



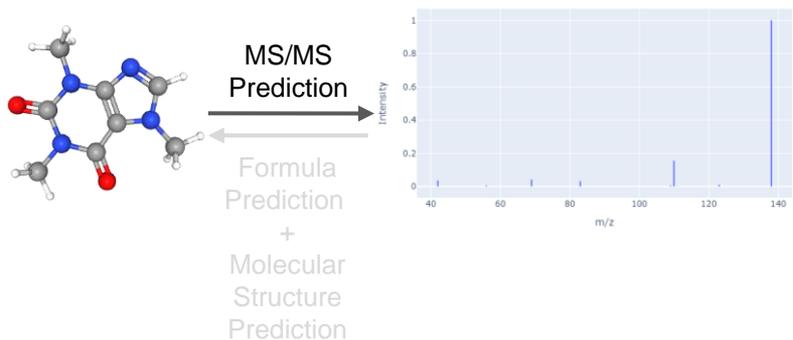
# A Machine Learning Model for Chemical Formula Prediction Using Tandem Mass Spectra of Compounds

Yuhui Hong, Haixu Tang

Luddy School of Informatics, Computing, and Engineering  
Indiana University, Bloomington, IN 47408, USA

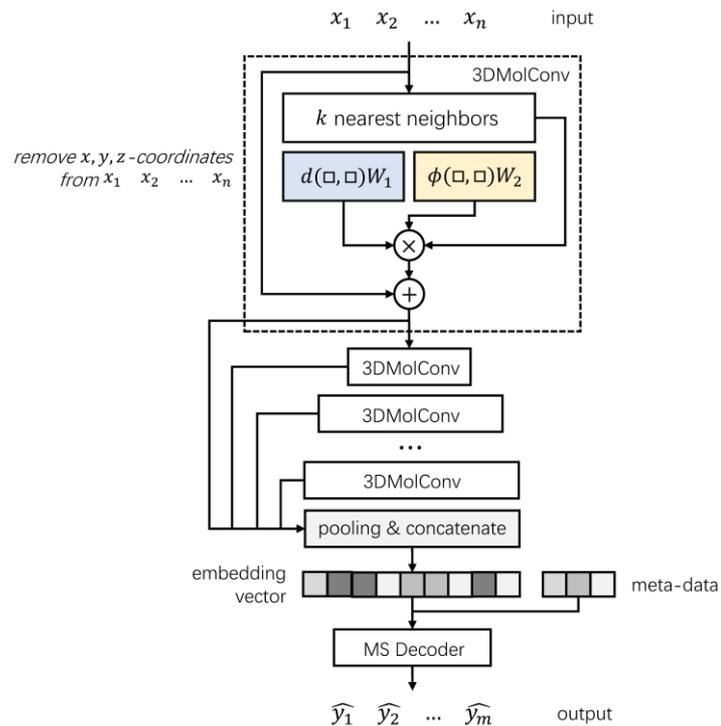
*The authors declare no competing financial interest.*

# From MS/MS Prediction to Formula Prediction



Codes are available on GitHub.

Online service are available on GNPS.

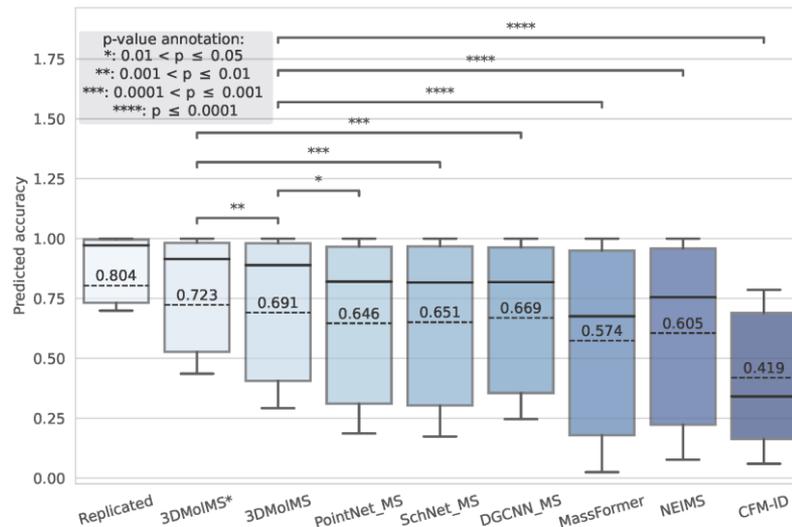


Architecture of 3DMoIMS

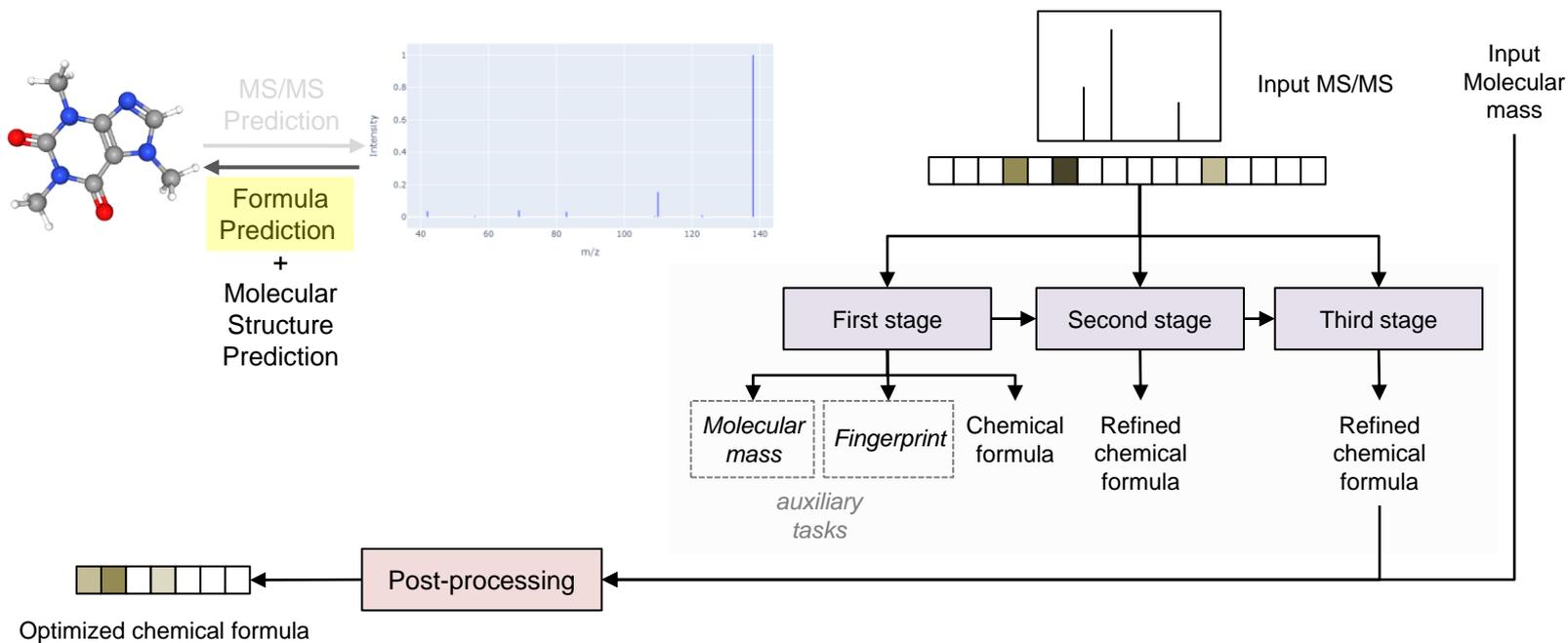
# From MS/MS Prediction to Formula Prediction

	Positive ion mode		Negative ion mode	
	# spectra	# compounds	# spectra	# compounds
NIST20	27085	2492	1749	193
Agilent PCDL	35373	11239	8362	2942
Unique	62458	13295	10111	3080

Performances of MS/MS Prediction on Agilent QTOF [M+H]<sup>+</sup>



# From MS/MS Prediction to Formula Prediction



# Related Work: SIRIUS4

- Fragmentation tree<sup>[1][2]</sup> can only process single-charged MS/MS because it relies on the neutral loss, e.g., H<sub>2</sub>O.

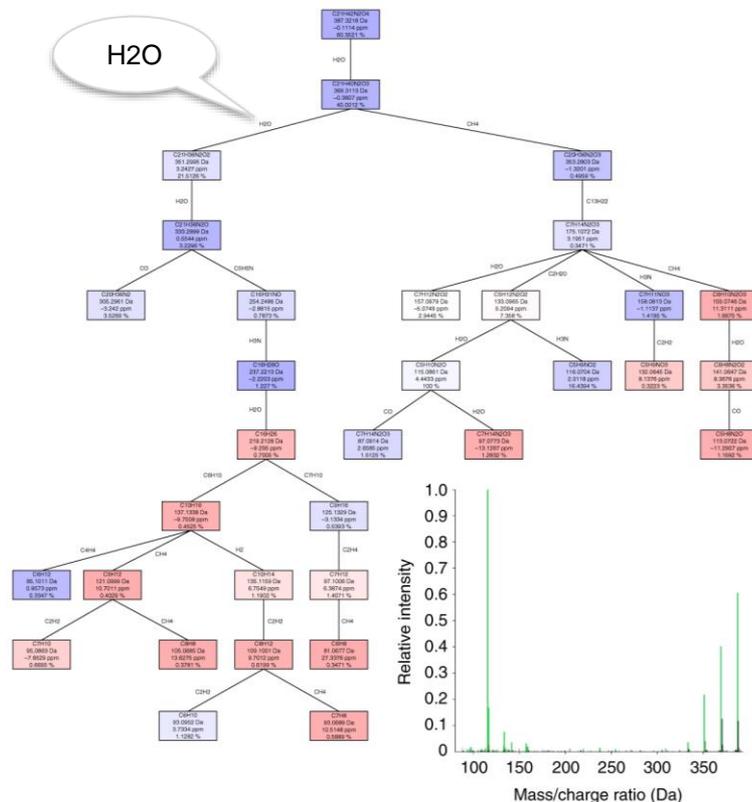
## Ion modes

Whenever SIRIUS requires the ion mode, it should be given in the following format:

```
[M+ADDUCT]+ for positive ions
[M+ADDUCT]- for negative ions
[M-ADDUCT]- for losses
[M]+ for intrinsically charged compounds
```

ADDUCT is the molecular formula of the adduct. The most common ionization modes are [M+H]<sup>+</sup>, [M+Na]<sup>+</sup>, [M-H]<sup>-</sup>, [M+Cl]<sup>-</sup>. **Currently, SIRIUS supports only single-charged compounds**, so [M+2H]<sup>2+</sup> is not valid. For intrinsic charged compounds [M]<sup>+</sup> and [M]<sup>-</sup> should be used.

- Computations are time consuming.

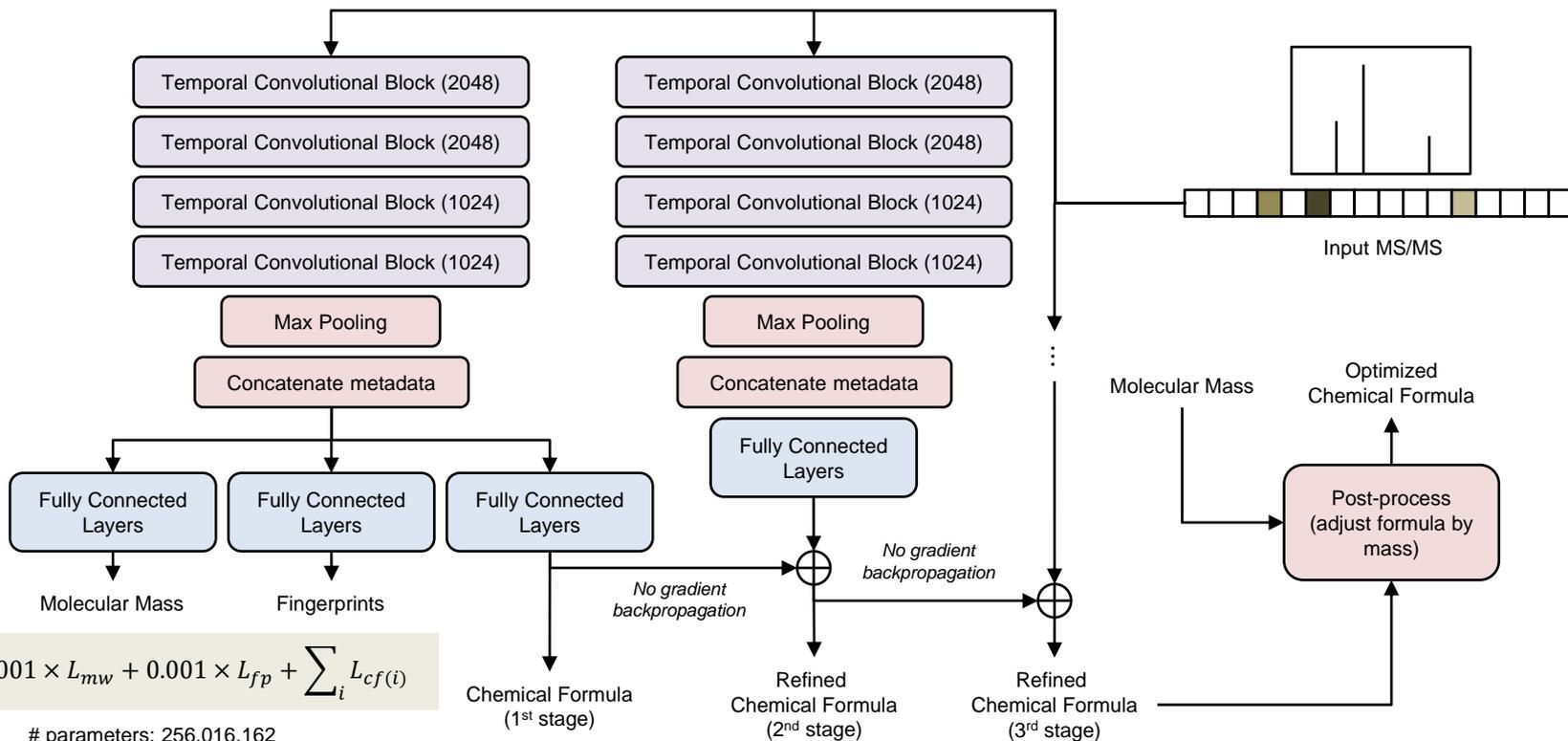


Fragmentation tree that explains the experimentally observed MS/MS fragmentation pattern of the ion with  $m/z$  387.322.

[1] Dührkop, Kai, et al. "SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information." *Nature methods* 16.4 (2019): 299-302.

[2] Rasche, Florian, et al. "Computing fragmentation trees from tandem mass spectrometry data." *Analytical Chemistry* 83.4 (2011): 1243-1251.

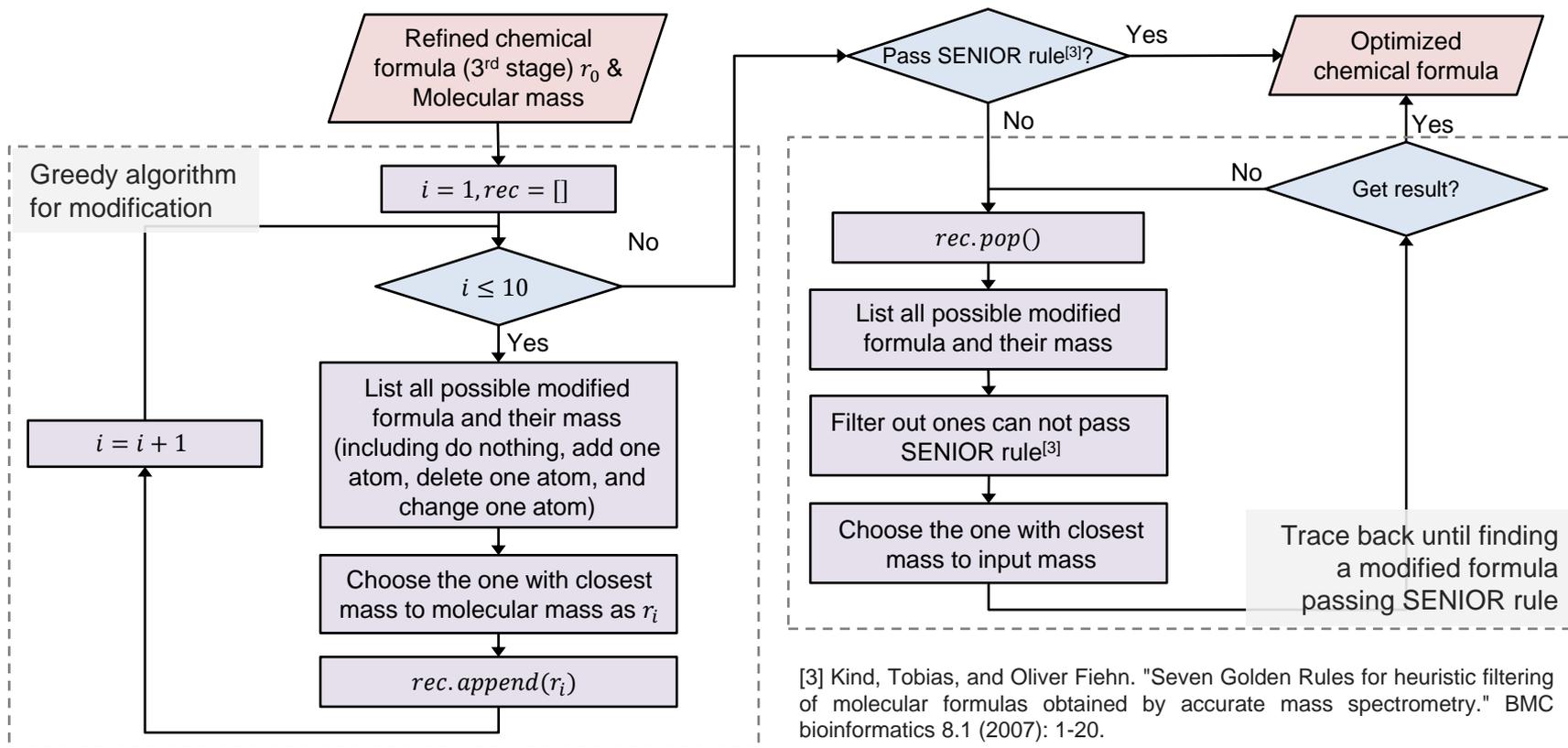
# Our Methods



$$\mathcal{L} = 0.001 \times L_{mw} + 0.001 \times L_{fp} + \sum_i L_{cf(i)}$$

# parameters: 256,016,162

# Our Methods (Post-processing)



[3] Kind, Tobias, and Oliver Fiehn. "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry." BMC bioinformatics 8.1 (2007): 1-20.

# Our Methods (SENIOR rule)

## SENIOR rule:

1. The sum of valences or the total number of atoms having odd valences is even;
2. The sum of valences is greater than or equal to twice the maximum valence;
3. The sum of valences is greater than or equal to twice the number of atoms minus 1.

e.g., C<sub>9</sub>H<sub>14</sub>O<sub>3</sub> passes SENIOR rule.

$4 \times 9 + 1 \times 14 + 2 \times 3 = 56$ ; The valences of C, H, and O are 4, 1, and 2, respectively.

$56 \geq 2 \times 4$ ;

$56 \geq 2 \times (9 + 14 + 3 - 1) = 50$ ;

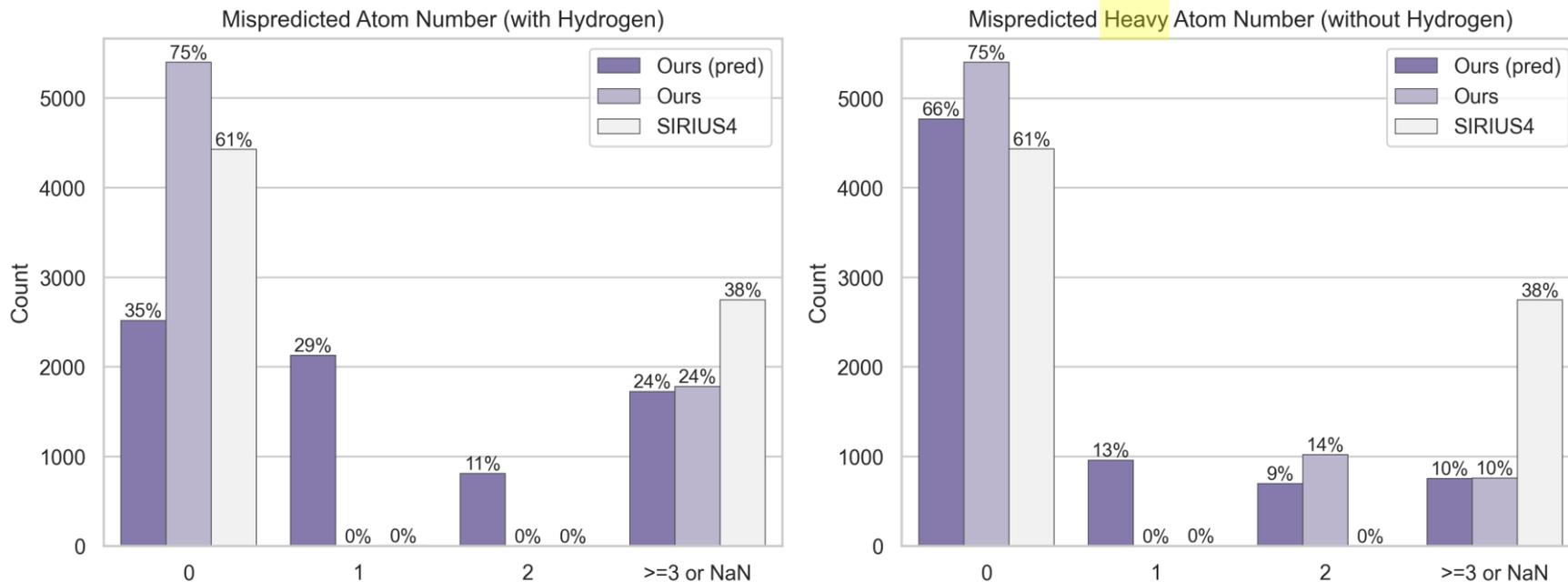
# Experiment Data Preprocessing

We collected **70,111** spectra of **14,376** compounds with masses from the Quadrupole Time-of-Flight (Q-TOF) MS/MS library of Agilent and NIST20.

The compounds are randomly split into training and test sets with a ratio of 9:1.

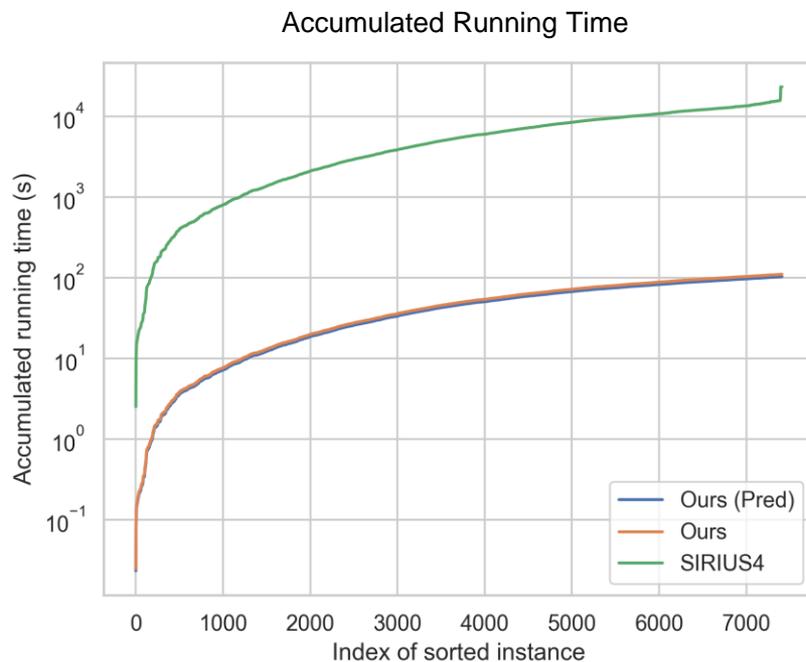
1. The spectra with less than 5 peaks are removed because they are typically unreliable;
2. The spectra with  $m/z$  greater than 1500 are discarded because only a few spectra are from such large molecules;
3. Only the spectra with the precursor types of  $M+H$  and  $M-H$  are retained;
4. Only the compounds with fewer than 300 atoms are retained because only a few compounds in the library have more than 300 atoms;
5. Only the molecules composed by the most common atoms (C, H, O, N, F, S, Cl, P, B, I and Br) are retained.

# Results on Single Charged MS/MS (accuracy)



Ours (pred) denotes the results from machine learning model without post-processing.

# Results on Single Charged MS/MS (speed)



The compounds have been arranged in ascending order based on their mass.

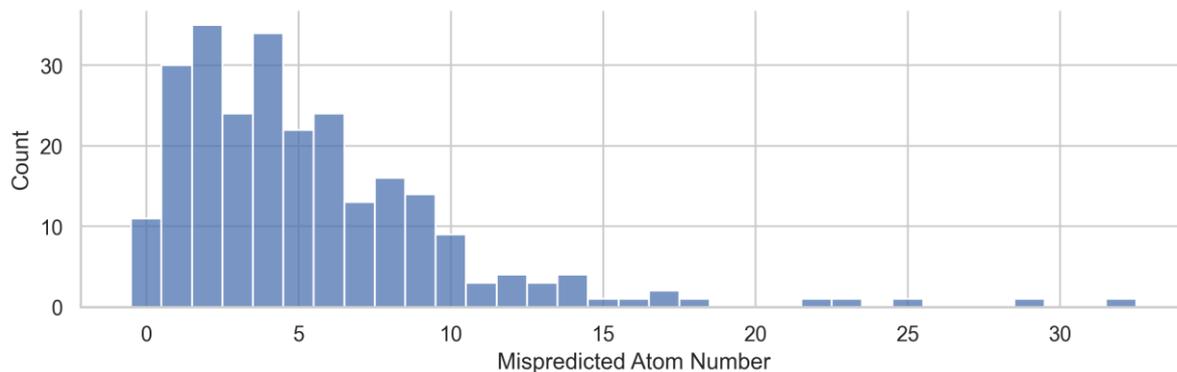
# Results on Double Charged MS/MS

The double charged MS/MS from Agilent PCDL and NIST20 are gathered as an additional test set.

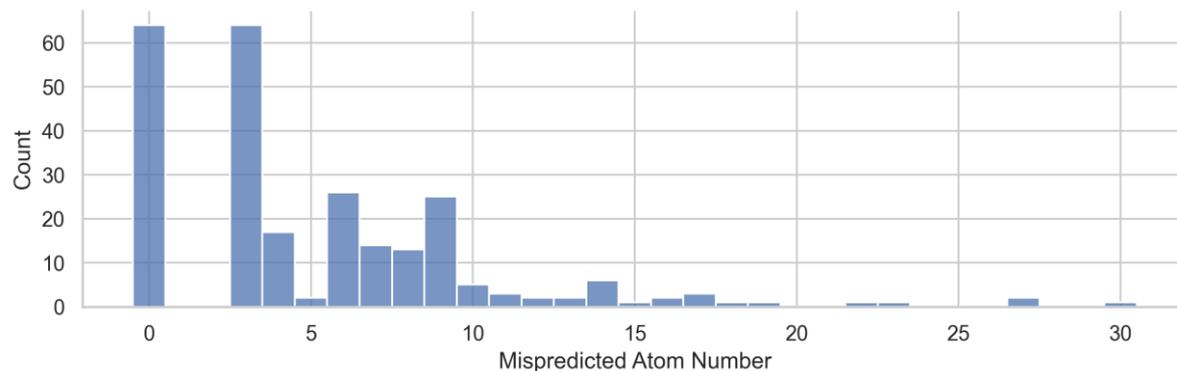
The model trained on single charged MS/MS are applied directly to the double charged MS/MS.

# Spectra	# Compound
256	197

Performance of Ours (pred) on  $[M+2H]^{2+}$



Performance of Ours on  $[M+2H]^{2+}$



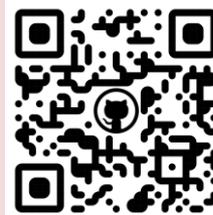
# Takeaways

- We presented a deep learning model with post-processing for chemical formula prediction achieving state-of-the-art performance on QTOF MS/MS.
- Our model is efficient, and it can be extended to MS/MS with different adducts.

Thank you!

Please find the codes of  
3DMolMS on GitHub.

We will release the codes  
for chemical formula  
prediction soon!



## 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. We are indebted to Dr. Mingxun Wang for his invaluable advice and his effort to build the online service of 3DMoIMS.

## Funding

We acknowledge the Center for Bioanalytical Metrology (CBM), an NSF Industry-University Cooperative Research Center, for providing funding under grant NSF IIP-1916645. This work was also partially supported by National Science Foundation grant DBI-2011271.