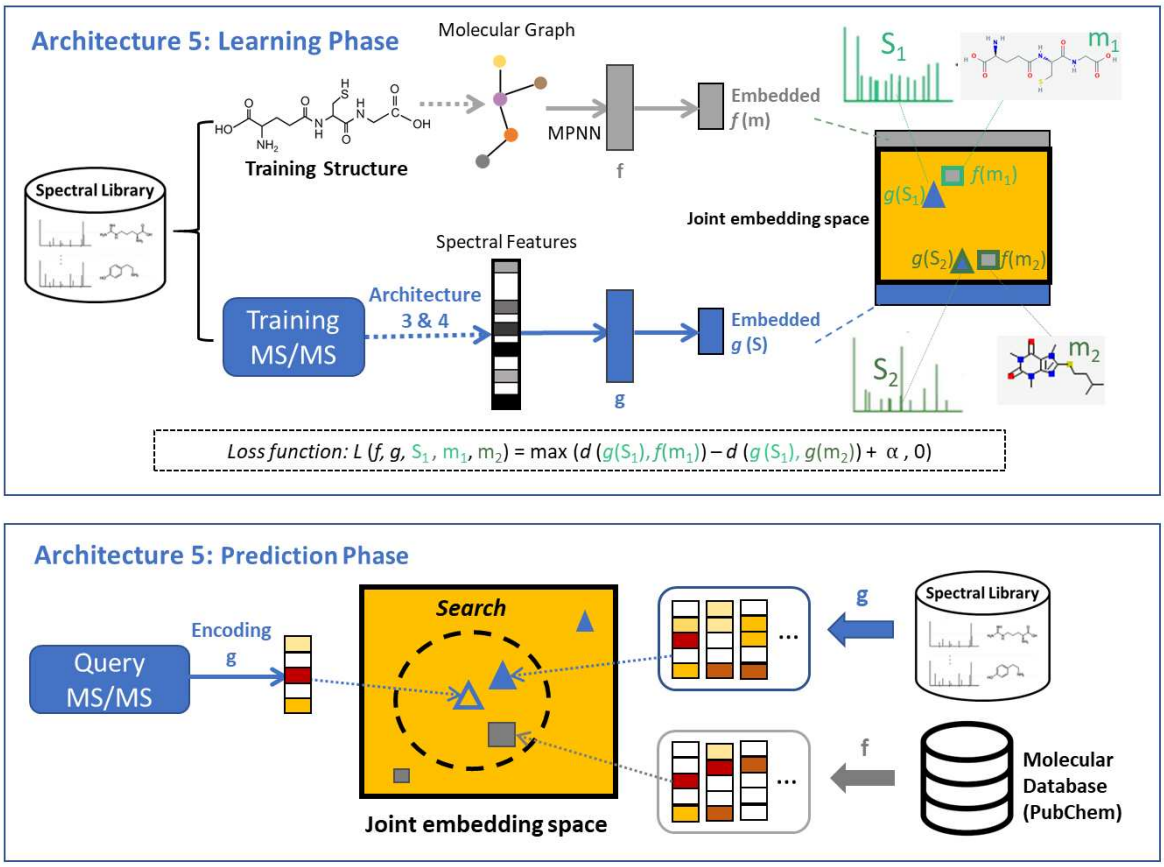

Figure 4. Deep learning architecture with structural encoding and cross-modal embedding of spectra and structure. Training structures can be converted to molecular graphs before applying MPNN to learn meaningful representations. Meanwhile, training spectra can be encoded using approaches described in Figure 3. The mapping functions for structures and spectra are noted as *f* and g, respectively. The learning phase maximizes the co-structure of two encodings so that the spectrum (S_1) should be closer to the corresponding structure (m_1) in the joint embedding space than any other structures (e.g. compound m_2). A triplet loss function *L* is adapted for this learning task. In the function *L*, *d* expresses the distance between the embeddings in the latent space, and α is a positive margin which controls the separation. The sum of losses over all triplets (i.e. an input spectrum, the corresponding structure, and an irrelevant structure) should be minimized to learn the parameters of *f* and *g*. In the prediction phase, a query spectrum can be mapped to the joint embedding space with *g*. Meanwhile, the database structures and library spectra can be encoded via *f* and *g* in the same joint space, which allows simultaneous structure candidate retrieval from molecular databases and spectral libraries.

Another important DL development in QSAR and drug discovery is structural feature engineering via unsupervised learning. The resulting molecular representations are not limited to the training task at hand but can instead be generated based on a large set of unrelated input structures. Some extracted features can be further be applied as inputs in a modelling task of interest. Popular architectures of unsupervised learning include Generative Adversarial Networks (GANs)[80] and Variational Autoencoders (VAEs)[81] based on RNNs[82], transformers[83] and graph models[84].

A VAE consists of an encoder and a decoder. The encoder compresses discrete molecular representations (e.g. SMILES strings) into latent continuous vectors (the latent space), and the decoder can recreate the input molecules from these vectors. The success of VAEs in drug discovery is mainly due to the ability for *de novo* generation of small molecule drugs. In fact, any given point in the latent space can be passed to the decoder networks in order to generate a chemical structure [85]. It can be noted that this generation process is non-deterministic.

This article is protected by copyright. All rights reserved.

Therefore, by performing operations such as random sampling, perturbing known molecules, or interpolating between molecules, the decoder can generate novel chemical structures. Some VAE architectures include a predictor that estimates chemical properties or biological activities from the continuous vectors[82,86,87]. These frameworks create or select latent representations that are expected to lead to higher values of desired properties or activities. The new representations can then be "decoded" to new drug candidates.

The *de novo* aspect of VAEs is interesting for MS-based structure elucidation. We recall that databases, either spectral libraries or molecular databases, are necessary to retrieve structural candidates in all DL frameworks previously presented. However, many molecules to be identified, such as process impurities or degradation products from drugs in early pharmaceutical development, are present in neither spectral libraries nor structure databases. The decoder networks of VAEs, on the other hand, although requiring an extensive database when training in an unsupervised manner, do not require this for the task at hand, and offer the possibility of *de novo* structure elucidation.

For instance, a pre-trained decoder network called *Neuraldecipher* has been recently introduced in the QSAR field[88]. *Neuraldecipher* aims to reconstruct or partially recreate molecular structures via reverse-engineering of discrete MFPs by learning a mapping from ECFPs to a latent space from which the SMILES representations of small molecules can be obtained through the decoder. To allow database-independent structure elucidation, instead of mapping from ECFPs, MS/MS spectral information can be used to map to such latent space. An example using input spectra and structures, could be mapping the combined feature layer in Architecture 3 (Figure 3A) to a continuous latent space *C* (instead of predefined molecular descriptors) and then mapping (with the decoder network pretrained in *Neuraldecipher*) from *C* to SMILES representations. In the prediction phase of such architectures, the trained decoder network could sample from *C* to predict *de novo* structure candidates without the need of spectral libraries or molecular databases. Furthermore, recent autoencoder architectures can incorporate user-defined constraints for sampling and prediction[89]. For instance, if we know an unknown compound is a degradation product of a known drug, instead of sampling from the entire latent space, we can predict the structure of the unknown by sampling from the latent space surrounding that drug. The output of the decoder will be structurally close to the known compound in the context of application.

As a *de novo* approach, auto-encoder frameworks can normally generate a high proportion of grammatically valid SMILES[87]. Moreover, there exist other types of structural representations, such as SELFIE[90], that can further prevent invalid structure outputs. On the other hand, the appropriateness of the proposed structures should be rigorously evaluated based on the context of application, domain knowledge, and ideally data from other analytical layers or dimensions.

7. Conclusion and General Discussions

The structure elucidation of small molecules is an essential step towards biological interpretation, biomarker discovery, process understanding, etc. in diverse application domains. Despite the increasing availability of library spectra and powerful software tools, the identification of unknowns still requires laborious manual interpretation and expert intervention. In fact, many spectra cannot be solved by current bioinformatics tools due to the lack of library spectra or appropriate algorithmic models.

Recently, deep learning has achieved remarkable success in various research areas. In the MS field, the increased availability of potential training data (public MS/MS spectra and community-wide curated structures) has stimulated early deep learning efforts for automated structure elucidation. These early-stage DL frameworks follow a two-step concept characterized by *m/z* binning, DNN prediction for individual molecular descriptors, and

This article is protected by copyright. All rights reserved.

descriptor-driven molecular database search. To potentially improve these models, we conceptualize and design new DL architectures that: 1) combine structurally related classification tasks (descriptor prediction) via MTL; 2) engineer information-rich features from training spectra and structures; 3) create joint embedding of spectra and structures. Like existing ML models, these architectures are adapted for structural candidate retrieval from molecular databases and spectral libraries.

The reason DL architectures might outperform classical ML models is their ability to learn complex interactions between input features. In contrast, DL frameworks are considered to have some limitations and flaws compared to classical ML models: A) Combining massive training data to learn from and parameters to compute, DL models can take quite a lot of time to train, sometimes up to several weeks, while training smaller ML models require much less computing power. B) DL models usually have many hyperparameters to fine-tune, and the process of optimizing hyperparameters can be time-consuming and tedious especially when done manually or when an exhaustive grid search approach is used. C) The "black box" nature of DNNs is a barrier in structure elucidation-based applications, where interpretability is essential. As an example, DNNs may achieve high accuracy in predicting a specific substructure, but they cannot directly pinpoint to end-users the subset of contributing features in the raw spectra.

On the other hand, as a rapidly emerging field, there are many techniques and toolboxes in DL to tackle above challenges. To tackle A), DL training time can be reduced by using specialized hardware, such as GPUs[20] or custom cloud-based processors[91] (e.g. Google's Tensor Processing Unit). Moreover, advanced training approaches, such as self-supervised learning[92], can be used to achieve state-of-the-art performance without the need for a huge amount of labeled data. Concerning B), advanced hyperparameter tuning techniques, such as Bayesian optimization[93,94] or early stopping[95], can more efficiently find optimal hyperparameters and significantly reduce the computational requirements. An extensive guideline for improving hyperparameter tuning and model selection was recently published[96]. For the interpretability issue in C), powerful techniques to elucidate the workings of DL models are becoming available. For example, SHAP (Shapley Additive exPlanations)[97] is an interpretability technique rooted in game theory that uses interpretable local models to assign a SHAP importance value to each input feature for a particular prediction. In this fashion, SHAP values can be used to identify important spectral features contributing to the predictions and interpret structure elucidation-based models.

Despite existing techniques to improve DL models, it is still unclear whether their performance and general applicability can outcompete classical ML frameworks. In terms of prediction accuracy, it is important to benchmark and evaluate several classical ML and DL frameworks on the same, previously unknown, validation set. As demonstrated by the ImageNet challenge[98], which initiated the DL revolution in image processing, community challenges can be a valuable venue to find the best model. Interestingly, community challenges are already launched for small molecule structure elucidation. A typical example is the CASMI contest[99], or the Critical Assessment of Small Molecule Identification, that enables comparison of diverse identification software and methods on a common dataset.

Meanwhile, common drawbacks of ML (including DL) raised in Nguyen et al.[15] have been partially addressed throughout the manuscript. Several related techniques, including MTL, feature engineering, and joint embedding, are not limited to DL architectures - they can be adopted by classical ML models to deal with dimensionality, higher-order interaction, and data sparsity issues.

This article is protected by copyright. All rights reserved.

After all, to fundamentally enhance automated MS-based structure elucidation, we believe that the following directions should not be overlooked next to developing new DL and even classical ML frameworks: 1) Increasing the amount, the diversity and the quality of training data, which includes the removal of low quality spectra, the detection of incorrect structure labels, and integrating data from orthogonal analytical dimensions (e.g. precursor *m/z*, charge, retention time, isotopic pattern, collision cross section values) into current frameworks. 2) Further exploring database-independent structure elucidation. Interestingly, the last direction will lead to potential discovery of novel small molecules from MS/MS data. This *de novo* aspect was recently tackled by CANOPUS[100], a DNN model that predicts 2,497 compound classes from any given MS/MS spectra. To move beyond compound classes, we introduce VAE due to its unique functionality to reconstruct a novel structure entirely. Although the appropriateness of reconstructed structures should be rigorously evaluated, mostly correct structural proposals (e.g. partial prediction of very large substructures) by VAEs are already informative and can be a big leap towards complete automation and compound discovery in diverse MS applications. 3) Bridging the developer-end-user gap. As a major challenge to both classical ML and DL-based methods, future directions should be developing user-friendly software based on a comprehensive structure elucidation pipeline that includes raw data processing, customizable reporting, model interpretation and batch processing. In the end, the proof of the pudding for the end-users will be in the eating.

References

1. De Vijlder T, Valkenborg D, Lemière F, Romijn EP, Laukens K, Cuyckens F. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrom Rev*. 2018;37(5):607-629. doi:10.1002/mas.21551

2. Suhre K, Schmitt-Kopplin P. MassTRIX: mass translator into pathways. *Nucl Acids Res*. 2008;36(suppl 2):W481-W484. doi:10.1093/nar/gkn194

3. Liang Q, Liu H, Xie L, Li X, Zhang A-H. High-throughput metabolomics enables biomarker discovery in prostate cancer. *RSC Adv*. 2017;7(5):2587-2593. doi:10.1039/C6RA25007F

4. Liu Y, Romijn EP, Verniest G, Laukens K, De Vijlder T. Mass spectrometry-based structure elucidation of small molecule impurities and degradation products in pharmaceutical development. *TrAC Trends in Analytical Chemistry*. 2019;121:115686. doi:10.1016/j.trac.2019.115686

5. Alvarez-Rivera G, Ballesteros-Vivas D, Parada-Alfonso F, Ibañez E, Cifuentes A. Recent applications of high resolution mass spectrometry for the characterization of plant natural products. *TrAC Trends in Analytical Chemistry*. 2019;112:87-101. doi:10.1016/j.trac.2019.01.002

6. Kind T, Tsugawa H, Cajka T, et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews*. 2018;37(4):513-532. doi:10.1002/mas.21535

7. Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with GNPS. *Nat Biotechnol*. 2016;34(8):828-837. doi:10.1038/nbt.3597

This article is protected by copyright. All rights reserved.

8. Vaniya A, Fiehn O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Analyt Chem*. 2015;69:52-61. doi:10.1016/j.trac.2015.04.002

9. Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*. 2010;45(7):703-714. doi:10.1002/jms.1777

10. Xue J, Guijas C, Benton HP, Warth B, Siuzdak G. METLIN MS 2 molecular standards database: a broad chemical and biological resource. *Nat Methods*. 2020;17(10):953-954. doi:10.1038/s41592-020-0942-5

11. Haug K, Cochrane K, Nainala VC, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res*. 2020;48(D1):D440-D444. doi:10.1093/nar/gkz1019

12. Hufsky F, Böcker S. Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. *Mass Spectrometry Reviews*. 2017;36(5):624-633. doi:10.1002/mas.21489

13. Peisl BYL, Schymanski EL, Wilmes P. Dark matter in host-microbiome metabolomics: Tackling the unknowns–A review. *Analytica Chimica Acta*. 2017;1037(2018):13-27. doi:10.1016/j.aca.2017.12.034

14. Silva RR da, Wang M, Nothias L-F, et al. Propagating annotations of molecular networks using in silico fragmentation. *PLOS Computational Biology*. 2018;14(4):e1006089. doi:10.1371/journal.pcbi.1006089

15. Nguyen DH, Nguyen CH, Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches (2019). *Brief Bioinform*. 2019;20(6):2028-2043. doi:10.1093/bib/bby066

16. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*. 2019;47(D1):D1102-D1109. doi:10.1093/nar/gky1033

17. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences of the United States of America*. 2015;112(41):12580–12585. doi: 10.1073/pnas.1509788112

18. Dahl GE, Jaitly N, Salakhutdinov R. Multi-task Neural Networks for QSAR Predictions. *arXiv.* 2014; arXiv: 1406.1231v1. http://arxiv.org/abs/1406.1231-live Accessed April 12, 2021.

19. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. *Front Environ Sci*. 2016;3. doi:10.3389/fenvs.2015.00080

20. Gawehn E, Hiss JA, Brown JB, Schneider G. Advancing drug discovery via GPU-based deep learning. *Expert Opin Drug Discov*. 2018;13(7):579-582. doi:10.1080/17460441.2018.1465407

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

22. Djoumbou-Feunang Y, Pon A, Karu N, et al. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction and Compound Identification. *Metabolites*. 2019;9(4). doi:10.3390/metabo9040072

This article is protected by copyright. All rights reserved.

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

24. Rogers S, Wei Ong C, Wandy J, Ernst M, Ridder L, Hooft JJJ van der. Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. *Faraday Discussions*. 2019;218(0):284-302. doi:10.1039/C8FD00235E

25. Yoshida H, Leardi R, Funatsu K, Varmuza K. Feature selection by genetic algorithms for mass spectral classifiers. *Analytica Chimica Acta*. 2001;446(1):483-492. doi:10.1016/S0003-2670(01)00910-2

26. Hummel J, Strehmel N, Selbig J, Walther D, Kopka J. Decision tree supported substructure prediction of metabolites from GC-MS profiles. *Metabolomics*. 2010;6(2):322-333. doi:10.1007/s11306-010-0198-7

27. Böcker S, Dührkop K. Fragmentation trees reloaded. *J Cheminform*. 2016;8. doi:10.1186/s13321-016-0116-8

28. Hooft JJJ van der, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *PNAS*. 2016;113(48):13738-13743. doi:10.1073/pnas.1608041113

29. Djoumbou Feunang Y, Eisner R, Knox C, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform*. 2016;8:61. doi:10.1186/s13321-016-0174-y

30. Shahlaei M. Descriptor Selection Methods in Quantitative Structure–Activity Relationship Studies: A Review Study. *Chem Rev*. 2013;113(10):8093-8103. doi:10.1021/cr3004339

31. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. John Wiley & Sons; 2008.

32. Halder AK, Moura AS, Cordeiro MNDS. QSAR modelling: a therapeutic patent review 2010-present. *Expert Opinion on Therapeutic Patents*. 2018;28(6):467-476. doi:10.1080/13543776.2018.1475560

33. Ewing T, Baber JC, Feher M. Novel 2D Fingerprints for Ligand-Based Virtual Screening. *J Chem Inf Model*. 2006;46(6):2423-2431. doi:10.1021/ci060155b

34. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *J Chem Inf Model*. 2010;50(5):742-754. doi:10.1021/ci100050t

35. McGregor MJ, Muskal SM. Pharmacophore fingerprinting. 2. Application to primary library design. *J Chem Inf Comput Sci*. 2000;40(1):117-125. doi:10.1021/ci990313h

36. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. *Methods*. 2015;71:58-63. doi:10.1016/j.ymeth.2014.08.005

37. Gao K, Nguyen DD, Sresht V, Mathiowetz AM, Tu M, Wei G-W. Are 2D fingerprints still valuable for drug discovery? *Phys Chem Chem Phys*. 2020;22(16):8373-8390. doi:10.1039/D0CP00305K

Accepted Article

This article is protected by copyright. All rights reserved.

38. Unterthiner T, Mayr A, Klambauer G, Hochreiter S. Toxicity Prediction using Deep Learning. *arXiv*. 2015; arXiv: 1503.01445v1. http://arxiv.org/abs/1503.01445-live Accessed April 12, 2021.

39. Ramsundar B, Liu B, Wu Z, et al. Is Multitask Deep Learning Practical for Pharma? *J Chem Inf Model*. 2017;57(8):2068-2076. doi:10.1021/acs.jcim.7b00146

40. Sturm N, Mayr A, Le Van T, et al. Industry-scale application and evaluation of deep learning for drug target prediction. *Journal of Cheminformatics*. 2020;12(1):26. doi:10.1186/s13321-020-00428-5

41. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discovery Today*. 2018;23(6):1241-1250. doi:10.1016/j.drudis.2018.01.039

42. Ji H, Deng H, Lu H, Zhang Z. Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. *Anal Chem*. 2020;92(13):8649-8653. doi:10.1021/acs.analchem.0c01450

43. Fan Z, Alley A, Ghaffari K, Ressom HW. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics*. 2020;16(10):104. doi:10.1007/s11306-020-01726-7

44. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J Chem Inf Model*. 2017;57(10):2490-2504. doi:10.1021/acs.jcim.7b00087

45. Caruana R. Multitask Learning. *Machine Learning*. 1997;28(1):41-75. doi:10.1023/A:1007379606734

46. Ben-David S, Schuller R. Exploiting Task Relatedness for Multiple Task Learning. In: Schölkopf B, Warmuth MK, eds. *Learning Theory and Kernel Machines*. Lecture Notes in Computer Science. Springer; 2003:567-580. doi:10.1007/978-3-540-45167-9_41

47. Zhang P, Chan W, Ang IL, et al. Revisiting Fragmentation Reactions of Protonated α-Amino Acids by High-Resolution Electrospray Ionization Tandem Mass Spectrometry with Collision-Induced Dissociation. *Scientific Reports*. 2019;9(1):6453. doi:10.1038/s41598-019-42777-8

48. Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*. 2020;12(1):43. doi:10.1186/s13321-020-00445-4

49. O'Boyle NM, Sayle RA. Comparing structural fingerprints using a literature-based similarity benchmark. *J Cheminform*. 2016;8. doi:10.1186/s13321-016-0148-0

50. Kasper PT, Rojas-Chertó M, Mistrik R, Reijmers T, Hankemeier T, Vreeken RJ. Fragmentation trees for the structural characterisation of metabolites. *Rapid Communications in Mass Spectrometry*. 2012;26(19):2275-2286. doi:10.1002/rcm.6340

51. Tolstikov VV, Fiehn O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal Biochem*. 2002;301(2):298-307. doi:10.1006/abio.2001.5513

Accepted Article

This article is protected by copyright. All rights reserved.

52. Deepak A, Fernández-Baca D, Tirthapura S, Sanderson MJ, McMahon MM. EvoMiner: frequent subtree mining in phylogenetic databases. *Knowledge and Information Systems*. 2014;41(3):559–590. doi: 10.1007/s10115-013-0676-0

53. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. *KDD*. 2016;2016:855-864. doi:10.1145/2939672.2939754

54. Rossi RA, Zhou R, Ahmed NK. Deep Feature Learning for Graphs. *arXiv:170408829 [cs, stat]*. Published online October 16, 2017. Accessed January 2, 2021. http://arxiv.org/abs/1704.08829

55. Lee JB, Rossi R, Kong X. Graph Classification using Structural Attention. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. Association for Computing Machinery; 2018:1666–1674. doi:10.1145/3219819.3219980

56. Ludwig M, Nothias L-F, Dührkop K, et al. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence*. 2020;2(10):629-641. doi:10.1038/s42256-020-00234-6

57. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J. Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics*. 2016;32(12):i28-i36. doi:10.1093/bioinformatics/btw246

58. Brouard C, Bassé A, d'Alché-Buc F, Rousu J. Improved Small Molecule Identification through Learning Combinations of Kernel Regression Models. *Metabolites*. 2019;9(8):160. doi:10.3390/metabo9080160

59. Nguyen DH, Nguyen CH, Mamitsuka H. SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics*. 2018;34(13):i323-i332. doi:10.1093/bioinformatics/bty252

60. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-288. https://www.jstor.org/stable/2346178?seq=1-live Accessed April 12, 2021

61. Huber F, Ridder L, Rogers S, Hooft JJJ van der. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol.* 2021;17(2):e1008724. doi: 10.1371/journal.pcbi.1008724

62. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. *arXiv.* 2013; 13104546v1. http://arxiv.org/abs/1310.4546-live Access April 12, 2021

63. Ji H, Xu Y, Lu H, Zhang Z. Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification. *Anal Chem*. 2019;91(9):5629-5637. doi:10.1021/acs.analchem.8b05405

64. Paul A, Jha D, Al-Bahrani R, Liao W-K, Choudhary A, Agrawal A. *CheMixNet: Mixed DNN Architectures for Predicting Chemical Properties Using Multiple Molecular Representations*. *arXiv*, 2018; arXiv:1811.08283v2. https://arxiv.org/abs/1811.08283-live Access April 12, 2021

65. Laufkötter O, Sturm N, Bajorath J, Chen H, Engkvist O. Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. *Journal of Cheminformatics*. 2019;11(1):54. doi:10.1186/s13321-019-0376-1

This article is protected by copyright. All rights reserved.

66. Bittremieux W, May DH, Bilmes J, Noble WS. Deep neural network embedding for efficient repository-scale analysis of hundreds of millions of mass spectra. Presented at the: ASMS 2020 Reboot; June 8, 2020. doi:10.5281/zenodo.3831052

67. Khan PM, Roy K. Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opinion on Drug Discovery*. 2018;13(12):1075-1089. doi:10.1080/17460441.2018.1542428

68. Alsenan SA, Al-Turaiki IM, Hafez AM. Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study. *IEEE Access*. 2020;8:78737-78752. doi:10.1109/ACCESS.2020.2990375

69. Duvenaud DK, Maclaurin D, Iparraguirre J, et al. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems*. 2015;28:2224-2232.

70. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. In: *International Conference on Machine Learning*. PMLR; 2017:1263-1272. http://proceedings.mlr.press/v70/gilmer17a.html-live Access April 12, 2021

71. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *arXiv.* 2017; arXiv: 1706.03762v5. http://arxiv.org/abs/1706.03762-live Access April 12, 2021

72. Goh G, Hodas N, Siegel C, Vishnu A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. *arXiv.* 2017; arXiv: 1712.02034v2. https://arxiv.org/abs/1712.02034v2- live Access April 12, 2021

73. Mayr A, Klambauer G, Unterthiner T, et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci*. 2018;9(24):5441-5451. doi:10.1039/C8SC00148K

74. Jastrzębski S, Leśniak D, Czarnecki WM. Learning to SMILE(S). *arXiv.* 2018; arXiv: 160.206289. http://arxiv.org/abs/1602.06289- live Access April 12, 2021

75. Zheng S, Yan X, Yang Y, Xu J. Identifying Structure–Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J Chem Inf Model*. 2019;59(2):914-923. doi:10.1021/acs.jcim.8b00803

76. Nguyen DH, Nguyen CH, Mamitsuka H. ADAPTIVE: leArning DAta-dePendenT, concIse molecular VEctors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*. 2019;35(14):i164-i172. doi:10.1093/bioinformatics/btz319

77. Carvalho M, Cadène R, Picard D, Soulier L, Thome N, Cord M. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. *arXiv*, 2018; arXiv: 1804.11146. http://arxiv.org/abs/1804.11146- live Access April 12, 2021

78. Wang G, Li C, Wang W, et al. Joint Embedding of Words and Labels for Text Classification. *arXiv*, 2018; arXiv:1805.04174. http://arxiv.org/abs/1805.04174- live Access April 12, 2021

79. Wang H, Sahoo D, Liu C, Lim E, Hoi SCH. Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images. *arXiv*, 2019; arXiv: 1905.01273. http://arxiv.org/abs/1905.01273- live Access April 12, 2021

This article is protected by copyright. All rights reserved.

80. Vanhaelen Q, Lin Y-C, Zhavoronkov A. The Advent of Generative Chemistry. *ACS Med Chem Lett*. 2020;11(8):1496-1505. doi:10.1021/acsmedchemlett.0c00088

81. Kingma DP, Welling M. Auto-Encoding Variational Bayes. *arXiv*, 2014; arXiv: 1312.6114. http://arxiv.org/abs/1312.6114- live Access April 12, 2021

82. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci*. 2018;4(1):120-131. doi:10.1021/acscentsci.7b00512

83. Liu D, Liu G. A Transformer-Based Variational Autoencoder for Sentence Generation. In: *2019 International Joint Conference on Neural Networks (IJCNN)*; 2019:1-7. doi:10.1109/IJCNN.2019.8852155

84. Simonovsky M, Komodakis N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *arXiv*, 2018;1802.03480. http://arxiv.org/abs/1802.03480- live Access April 12, 2021

85. Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci*. 2018;4(2):268-276. doi:10.1021/acscentsci.7b00572

86. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*. 2019;37(9):1038-1040. doi:10.1038/s41587-019-0224-x

87. Winter R, Montanari F, Noé F, Clevert D-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci*. 2019;10(6):1692-1701. doi:10.1039/C8SC04175J

88. Le T, Winter R, Noé F, Clevert D-A. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chem Sci*. 2020;11(38):10378-10389. doi:10.1039/D0SC03115A

89. Engel J, Hoffman M, Roberts A. Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models. *arXiv*, 2017; 1711.05772. http://arxiv.org/abs/1711.05772- live Access April 12, 2021

90. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Mach Learn: Sci Technol*. 2020;1(4):045024. doi:10.1088/2632-2153/aba947

91. Hwang K. *Cloud Computing for Machine Learning and Cognitive Applications*. The MIT Press; 2017.

92. Liu X, Zhang F, Hou Z, et al. Self-supervised Learning: Generative or Contrastive. *arXiv,* 2020; 2006.08218. http://arxiv.org/abs/2006.08218- live Access April 12, 2021

93. Snoek J, Larochelle H, Adams RP. Practical Bayesian Optimization of Machine Learning Algorithms. *arXiv*, 2012; 1206.2944. http://arxiv.org/abs/1206.2944- live Access April 12, 2021

94. Frazier PI. A Tutorial on Bayesian Optimization. *arXiv*, 2018; 1807.02811. http://arxiv.org/abs/1807.02811- live Access April 12, 2021

This article is protected by copyright. All rights reserved.

95. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *arXiv*, 2018; 1603.06560. http://arxiv.org/abs/1603.06560- live Access April 12, 2021

96. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv,* 2020; 1811.12808. http://arxiv.org/abs/1811.12808- live Access April 12, 2021

97. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv,* 2017;1705.07874. http://arxiv.org/abs/1705.07874- live Access April 12, 2021

98. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*; 2009:248-255. doi:10.1109/CVPR.2009.5206848

99. Schymanski EL, Neumann S. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites*. 2013;3(3):517-538. doi:10.3390/metabo3030517

100. Dührkop K, Nothias L-F, Fleischauer M, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*. 2020; 39: 462-471. doi:10.1038/s41587-020-0740-8

Accepted Article

This article is protected by copyright. All rights reserved.