Predicting compositional fragments of compounds from their tandem mass spectra using deep neural networks

Yuhui Hong[1], Sujun Li[1,2], Yuzhen Ye[1], and Haixu Tang[1*]
[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, USA
[2]GlycoMS LLC, Bloomington, IN, USA
*Corresponding author. E-mail: hatang@indiana.edu

Tandem mass spectrometry (MS/MS) serves as a pivotal tool for identifying small molecules, traditionally by searching experimental MS/MS spectra against a reference MS/MS library of previously identified compounds. However, this approach is limited to those identified compounds and cannot be extended to the identification of novel compounds. Here, we aim to compute the chemical structure of compounds directly from their MS/MS spectra by first predicting their compositional fragments.
To achieve this goal, we trained a deep neural network based on temporal convolutions that takes as input MS/MS spectra and outputs the presence/absence of chemical fragments in the compounds. When trained and tested on a large dataset of compound spectra, the model can correctly predict all fragments for 31.4% of the spectra, while for 77.5% of the spectra, the model only makes false predictions on only seven or fewer fragments.

METHODOLOGY

1. Data preprocessing

For all four libraries, the LC-MS/MS acquired by QTOF mass spectrometer were first filtered with the following steps:

i. The mass spectra have less than five peaks are filtered out because they are typically unreliable.
ii. The m/z range is limited in (0, 1500], because few spectra have m/z above 1500.
iii. Only the molecules composite by the high-frequency atoms (C, H, O, N, F, S, Cl, P, B, I, Br) are retained.
iv. Only the spectra with high-frequency precursor types ([M+H]+, [M-H]-, [M+Na]+, [M+H-H2O]+) are retained.

The statistics of the datasets after filtration are summarized in the following table.

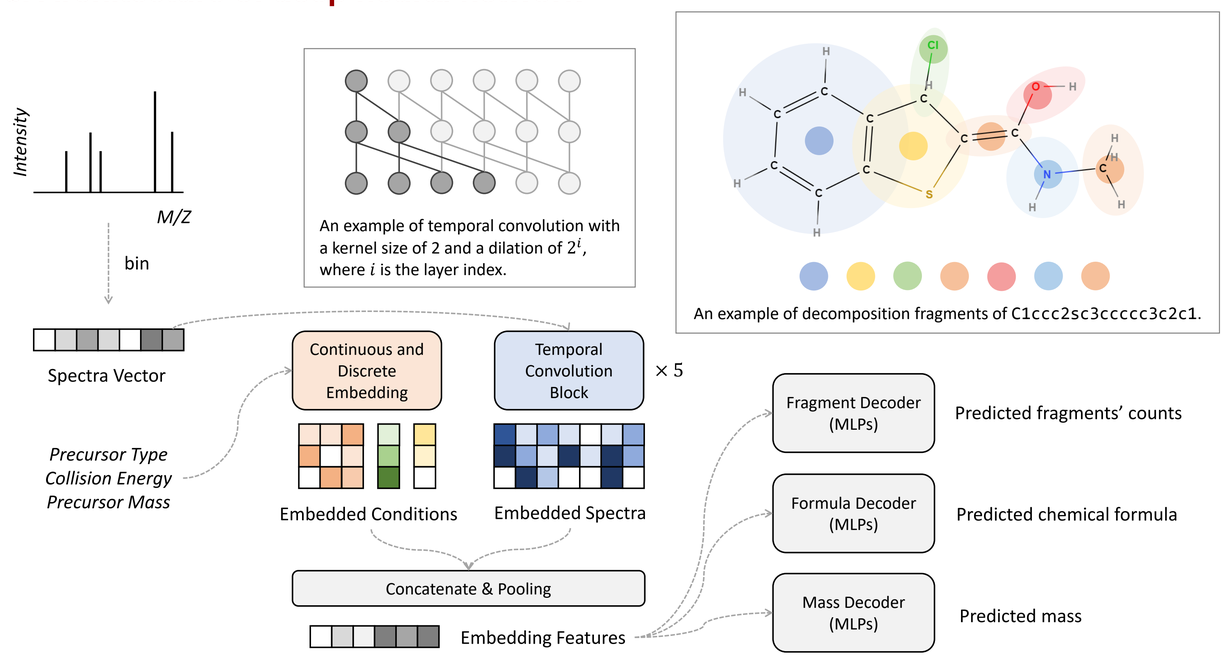
MS/MS Library	# Compound	# Mass Spectra
Agilent DPCL	12,133	42,410
NIST23	2,532	31,749
NIST20	2,544	32,057
MoNA	2,999	21,609
GNPS	1,270	2,642
Waters QTOF	612	757
In Total	17,725	131,224

2. Compounds decomposition algorithm

We adopted the first two steps of classical Junction Tree construction algorithm to decompose compounds, involving the following two steps:

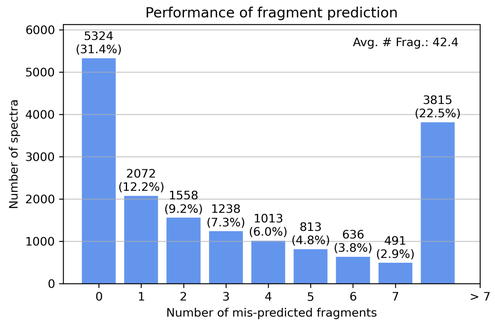
a. Moralization: The original molecular graph is moralized by adding edges between any two nodes that have a common neighbor and are not already connected. This step ensures that the graph is triangulated, meaning that every cycle of length 4 or greater has a chord (an edge connecting two non-adjacent nodes in the cycle).
b. Clique Finding: Maximal cliques (fully connected subgraphs) are identified in the moralized graph using the Bron-Kerbosch algorithm.

3. Architecture of deep neural network



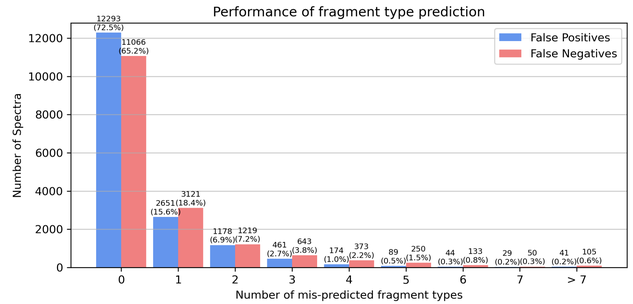
RESULTS

Fragment prediction



The fragments of 5,324 (31.4%) spectra are perfectly predicted. For 77.5% of the spectra, our model makes false predictions on seven or fewer fragments.

Fragment type prediction



On average, one compound is decomposed into 7.4 types of fragments. In total, the fragment types of 9,603 (56.6%) spectra are perfectly predicted.

Takeaway

- We presented a deep learning model to predict the decomposition fragments of compounds from their MS/MS spectra.
- Temporal convolution with large kernels is applied to extract features from MS/MS spectra, providing a large receptive field.
- In the future, the predicted fragments will be constructed into a complete molecular graph.

Acknowledgement

We appreciate Dr. Sujun Li, Dr. Christopher J. Welch, and Dr. Shane Tichy for their contribution of MS/MS data collection and invaluable advice on data preprocessing.

INDIANA UNIVERSITY