Accurate MS/MS Spectral Prediction with CFM-ID 4.0

Fei Wang[1,3], Siyang Tian[1,2], Jaanus Liggand[2,4], David Arndt[2], Russell Greiner[1,3], David S. Wishart[1,2]

[1]Department of Computing Science, [2]Department of Biological Science, University of Alberta, Edmonton, Alberta, Canada
[3]Alberta Machine Intelligence Institute (AMII), Edmonton, Alberta, Canada
[4]Institute of Chemistry, University of Tartu, Tartu, Estonia

TMIC
The Metabolomics Innovation Centre

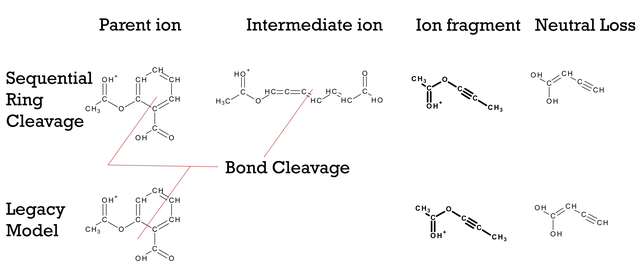
Introduction

CFM-ID is a software package that uses machine learning to accurately predict the Electrospray Ionization Tandem Mass Spectrometry (ESI-MS/MS) spectra for organic compounds. It can perform the following three tasks: (1) **Predict the spectra** for a given chemical structure, (2) **annotate the peaks** in a set of given spectra of a known chemical structure, and (3) **classify the structure** for a target spectrum. CFM-ID 3.0 is a state-of-the-art program for ESI-MS/MS spectral prediction, as well as spectrum-to-compound classification. In this work, we are introducing the significantly improved version, CFM-ID 4.0. It is freely available as both, a web version (cfmid4.wishartlab.com), and a downloadable software package.

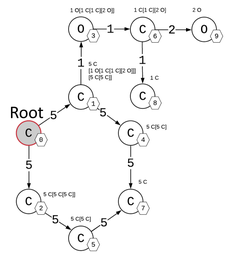
CFM-ID 4.0 Improvements

- Learned parameters via Deep Neural Networks
- Learned parameters from molecular topology
- Improved ring cleavage model
- Expanded training set by 3x (QToF)
- Added rule-based schema for
  - Acylcarnitines
  - Acylcholines
  - Polyphenols
  - Flavonols

Ring Cleavage Modeling



Graph Based Feature Representation



Rooted Graph
same graph = same indexing

Create Labels for node

Bond Type
1 – single
2 – double
5 – aromatic

Atom Type

Children labels sorted alphanumerically

Expanding Rule Based Predictors

Compound class	Number of Covered Rules	Covered Adduct Types
acylcarnitines	9	[M]+
acylcholines	11	[M]+
flavonols	47	$[M+H]^+$, $[M-H]^-$
flavones	20	$[M+H]^+$, $[M-H]^-$
flavanones	15	$[M+H]^+$, $[M-H]^-$
flavonoid-3-O-glycosides	36	$[M+H]^+$, $[M-H]^-$
flavonoid-7-O-glycosides	63	$[M+H]^+$, $[M-H]^-$
flavonoid-7-O-glucuronides	12	$[M+H]^+$, $[M-H]^-$
4'-O-methylated flavonoids	53	$[M+H]^+$, $[M-H]^-$
7-O-methylated flavonoids	36	$[M+H]^+$, $[M-H]^-$
3'-O-methylated flavonoids	11	$[M-H]^-$

Predicted Spectra

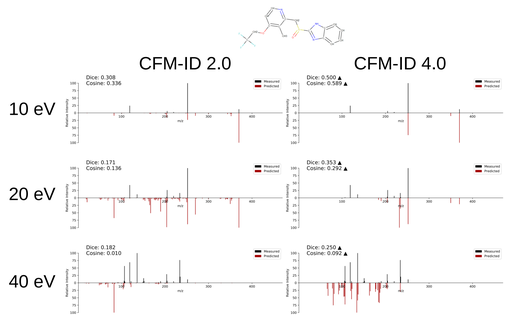


Figure1. CFM-ID 4 Machine Learned Prediction

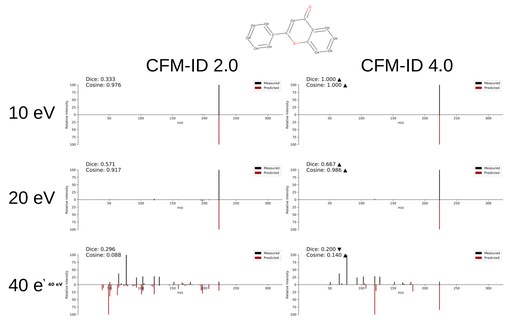
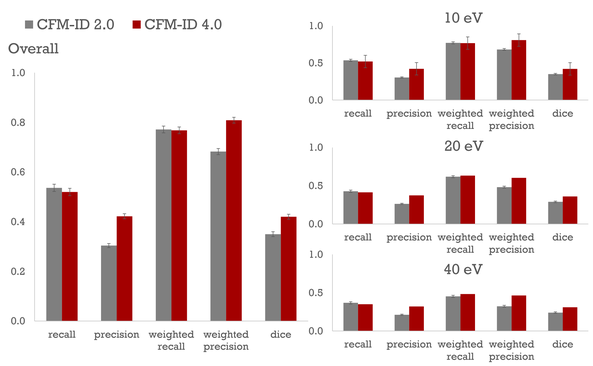


Figure2. CFM-ID 4 Rule Based Prediction

Results

Spectral Prediction Performance

10-Fold Cross Validation Results on Metlin 2015 $[M+H]^+$ Set



Spectral Classification Performance CASMI 2016 (Category 3)

208 compounds

	TOP 1	TOP 3	TOP 10
CFM-ID 2.0 Alone	120	160	182
CFM-ID 4.0 Alone	**147**	**178**	**203**
CFM-ID 3.0 + Expt. Spectra + Meta data	149	194	204
CFM-ID 4.0 + Expt. Spectra + Meta data	**165**	**197**	**207**
MS-Finder (post-hoc)	146	162	174

Conclusions

In this work, we introduced a novel tensor representation for describing chemical structures and used it to extend the capabilities of existing CFM-ID machine learning methods in ESI-MS/MS spectral prediction tasks. Alongside machine learning based improvements, we proposed new rule-based methods to further enhance the CFM-ID 4.0 system's predictive ability for specific classes of chemicals, where the machine learning model suffers. The *in-silico* spectra prediction performance of these novel methods was examined against empirical results on multiple ESI-MS/MS data sets, encompassing a wide range of chemical classes, in both positive and negative ionization modes. While still imperfect, our proposed method outperformed the legacy CFM-ID model by a significant margin across all data sets. In addition, we demonstrated CFM-ID 4.0's *in-silico* compound identification ability via the CASMI 2016 competition (category 3), where CFM-ID 4.0 achieved better identification results then all existing approaches.

CFM-ID Webservice

Web Service:
http://cfmid4.wishartlab.com
Docker Hub Access:
https://hub.docker.com/r/wishartlab/cfmid
Source Code:
https://bitbucket.org/wishartlab/cfm-id-code/
https://bitbucket.org/wishartlab/msrb-fragmenter

Acknowledgements

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
GenomeCanada
GenomeAlberta
GenomeBritishColumbia
UNIVERSITY OF ALBERTA
University of Victoria
McGill
McMaster University
NIH National Institutes of Health
CIHR IRSC
NSERC CRSNG
amii
compute canada | calcul canada