Deep Learning Models for LC-MS Untargeted Metabolomics Data Analysis

Francesco Russo[1], Filip Ottosson[1], Justin J. J. van der Hooft[2], and Madeleine Ernst[1(✉)]

[1] Section for Clinical Mass Spectrometry, Danish Center for Neonatal Screening, Department of Congenital Disorders, Statens Serum Institut, Copenhagen, Denmark
maet@ssi.dk

[2] Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

Abstract. Metabolomics, the measurement of all metabolites in a given system, is a growing research field with great potential and manifold applications in precision medicine. However, the high dimensionality and complexity of metabolomics data requires expert knowledge, the use of proper methodology, and is largely based on manual interpretation. In this book chapter, we discuss recent published approaches using deep learning to analyze untargeted metabolomics data. These approaches were applied within diverse stages of metabolomics data analysis, e.g. to improve preprocessing, feature identification, classification, and other tasks. We focus our attention on deep learning methods applied to liquid chromatography mass spectrometry (LC-MS), but these models can be extended or adjusted to other applications. We highlight current deep learning-based computational workflows that are paving the way toward high(er)-throughput use of untargeted metabolomics, making it effective for clinical, environmental and other types of applications.

Keywords: Metabolomics · Deep learning · Mac

1 Introduction

In recent years, biological research has become increasingly rich in data leading to several new subfields of biology denoted by the suffix "-omics". These research fields aim at characterizing and quantifying entire categories of biologically relevant molecules, defined by the suffix "-ome". As such, genomics is the field of research where the entire set of genes (i.e. genome) is characterized or analogously proteomics aims at characterizing and quantifying the entire set of proteins (i.e. proteome).

One of the relatively newcomers in the omic-era is metabolomics. Thousands of small molecules (<1500 Da), i.e. metabolites, are measured in a metabolomics experiment in a given biological sample [1]. By measuring the metabolome, an overview of diverse metabolic processes in matrices as varied as plants, environmental samples, tissues, cell or bacterial cultures, plasma, serum, skin or feces are obtained [2–7]. As

D. Cantone and A. Pulvirenti (Eds.): *From Computational Logic to Computational Biology*, LNCS 14070, pp. 128–144, 2024.
https://doi.org/10.1007/978-3-031-55248-9_7

such, among all the omics, metabolomics is the closest to the phenotype of an organism and thus has enormous potential in the field of precision medicine. The metabolome is a read-out of the combination of genetic, environmental, microbial and dietary factors that all influence the metabolic state of an individual. The metabolic state of an individual is closely related to the overall health status and an improved understanding could aid in prediction, diagnosis, prognosis and elucidation of molecular mechanisms of diseases [8]. However, the complexity and high dimensionality of data retrieved from metabolomics experiments, as well as the lack of sufficient public databases and automated data analysis workflows has hampered the integration of metabolomics data into the clinic. Deep learning methods applied to metabolomics data could aid the effective integration, and application of metabolomics data within a clinical context. This task is currently still largely dependent on expert knowledge with considerable investment in time for manual analysis, annotation and interpretation of the data.

In this book chapter we review deep learning methods applied to liquid chromatography mass spectrometry (LC-MS) metabolomics data, and highlight computational workflows that are paving the way towards a high(er)-throughput use of untargeted metabolomics data, making it effective for clinical, environmental and other types of applications.

1.1 Metabolomics

Metabolomics approaches can be grouped into targeted and untargeted measurements. In targeted metabolomics experiments certain metabolites of interest are pre-defined, while the aim of untargeted metabolomics approaches is to measure the entirety of all small molecules present in a given sample [9]. Advantages of targeted metabolomics approaches include higher reproducibility and ease of high throughput during data analysis and interpretation. The approach is however unable to realize the ambitious aims of metabolomics research, to measure the entire set of metabolites in a given sample. To get closer to this goal, untargeted metabolomics experiments are performed. Analysis and interpretation of untargeted metabolomics data is more challenging but allows for more explorative research and de-novo hypothesis generation. The number of metabolites measurable within the human body is unknown, but estimated at several thousand or more. The Human Metabolome Database (HMDB), an online database that aims to provide documentation for all known metabolites, has over 100,000 metabolite entries [10].

1.2 Liquid Chromatography Mass Spectrometry

Mass spectrometry (MS) is the analytical work horse of metabolomics measurements and remarkable developments in instrumentation technology have led to a rapid increase in the application of MS-based metabolomics platforms in clinical and research laboratories over the past decade [11]. High-resolution mass spectrometers allow for the simultaneous chemical structural elucidation and quantification of hundreds to thousands of small molecules in a sample with high sensitivity, specificity, throughput, and low sample consumption in a cost effective manner [11, 12]. Chemical characterization is further aided by integrating liquid chromatography with mass spectrometry [13]

(LC-MS), enabling quantitative and qualitative analysis of an increasing number of metabolites. Consequently, LC-MS techniques are widely used in clinical research for biomarker discovery [14], elucidation of molecular mechanisms of diseases [15] or monitoring disease therapy [16].

MS-based methods provide two levels of information, the mass-to-charge ratio (*m/z*) and abundance, which is summarized in a mass spectrum. When utilizing high resolution MS, the precision of the measured *m/z* is typically high enough for it to be mapped to a specific molecular formula, in particular when specific filters on the occurrence of elements are used, which in turn can be assigned to candidate molecules. The abundance is the measure of the number of times a specific *m/z* hits the detector and can be regarded as the relative concentration. However, measuring the exact mass (*m/z*) is often not enough to properly identify a metabolite, since numerous isomeric compounds could give rise to the same *m/z* [17]. To resolve the identity, it is usually necessary to perform MS/MS experiments, where a specific *m/z* is isolated, fragmented and subsequently subjected to an additional MS analysis. Fragmentation of the molecule of interest is useful because the fragmentation pattern is dependent on the molecular structure. The resulting fragmentation spectrum shows the *m/z* of the molecule fragments, which can be matched against databases of experimentally determined fragmentation patterns for different metabolites [18]. LC results in one additional dimension of data, the chromatographic retention time (RT). Mass spectra are continuously generated throughout the chromatographic separation of the samples, typically acquiring one or several mass spectra each second. Since the LC separates the compounds in the sample based on physico-chemical properties, each metabolite has a characteristic RT for the specific chromatographic setup it was generated on. Therefore, RT can be an additional variable that can be utilized when annotating untargeted metabolomics data [19–21].

Targeted metabolomics experiments typically measure up to a few hundred metabolites with known *m/z* and fragmentation patterns but are unable to detect metabolites that were not selected as targets. For untargeted metabolomics, high-resolution mass spectra are generated and are typically able to measure thousands of mass spectral features (i.e. metabolites before the identification) [13]. Since untargeted metabolomics is not hypothesis driven, much effort has to be put into detecting, aligning, and annotating each measured feature, noting its *m/z* and RT, relative abundance in a feature quantification table, and finally linking each feature to its fragmentation spectrum.

1.3 LC-MS Metabolomics Data Analysis and Interpretation

LC-MS approaches are powerful methods but the data that they generate are highly dimensional and extremely complex hindering several stages of metabolomics data treatment, including data preprocessing (the transformation of the multiple chemical signal dimensions into an easy-to-use format for subsequent statistical analysis) [22], analysis, chemical structural annotation and interpretation. Non-preprocessed LC-MS data files contain several thousands of MS spectra one after another and each spectrum is characterized by a sequence number which increases with the RT. Currently, several pipelines exist for automated LC-MS data preprocessing [23–27], peak detection and alignment across samples remains however challenging and often produces false positives. In addition, available modular framework algorithms are not scalable, but feasible

for a few hundred samples, making them unsuitable for the analysis and processing of population-size clinical metabolomics cohorts. Another bottleneck is the chemical structural annotation of unknowns in untargeted metabolomics data. Compared to other omics sciences, chemical structural identification of metabolomics data is more challenging, as the estimated chemical space of small molecules is vast. Over 10 [60] possible carbon-based small molecules (<500 Da) are thought to exist [28], as compared to 20 unique amino acids, or four nucleotides, which form the building blocks of all proteins and DNA measured in proteomics and genomics experiments respectively. On average only 2–5% of the data collected in a typical LC-MS metabolomics experiment can be matched to known molecules [29] through private and public spectral libraries and the ever-increasing accumulation of unidentified metabolites reported in clinical metabolomics studies is a major bottleneck [30]. In more recent years, computational metabolomics workflows have been proposed based on substructure discovery, chemical compound class annotation, and mass spectral networking to make an inroad into the vast amount of yet unknown metabolite features in untargeted metabolomics profiles [31–38].

1.4 Machine Learning Applied to Untargeted Metabolomics

In the last few years, machine learning (ML)-based technologies entered every aspect of society, helping image and speech recognition, recommendation systems and many more tools that we use in our daily life. A typical form of ML is supervised learning, where the machine is trained with a large amount of data having labeled information and it gives as output a vector of scores for each category. Supervised learning includes models such as Random Forest [39] and Support Vector Machine (SVM) [40]. After training, the performance is evaluated on a different dataset, never seen during training (which is called test set) to verify the generalization ability of the ML model. In contrast, unsupervised learning is a form of ML where an algorithm learns patterns from unlabeled data. Typical examples of those methods are k-means clustering, Principal Component Analysis and Hierarchical Clustering [41].

Several ML-based approaches have been proposed for improving untargeted metabolomics data analysis (see Table 1). Taking inspiration from other fields of computer science and statistics, some supervised and unsupervised ML methods have been applied to substructure discovery and annotation (e.g., MS2LDA substructure discovery [38], MAGMa - ClassyFire - MS2LDA integration & MotifDB [42]), for improving spectral similarity scores (e.g., Spec2Vec [43]), and for large-scale mass spectral clustering and networking (e.g., falcon [44] and Molecular Networking [35]).

1.5 Deep Learning Applied to Untargeted Metabolomics

ML-based methods usually require careful feature engineering and a high level of knowledge of the specific field, in order to apply proper transformations to the raw data [46]. This step is crucial for most of the classical ML approaches and is needed for giving the proper inputs to ML methods. Therefore, alternative ML models have been proposed based on representation learning which are able to use data in the raw form and discover meaningful patterns. Deep learning (DL) models are a type of representation

Table 1. Recent ML methods applied to untargeted metabolomics.

Task	Name	Model	Availability	Year	Ref.
Spectral similarity score	Spec2Vec	Natural language processing algorithm (inspired by Word2Vec)	https://github.com/iomega/spec2vec	2021	[43]
Metabolite identification	MS2LDA	Topic modeling (latent Dirichlet allocation)	https://ms2lda.org/	2016	[38]
Metabolite identification	MAGMa - ClassyFire - MS2LDA integration	Unsupervised and supervised ML model - Multilayer Neural Network	https://github.com/sdrogers/lda https://github.com/iomega/motif_annotation https://github.com/sdrogers/nnpredict https://github.com/sdrogers/ms2ldaviz	2019	[42]
Metabolite identification	SIRIUS 4	Support vector machine (SVM)	https://bio.informatik.uni-jena.de/sirius/	2019	[45]
Spectrum clustering	Falcon	Tandem mass spectrum clustering using fast nearest neighbor searching	https://github.com/bittremieux/falcon	2021	[44]
Multiple tasks including metabolite identification and spectrum clustering	GNPS	Web-based MS platform including ML approaches for metabolite identification and clustering	https://gnps.ucsd.edu/	2016	[33, 35]

learning methods that are gaining more visibility in recent years. To improve some of the mentioned issues related to preprocessing and, overall, metabolomics data analyses, DL models have been proposed as a valid alternative for peak detection [47], identification of molecular structures [48] and, among other tasks, for batch effect removal [49].

In the last few years, DL has seen an impressive increase of applications in many scientific fields and we are observing a tremendous impact on society. In particular,

DL has been applied for improving classical ML-based approaches which are related to image and speech recognition and natural language processing. The computational biology field has shown a great interest in DL methods as well, particularly for discovering patterns in the data that could discriminate disease status and stratify patients according to meaningful features. However, for many years the low volume of data available and the challenging interpretation of DL models have limited the applications in computational biology and medicine. The increased availability of public datasets in the mass spectrometry field has opened new opportunities. However, in many cases it is still not enough for building proper DL models. In this context, we foresee more applications of transfer learning in the coming years, which is the capability to reuse knowledge obtained from other learning tasks for new and often unrelated tasks. A successful example is the use of the large ImageNet database (https://www.image-net.org/) containing 14,197,122 images, which has been instrumental for transfer learning applied to image recognition and advancing the DL field.

In the following, we present an overview of recent computational methods that implement DL models for different steps of metabolomics data analysis. These approaches were applied for improving preprocessing, feature identification and other tasks. Particularly, we focus our attention on DL methods applied to LC-MS.

2 Overview of DL Methods

Here, we present an overview of some of the recently published papers on DL models applied to the metabolomics field, in particular LC-MS methods (see Table 2). Overall, only few approaches have been published in the last few years, indicating that this is a novel and unexplored aspect of metabolomics research.

These methods include different steps of untargeted metabolomics data analysis such as batch effect correction, peak detection, spectra similarity and prediction of metabolite classes (Fig. 1). In the following sections, we will present each individual method in detail showing the innovative aspects of these approaches compared to traditional methods.

3 DL Methods for Pre- and Postprocessing Metabolomics Data

3.1 Peakonly: A DL Model for Detecting and Integrating Peaks

One of the main preprocessing tasks in untargeted metabolomics is peak detection and integration of LC-MS data, which is currently prone to several false positive signals. To overcome this challenging issue, Melnikov and co-authors developed peakonly [47], an algorithm which is able to detect and exclude low-intensity noisy peaks starting from raw data. Peakonly is based on a convolutional neural network (CNN) which classifies regions of interest (ROIs). ROIs were detected using a modified version of the *centWave* algorithm [54]. The algorithm classifies three classes: 1) noise (ROIs do not contain peaks), 2) ROI contains one or more peaks, 3) ROI might contain a peak but a particular attention is required and in this case the decision has to be taken by an expert. Furthermore, the peak integration step was considered as a segmentation problem using an additional CNN, which allowed the authors to define whether a point in a ROI belongs to a peak.

Table 2. DL methods published recently for improving different LC-MS based metabolomics data analyses.

Task	Name	Model	Availability	Year	Ref
Similarity measure to compare tandem mass spectra	DeepMASS	Customized deep neural network	https://github.com/hcji/DeepMASS	2019	[50]
Batch effect removal	NormAE	Deep Adversarial Learning Model	https://github.com/luyiyun/NormAE	2020	[49]
Peak detection	Peakonly	Convolutional neural network	https://github.com/arseha/peakonly	2020	[47]
Peak detection	NeatMS	Convolutional neural network	https://github.com/bihealth/NeatMS	2022	[51]
Prediction of compound classes	CANOPUS	Customized deep neural network	https://bio.informatik.uni-jena.de/software/canopus/	2020	[48]
Characterization and expansion of reference libraries for small molecule identification	DarkChem	Variational autoencoder (VAE)	https://github.com/pnnl/darkchem	2020	[52]
Similarity measure to compare tandem mass spectra	MS2DeepScore	Siamese neural network architecture	https://github.com/matchms/ms2deepscore	2021	[53]
Prediction of liquid chromatographic retention times (RTs)	GNN-RT	Graph neural network	https://github.com/Qiong-Yang/GNN-RT	2021	[20]

This method showed high precision for solving the hard task of detecting and defining a peak area, discovering true positive peaks using in-house and publicly available datasets. The approach achieved a precision of 97% which is a very high value considering that the precision of existing algorithms without additional noise filtering ranges from 0.5 to 0.8 [47, 55]. The approach has been developed specifically for LC-MS but it is potentially applicable to other MS systems, even though some adjustments might be needed, especially the training of the CNNs.

The method was written in pytorch and it is available at https://github.com/arseha/peakonly.

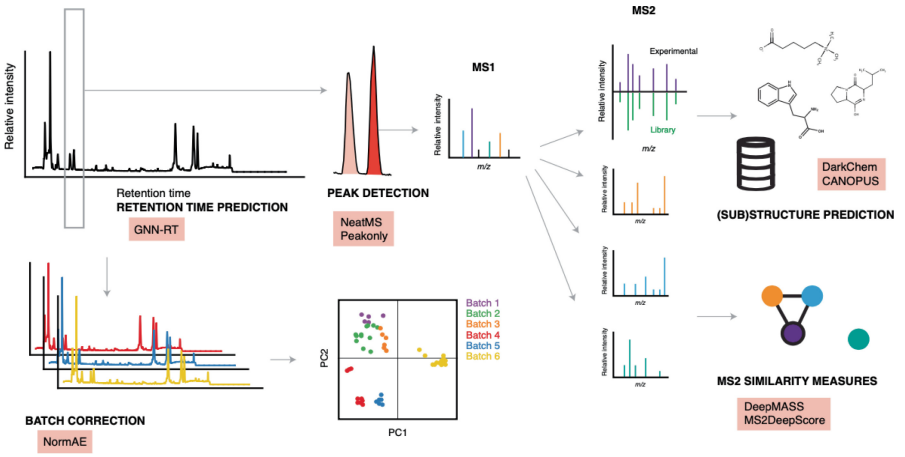

Fig. 1. Overview of recent deep learning methods applied to untargeted LC-MS MS1 and MS2 (MS/MS) metabolomics data analysis. Deep learning methods have been applied within four different subareas of untargeted metabolomics data analysis, including retention time prediction (GNN-RT), peak detection (NeatMS, Peakonly), (sub)structure prediction (DarkChem, CANOPUS) and tandem mass spectral similarity measures (DeepMASS, MS2DeepScore).

3.2 NeatMS: A DL-Based Post Processing Tool to Remove Low Quality Peaks

NeatMS is a tool to perform peak curation for LC-MS based metabolomics [51]. Where Peakonly removes low quality peaks from the raw data, NeatMS uses an alternative approach by filtering out poor peaks from data already preprocessed by conventional algorithms. NeatMS is based on a convoluted neural network architecture that was trained on datasets with a wide range of peak shapes. The algorithm comes with a pre-trained model but transfer learning functionality is also available. Peak filtering with NeatMS can be implemented after an existing metabolomics preprocessing workflow by categorizing detected peaks as high, acceptable or poor quality. The method managed to retain chemical standards that were spiked into a complex data matrix with high precision (94–97%). In a head-to-head comparison to Peakonly, the corresponding numbers were 79%. Also, NeatMS included a larger number of total peaks (acceptable quality or higher).

NeatMS is written in python and available at https://github.com/bihealth/NeatMS.

3.3 NormAE: The Normalized Autoencoder for Removing Batch Effects

Untargeted metabolomics data are usually affected by strong batch effects [56, 57], which is defined as systematic technical differences between samples due to the measurement of metabolites in several batches or analytical plates. To be considered as a technical difference, the variation observed between samples can not be explained by any biological difference. Batch effects limit downstream analyses and the interpretability of the data since they mask the real biological effect. This is a fundamental aspect that has to be taken into account when performing an untargeted metabolomics study, since this type of metabolomics gives global information about the metabolome without

knowing absolute concentrations. Therefore, it is crucial to reduce batch effects in order to discover strong biological processes and biomarkers to be translated into the clinic.

Several methods aiming to remove batch effects have been proposed. Many of these methods take inspiration from other omics data and are based on linear models [58–61], which are not always based on the correct assumptions for metabolomics data. Moreover, some of the methods usually take into account inter-batch effects but they ignore intra-batch effects, which are commonly observed in untargeted metabolomics studies. To overcome these limitations, Rong and collaborators developed the normalized autoencoder (NormAE) [49], which is inspired by adversarial learning and autoencoders. The authors demonstrated that NormAE was able to remove non-linear batch effects from LC-MS untargeted metabolomics datasets without overfitting, better than the current state of the art methods.

NormAE takes as input preprocessed untargeted metabolomics data files and applies non-linear autoencoders and adversarial generative networks (GANs), with the goal to encode the original intensities of the peaks with latent representations. In this process, only the actual biological information is retained and during the decoding step the data are reconstructed without inter-batch effects. NormAE achieves the goal of removing the inter-batch effects by implementing an adversarial training step which aims to optimize the autoencoder by classifying the labels (i.e. batches) based on the latent representation. Furthermore, metabolomics-specific intra-batch effects, due for instance to injection order, are also taken into account and added to the model.

The authors evaluated the method on two datasets and based on the removal of the batch effect and the ability of the method to retain biological information, it outperformed the current state of the art approaches reducing the variance due to batch effect.

NormAE is publicly available at https://github.com/luyiyun/NormAE and it has been implemented in pytorch.

4 DL Methods for Metabolite Annotation

4.1 GNN-RT: Improving Metabolite Annotation by Predicting RT Using Graph Neural Networks

The chromatographic RT is strongly correlated to the molecular structure of metabolites and, therefore, it can be used for improving molecular identification and reducing false positives. However, in untargeted metabolomics RT is not commonly used for identifying small molecules because the current computational approaches are usually instrument-specific and most of the methods have been applied to small datasets. In order to improve RT prediction methods, the large METLIN small molecule retention time (SMRT) dataset [19] was released for boosting machine learning-based models and improving metabolite annotation.

GNN-RT is a method based on graph neural networks (GNNs), which have a high learning ability [20]. Recently, GNNs gained visibility in the metabolomics field because of the possibility of representing small molecules as graphs, which better describe the relation between atoms and bonds compared to fixed molecular descriptors and fingerprints. Therefore, methods based on graphs are more accurate and precise by taking into account the topological and chemical properties of small molecules [20].

The input of GNN-RT is a set of molecular graphs, which are based on the international chemical identifier (InChI) [62] and generated using the RDKit (https://www.rdkit.org/). Each graph can be described as vertices (i.e., atoms) and edges (i.e., chemical bonds between atoms). Additionally, the number of bonds was encoded as well. The molecular representation can be seen as a low dimensional vector which is learned by backpropagation. The learned representation is then used to predict RT and support metabolite annotation. Furthermore, an interesting and useful aspect of GNN-RT is its ability to perform transfer learning between different chromatographic systems. This is of particular importance because different systems will generate different RTs and makes it challenging to compare them. To overcome this issue, GNN-RT can be pretrained with a large datasets such as SMRT, and used on a new dataset (even if generated using a different chromatographic system) by transfer learning, allowing to obtain high performance and avoiding the need of training on larger datasets.

Overall, GNN-RT achieved the highest performance based on different metrics including MAE = 39.87, MedAE = 25.24, and MRE = 4.9%, compared to multichannel-CNN (MC-CNN) [63], single channel-CNN (SC-CNN) [63], Bayesian ridge regression (BRR) [21] and Random Forest (RF) [21].

GNN-RT has been written in pytorch and is available at https://github.com/Qiong-Yang/GNN-RT.

4.2 DeepMASS: Structural Similarity Scoring of Unknown Metabolites Using Deep Neural Networks

As we mentioned in the previous paragraphs, the metabolomics field needs novel and advanced tools for metabolite annotation. MS/MS spectra generated in tandem mass spectrometry are a large source of knowledge, representing substructure information of metabolites. A classical approach for identifying metabolites consists of searching MS/MS spectra in a publicly available database, hoping to find a match. The databases containing MS/MS spectra are undoubtedly increasing, however they are still limited compared to the hypothetical large number of metabolites existing in nature.

In this context, DeepMASS [50] tries to solve the above limitations of current approaches using deep neural networks. The idea behind this model is based on biotransformations and the fact that reactant and product metabolites have similar substructures. Therefore, it is possible to search unknown metabolites in MS/MS spectra databases looking for transformational products of known metabolites.

The DeepMASS model trains and validates metabolite pairs characterized by high structure similarity (i.e., 'positive metabolite pair') retrieved from the KEGG database and random metabolite pairs (i.e., 'negative metabolite pair') generated randomly. Then, for each positive or negative metabolite pair their spectra are searched against MS/MS databases. Due to the limited number of experimental MS/MS spectra after matching, an additional tool (CFM-ID) was used for generating more spectra, increasing the total number of spectra needed for applying DeepMASS. Then, theoretical spectra pairs were collected in the same way as experimental spectra pairs and used for pretraining the deep neural network, while the experimental spectra pairs were used for fine-tuning the model. Using pretrained networks allows to apply DL models overcoming the issue of having a small number of experimental spectra pairs.

To validate the identification performance of DeepMASS, the authors performed a cross validation test based on 662 spectra. The percentage that the correct structure was found among the top 3, top 5, and top 10 hits was reported as 74.9%, 85.3%, and 92.0%, respectively, achieving remarkable performance compared to other methods (MetFrag [64] and CFM-ID [65]). Additionally, the authors expanded the search with the entire PubChem compound database achieving the highest percentage of identification compared to MetFrag and CFM-ID.

DeepMASS was implemented in Keras and Tensorflow backend and it is publicly available at https://github.com/hcji/DeepMASS.

4.3 MS2DeepScore: A Siamese Neural Network for Predicting the Structural Similarity of MS/MS Fragmentation Spectra

Following the recent development and improvements for predicting the structural similarity of MS/MS spectra using DL models, Huber and colleagues proposed a novel approach inspired by DeepMASS [50]. This new method, MS2DeepScore [53], uses a simpler architecture based on a Siamese neural network and it only relies on peak *m/z* positions and intensities instead of using mass and chemical formulas. An additional advantage of MS2DeepScore is that it creates mass spectral embeddings which can be the input for additional spectral clustering. The method has been evaluated on curated and cleaned spectra retrieved from the Global Natural Product Social Molecular Networking (GNPS) [35], reaching high performance.

The inputs of MS2DeepScore are pairs of MS/MS spectra. The model is based on a Siamese network which consists of two components, a base network creates the embeddings from the input spectra and another part of the network which is a cosine calculation of the embeddings. Additionally, MS2DeepScore applies Monte-Carlo Dropout ensembles to estimate the uncertainty of a prediction.

The advantage of using MS2DeepScore relies on the fact that it is trained to predict structural similarities, usually obtained by applying Tanimoto or Dice scores based on molecular fingerprints, directly from pairs of MS/MS spectra without the need to compute molecular fingerprints. In the introducing paper, the authors show that MS2DeepScore can predict, based on MS/MS spectral pairs, the Tanimoto scores (between 0.1 and 0.9) of the fragmented molecules with an RMSE between 0.13 and 0.2.

MS2DeepScore is available as a python library at https://github.com/matchms/ms2deepscore.

4.4 CANOPUS: A Computational Tool for Systematic Compound Class Annotation

One of the most recent works on deep learning applied to metabolite annotation is CANOPUS [48]. CANOPUS is based on deep neural networks. It uses predicted fingerprints as input for assigning classes and ontology to metabolites. The authors used the probabilistic molecular fingerprint predicted by CSI:FingerID as well as the molecular formula computed by SIRIUS [45] as input. In particular, MS/MS spectra are the input of support vector machines (SVMs), which are trained with reference MS/MS spectra and used to predict a probabilistic fingerprint. Then, the probabilistic fingerprint is used

as the input of a deep neural network trained on 4.1 million compounds without needing MS/MS spectra as input and gives predicted classes as output.

CANOPUS showed a very high prediction performance with an average accuracy of 99.7% in cross-validation, using reference data.

Advantages of CANOPUS include the ability to assign putative compound classes to every mass spectral feature in a LC-MS experiment, including molecules which have not been previously reported in any database. The algorithm is available as source code and software, making it also available for users with limited bioinformatics skills.

The source code of CANOPUS is available at https://github.com/boecker-lab/sirius-libs and the software implementation at https://bio.informatik.uni-jena.de/software/canopus/.

4.5 DarkChem: A Variational Autoencoder for Creating a Massive in Silico Library

One of the most important approaches for small metabolite annotation relies on the comparison of experimental features characterized by *m/z* and RT to libraries containing reference values, which help the identification of known molecules. However, these libraries are not able to identify unknown molecules and therefore most of the available commercial reference standards are not enough for untargeted metabolomics experiments. On the other hand, modern *in silico* approaches have the potential of building large libraries and boosting the identification process.

A recent approach, DarkChem [52], aims to create a massive *in silico* library using a variational autoencoder (VAE). One of the key qualities of DarkChem is the ability to include collision cross section (CCS) in the model, which is obtained from ion mobility spectrometry and measures the interaction between the ionized molecule and a buffer gas. CCS represents an additional dimension for small metabolite annotation, enabling the measuring of the mobility of the molecule in the mass spectrometer. Furthermore, DarkChem contains a 3-stage transfer learning method which allows it to learn important molecular structure representations from millions of molecules. Then, an optimization step improves the ability of the model to predict chemical properties. As we mentioned in the previous paragraphs, transfer learning is a useful approach in cases where the experimental datasets are too small and the risk of overfitting is too high.

The network architecture of DarkChem included an encoder consisting of canonical SMILES, a character embedding and convolutional layers. The latent representation was a fully connected dense layer and the decoder was convolutional layers and a linear layer with softmax activation for giving the outputs (i.e. canonical SMILES).

The model achieved a validation reconstruction accuracy of 98.9% for the experimental dataset and 99.0% for the in silico dataset, demonstrating the ability of transfer learning to improve performances in case of small experimental datasets.

DarkChem was written in python using Keras and Tensorflow backend and it is available at https://github.com/pnnl/darkchem.

5 Conclusions

In this book chapter, we presented the most recent works regarding DL and its applications to untargeted metabolomics. We introduced the main characteristics of the DL models and their strength for solving several bottlenecks in the metabolomics field. As we have shown, different deep neural network architectures share a common strategy based on transfer learning. The amount of publicly available datasets in this field is growing through platforms such as MassBank [66], GNPS [35] or MetaboLights [67] where raw, processed, or annotated mass spectrometry data are shared. However, DL methods require a large amount of data to perform and to avoid overfitting. Transfer learning is an effective approach for learning fundamental aspects of the data and for generalizing models that can be applied to different and often unrelated tasks.

We observe the development of DL models spanning several topics related to computational metabolomics, but undoubtedly the small metabolite annotation has seen enormous progress in recent years. With the ever growing amount of mass spectrometry data being collected and shared in public repositories, we and others [68] foresee the development of many more DL methods aiming to identify the unknown molecules and finally reconstruct complex metabolic pathways related to biological processes and diseases.

In this book chapter, we focused our attention on DL methods applied to LC-MS-based metabolomics. However, DL is having an impact on other techniques and research fields such as gas chromatography mass spectrometry (GC-MS) [69], nuclear magnetic resonance (NMR) spectroscopy [70, 71], Matrix Assisted Laser Desorption-Ionization Time-of-Flight mass spectrometry (MALDI-TOF) [72] and proteomics [73, 74]. We believe that these approaches can inspire each other to improve future methods and generate new ideas, for solving bottlenecks within and between the research fields.

Sharing metabolomics data will help increase the number of available datasets for training ML models and become better at performing classification tasks [75]. However, in the context of translational research several ethical considerations remain to be addressed, therefore data sharing remains challenging. This is an aspect of research which has been faced by genetic data [76, 77], which are considered personal information and therefore have to follow the data protection regulations [78]. However, the ability of metabolomics studies to identify research participants is largely unknown [79]. It has been proposed that a more extensive controlled data access might be needed for metabolomics data in order to be shared [79]. This requires that the availability of data will be limited to researchers that have been authorized by appropriate data access committees and a proper infrastructure will be needed.

References

1. Fiehn, O.: Metabolomics — the link between genotypes and phenotypes. In: Functional Genomics, pp. 155–171, Springer, Netherlands (2002). https://doi.org/10.1007/978-94-010-0448-0_11
2. Zierer, J., et al.: The fecal metabolome as a functional readout of the gut microbiome. Nat. Genet. **50**, 790–795 (2018)
3. Psychogios, N., et al.: The human serum metabolome. PLoS ONE **6**, e16957 (2011)

4. Dame, Z.T., et al.: The human saliva metabolome. Metabolomics **11**, 1864–1883 (2015)
5. Beltran, A., et al.: Assessment of compatibility between extraction methods for NMR- and LC/MS-based metabolomics. Anal. Chem. **84**, 5838–5844 (2012)
6. Dietmair, S., Timmins, N.E., Gray, P.P., Nielsen, L.K., Krömer, J.O.: Towards quantitative metabolomics of mammalian cells: development of a metabolite extraction protocol. Anal. Biochem. **404**, 155–164 (2010)
7. Elpa, D.P., Chiu, H.-Y., Wu, S.-P., Urban, P.L.: Skin Metabolomics. Trends Endocrinol Metab **32**, 66–75 (2021)
8. Beger, R.D., et al.: Metabolomics enables precision medicine: 'a white Paper. Community Perspect. Metabolomics **12**, 149 (2016)
9. Fiehn, O.: Metabolomics–the link between genotypes and phenotypes. Plant Mol. Biol. **48**, 155–171 (2002)
10. Wishart, D.S., et al.: HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. **46**, D608–D617 (2018)
11. Jannetto, P.J., Fitzgerald, R.L.: Effective use of mass spectrometry in the clinical laboratory. Clin. Chem. **62**, 92–98 (2016)
12. Chace, D.H., Kalas, T.A., Naylor, E.W.: Use of tandem mass spectrometry for multianalyte screening of dried blood specimens from newborns. Clin. Chem. **49**, 1797–1817 (2003)
13. Wishart, D.S.: Metabolomics for investigating physiological and pathophysiological processes. Physiol. Rev. **99**, 1819–1875 (2019)
14. Liang, Q., Liu, H., Xie, L.-X., Li, X., Zhang, A.-H.: High-throughput metabolomics enables biomarker discovery in prostate cancer. RSC Adv. **7**, 2587–2593 (2017)
15. Johnson, C.H., Ivanisevic, J., Siuzdak, G.: Metabolomics: beyond biomarkers and towards mechanisms. Nat. Rev. Mol. Cell Biol. **17**, 451–459 (2016)
16. van der Hooft, J.J.J., Padmanabhan, S., Burgess, K.E.V., Barrett, M.P.: Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. Metabolomics **12**, 125 (2016)
17. Sumner, L.W., et al.: Proposed minimum reporting standards for chemical analysis chemical analysis working Group (CAWG) metabolomics standards initiative (MSI). Metabolomics **3**, 211–221 (2007)
18. Xiao, J.F., Zhou, B., Ressom, H.W.: Metabolite identification and quantitation in LC-MS/MS-based metabolomics. Trends Analyt. Chem. **32**, 1–14 (2012)
19. Domingo-Almenara, X., et al.: The METLIN small molecule dataset for machine learning-based retention time prediction. Nat. Commun. **10**, 5811 (2019)
20. Yang, Q., Ji, H., Lu, H., Zhang, Z.: Prediction of liquid chromatographic retention time with graph neural networks to assist in small molecule identification. Anal. Chem. **93**, 2200–2206 (2021)
21. Bouwmeester, R., Martens, L., Degroeve, S.: Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. Anal. Chem. **91**, 3694–3703 (2019)
22. Katajamaa, M., Oresic, M.: Data processing for mass spectrometry-based metabolomics. J. Chromatogr. A **1158**, 318–328 (2007)
23. Pluskal, T., Castillo, S., Villar-Briones, A., Oresic, M.: MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf. **11**, 395 (2010)
24. Tsugawa, H., et al.: MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat. Methods **12**, 523–526 (2015)
25. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., Siuzdak, G.: XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. **78**, 779–787 (2006)

26. Röst, H.L., et al.: OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat. Methods **13**, 741–748 (2016)
27. Protsyuk, I., et al.: 3D molecular cartography using LC-MS facilitated by optimus and 'ili software. Nat. Protoc. **13**, 134–154 (2018)
28. Bohacek, R.S., McMartin, C., Guida, W.C.: The art and practice of structure-based drug design: a molecular modeling perspective. Med. Res. Rev. **16**, 3–50 (1996)
29. da Silva, R.R., Dorrestein, P.C., Quinn, R.A.: Illuminating the dark matter in metabolomics. Proc. Natl. Acad. Sci. U.S.A. **112**, 12549–12550 (2015)
30. Wood, P.L.: Mass spectrometry strategies for clinical metabolomics and lipidomics in psychiatry, neurology, and neuro-oncology. Neuropsychopharmacology **39**, 24–33 (2014)
31. Beniddir, M.A., et al.: Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. Nat. Prod. Rep. **38**, 1967–1993 (2021)
32. Ernst, M., et al.: MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. Metabolites **9**, 144 (2019)
33. Nothias, L.-F., et al.: Feature-based molecular networking in the GNPS analysis environment. Nat. Methods **17**, 905–908 (2020)
34. Wang, M., et al.: Mass spectrometry searches using MASST. Nat. Biotechnol. **38**, 23–26 (2020)
35. Wang, M., et al.: Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol. **34**, 828–837 (2016)
36. Mohimani, H., et al.: Dereplication of microbial metabolites through database search of mass spectra. Nat. Commun. **9**, 4035 (2018)
37. Scheubert, K., et al.: Significance estimation for large scale metabolomics annotations by spectral matching. Nat. Commun. **8**, 1494 (2017)
38. van Der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V., Rogers, S.: Topic modeling for untargeted substructure exploration in metabolomics. Proc. Natl. Acad. Sci. **113**, 13738–13743 (2016)
39. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
40. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
41. Malik, A., Tuckfield, B.: Applied unsupervised learning with R: uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA. Packt Publishing Ltd (2019)
42. Rogers, S., et al.: Deciphering complex metabolite mixtures by unsupervised and supervised substructure discovery and semi-automated annotation from MS/MS spectra. Faraday Discuss. **218**, 284–302 (2019)
43. Huber, F., et al.: Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. PLoS Comput. Biol. **17**, e1008724 (2021)
44. Bittremieux, W., Laukens, K., Noble, W.S., Dorrestein, P.C.: Large-scale tandem mass spectrum clustering using fast nearest neighbor searching. Rapid Commun. Mass Spectrom. e9153 (2021)
45. Dührkop, K., et al.: SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat. Methods **16**, 299–302 (2019)
46. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
47. Melnikov, A.D., Tsentalovich, Y.P., Yanshole, V.V.: Deep learning for the precise peak detection in high-resolution LC–MS data. Anal. Chem. **92**, 588–592 (2020)
48. Dührkop, K., et al.: Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat. Biotechnol. **39**, 462–471 (2020)
49. Rong, Z., et al.: NormAE: deep adversarial learning model to remove batch effects in liquid chromatography mass spectrometry-based metabolomics data. Anal. Chem. **92**, 5082–5090 (2020)

50. Ji, H., Xu, Y., Lu, H., Zhang, Z.: Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification. Anal. Chem. **91**, 5629–5637 (2019)
51. Gloaguen, Y., Kirwan, J.A., Beule, D.: Deep learning-assisted peak curation for large-scale LC-MS metabolomics. Anal. Chem. **94**, 4930–4937 (2022)
52. Colby, S.M., Nuñez, J.R., Hodas, N.O., Corley, C.D., Renslow, R.R.: Deep learning to generate chemical property libraries and candidate molecules for small molecule identification in complex samples. Anal. Chem. **92**, 1720–1729 (2020)
53. Huber, F., van der Burg, S., van der Hooft, J.J.J., Ridder, L.: MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. J. Cheminform. **13**, 84 (2021)
54. Tautenhahn, R., Böttcher, C., Neumann, S.: Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics **9**, 504 (2008)
55. Tengstrand, E., Lindberg, J., Åberg, K.M.: TracMass 2–a modular suite of tools for processing chromatography-full scan mass spectrometry data. Anal. Chem. **86**, 3435–3442 (2014)
56. Liu, Q., et al.: Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. Sci. Rep. **10**, 13856 (2020)
57. Wehrens, R., et al.: Improved batch correction in untargeted MS-based metabolomics. Metabolomics **12**, 88 (2016)
58. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**, 118–127 (2007)
59. Zhang, Y., Parmigiani, G., Johnson, W.E.: ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform **2**, lqaa078 (2020)
60. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. **42**, e161–e161 (2014)
61. Pang, Z., Chong, J., Li, S., Xia, J.: MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. Metabolites **10**, 186 (2020)
62. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC international chemical identifier. J. Cheminform. **7**, 23 (2015)
63. Matyushin, D.D., Sholokhova, A.Y., Buryak, A.K.: A deep convolutional neural network for the estimation of gas chromatographic retention indices. J. Chromatogr. A **1607**, 460395 (2019)
64. Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S.: MetFrag relaunched: incorporating strategies beyond in silico fragmentation. J. Cheminform. **8**, 3 (2016)
65. Allen, F., Pon, A., Wilson, M., Greiner, R., Wishart, D.: CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. Nucleic Acids Res. **42**, W94–W99 (2014). https://doi.org/10.1093/nar/gku436
66. Horai, H., et al.: MassBank: a public repository for sharing mass spectral data for life sciences. J. Mass Spectrom. **45**, 703–714 (2010)
67. Haug, K., et al.: MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. **48**, D440–D444 (2020)
68. Liu, Y., De Vijlder, T., Bittremieux, W., Laukens, K., Heyndrickx, W.: Current and future deep learning algorithms for tandem mass spectrometry (MS/MS)-based small molecule structure elucidation. Rapid Commun. Mass Spectrom. e9120 (2021)
69. Li, M., Wang, X.R.: Peak alignment of gas chromatography–mass spectrometry data with deep learning. J. Chromatogr. A **1604**, 460476 (2019)
70. Qu, X., et al.: Accelerated nuclear magnetic resonance spectroscopy with deep learning. Angew. Chem. Int. Ed. Engl. **59**, 10297–10300 (2020)
71. Hansen, D.F.: Using deep neural networks to reconstruct non-uniformly sampled NMR spectra. J. Biomol. NMR **73**(10–11), 577–585 (2019). https://doi.org/10.1007/s10858-019-00265-1
72. Normand, A.-C., et al.: Identification of a clonal population of aspergillus flavus by MALDI-TOF mass spectrometry using deep learning. Sci. Rep. **12**, 1575 (2022)

73. Meyer, J.G.: Deep learning neural network tools for proteomics. Cell Reports Methods **1**, 100003 (2021)
74. Mund, A., et al.: AI-driven deep visual proteomics defines cell identity and heterogeneity. bioRxiv 2021.01.25.427969 (2021). https://doi.org/10.1101/2021.01.25.427969
75. Jarmusch, S.A., van der Hooft, J.J.J., Dorrestein, P.C., Jarmusch, A.K.: Advancements in capturing and mining mass spectrometry data are transforming natural products research. Nat. Prod. Rep. **38**, 2066–2082 (2021)
76. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**, 321–324 (2013)
77. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**, 409–421 (2014)
78. Shabani, M., Borry, P.: Rules for processing genetic data for research purposes in view of the new EU general data protection regulation. Eur. J. Hum. Genet. **26**, 149–156 (2018)
79. Keane, T.M., O'Donovan, C., Vizcaíno, J.A.: The growing need for controlled data access models in clinical proteomics and metabolomics. Nat. Commun. **12**, 5787 (2021)