Prediction of Molecular Tandem Mass Spectra Using 3-Dimensional Conformers

Yuhui Hong[1]; Sujun Li[1]; Mingxun Wang[2]; Haixu Tang *[1]
[1]Indiana University, [2]University of California, San Diego

Novel Aspect

A deep neural network is proposed to predict molecular mass spectra and chemical properties from 3D conformers.

Introduction

Tandem mass (MS/MS) spectrometry is an essential technology for identifying and studying chemical compounds at high throughput, and thus is commonly adopted in metabolomics, drug discovery, and environmental chemistry. However, computational methods for automated compound identification from their MS/MS spectra are still limited, because the huge quality of molecular MS/MS has not been measured. Manual experiments are time-comsuming, so we established a deep learning algorithm to learn complex patterns from 3-dimensional molecular conformers and predict their MS/MS.

Different with molecular fingerprints or molecular graphs, we treat the 3D molecules as points of sets, which contains more geometric information. We designed an elemental operation, named **MolConv**, on 3D molecular conformers, from which we developed an efficient deep neural network (DNN) to predict MS/MS spectra.

In addition, we show that using transfer learning, the representation learned in spectra prediction can be transferred to the learning of other chemical properties such as retention time and collision cross section.

Preliminary Data

NIST20 [1], MassBank [2], and GNPS [3] are high-quality tandem mass spectral libraries containing 31k, 10k, and 6k compounds, respectively. Because the mass spectra resulting from different fragmentation methods are different, we only consider the spectra from the three most common methods:

- **HCD** (higher-energy collisional dissociation)
- **Q-TOF** (quadrupole/time-of-flight)
- **QqQ** (triple quadrupole)

We then merged the spectra of the same fragmentation methods from all libraries, and then randomly split compounds into the training (90%) and testing (10%) datasets, respectively.

Dataset	Instrument Type	# Mass Spectra	# Compounds
GNPS	QTOF	21,112	4,730
	QqQ	7,563	1,207
NIST20	HCD	535,283	21,037
	QTOF	30,870	2,167
	QqQ	21,285	1,700
MassBank	HCD	18,595	1,913
	QTOF	15,650	2,776
	QqQ	4,112	707
	Unknow	7,720	3,861

Methods

Data Encoding: The 3D conformers of molecules are encoded into sets of atomic coordinates and attributes. The mass spectra are encoded into vectors of a fixed length.

MS/MS Prediction: An end-to-end deep neural network based on sequential MolConv embeds a set of atoms into a latent vector, from which (as well as the metadata e.g., adducts, instruments and collision energy) MS/MS spectra are predicted. MolConv and max-pooling are insensitive to the order and can process atoms in any order and reach invariant output.

Chemical Properties Prediction: Furthermore, the encoder for embedding the input into the latent space is pre-trained using a massive training dataset and can be reloaded for transfer learning of other prediction tasks.

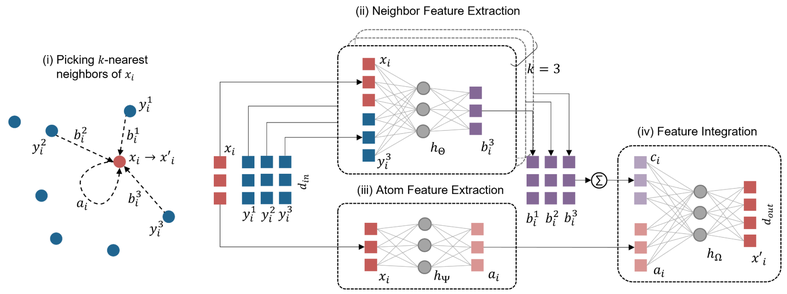

Figure 1. The convolution operation of MolConv.

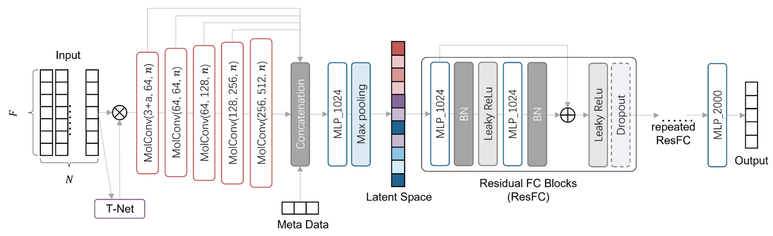

Figure 2. The architecture of Mol3DNet.

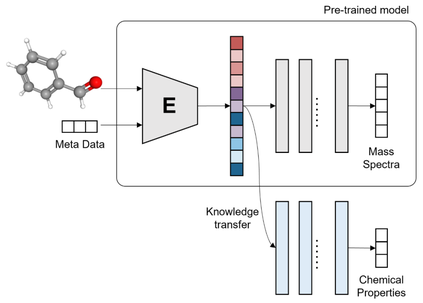

Figure 3. Transfer learning for chemical properties prediction.

Results – MS/MS Prediction

Cosine similarity is used to measure the accuracy between the predicted mass spectra and experimental spectra.

$$\cos(y, \hat{y}) = \frac{y \cdot \hat{y}}{\|y\|\|\hat{y}\|}$$

In all the precursor types, we got the following results:

Dataset	Instrument Type	Ours	Ours-TL
NIST20	HCD	0.539	-
MassBank	HCD	0.551	-
GNPS	QTOF, QqQ	0.538	0.607
NIST20	QTOF, QqQ	0.558	0.648
MassBank	QTOF, QqQ, Unknow	0.567	0.617

To compare with other reported methods, we evaluate our model on [M+H]+ and [M-H]- . It shows that our deep learning model achieves the stat-of-the-art performance on large dataset. The results on NIST20 dataset are:

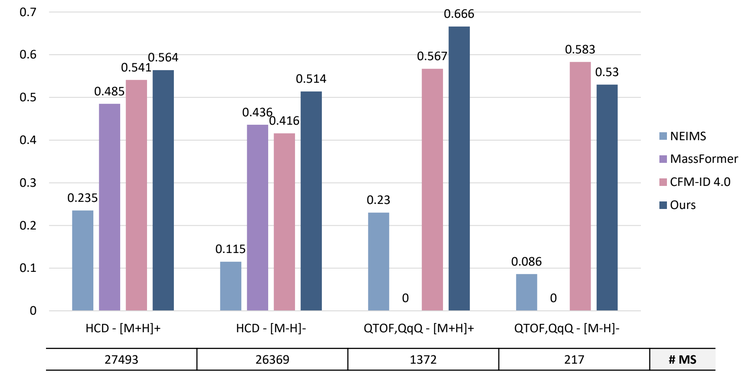

Figure 4. Comparison results on NIST20

Results – Chemical Properties Prediction

Using transfer learning the mean relative error (MAE) of retention time prediction decreases from 3.5% to 2.9%, and the MAE of collision cross section prediction decreases from 9.5% to 9.5%. It shows that the pattern learnt from MS/MS prediction can be used in other chemical properties prediction.

Conclusion

In this work, we designed a convolution operation, MolConv, which could learn the information from 3D molecular conformers. Based on MolConv, we established a deep neural network to predict MS/MS and chemical properties. Using the transfer learning technology, our model could achieve 0.549 and 0.621 cosine similarity on HCD and QTOF MS/MS, respectively.

Contact

Yuhui Hong: yuhhong@iu.edu
Haixu Tang: hatang@indiana.edu

References

1. Xiaoyu Yang, Pedatsur Neta, and Stephen E Stein. Extending a tandem mass spectral library to include ms2 spectra of fragment ions produced in-source and msn spectra. Journal of The American Society for Mass Spectrometry, 28(11):2280–2287, 2017.
2. Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. Journal of mass spectrometry, 45(7):703–714, 2010.
3. Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nature biotechnology, 34(8):828–837, 2016.