quantemol

Machine learning estimator for electron impact ionization fragmentation patterns

Dr. Kateryna Lemishko

May 10, 2024

Overview

Chemistry sets used in plasma modeling contain many species involved in a significant number of reactions

Problem: limited availability of input data for plasma models (cross sections and rate coefficients)

Experiments and simulations

Experiments and first principles calculations provide relatively accurate information.

They are time-consuming or not always possible.

Data-driven approach

Machine learning provides inexpensive, reasonably accurate and fast estimations.

Low accuracy in comparison to experiment or simulations.

Accuracy depends on the volume and quality of the input data.

Machine learning in plasma and chemistry

quantemol

Examples of implementation of machine learning in plasma and chemistry:

Computational chemistry:

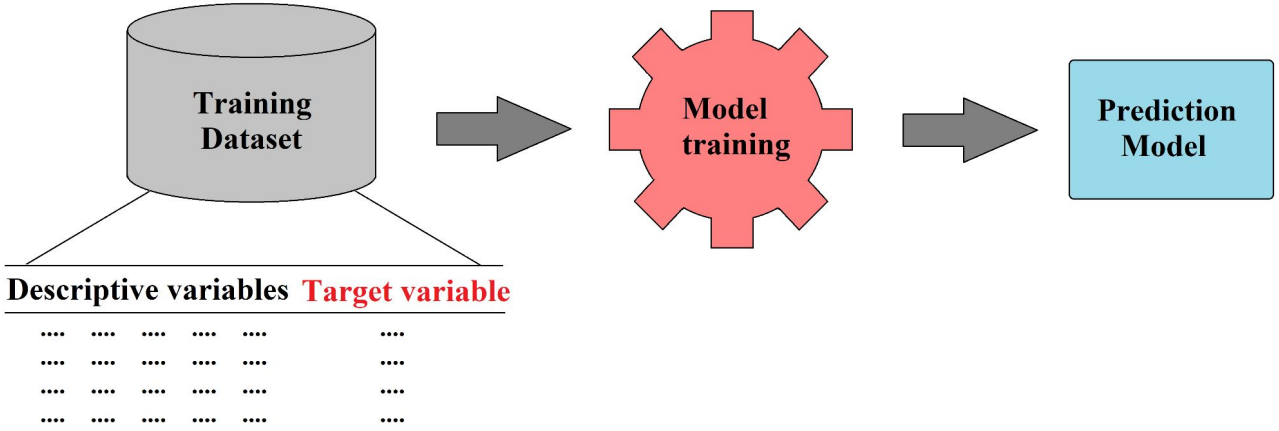
reaction yields,
molecular structure,
partial charges,
physicochemical properties, e.g. toxicity, bioactivity, solubility,
melting points, hydration free energies, atomisation energies, dipole moments,
etc.

Plasma modelling:

total electron impact ionization cross sections,
particle properties in plasma spraying,
sputtered particle distributions,
characteristics of plasma-deposited films,
plasma etch data, etc.

Supervised machine learning

Learning a model from known data



Supervised machine learning models use known data to identify and learn **patterns** and **relationships** between a set of descriptive variables and a target variable.

The process of learning these patterns in data is called **model training**.

Prediction of electron impact ionization fragmentation patterns

quantemol

In plasma physics, it is often desirable to know electron impact ionization fragmentation patterns.

Partial ionization cross sections can be calculated using branching ratios for the production of fragments which can be obtained from mass spectrometry data.

Problem:

For many compounds experimental mass spectrometry data is unavailable.

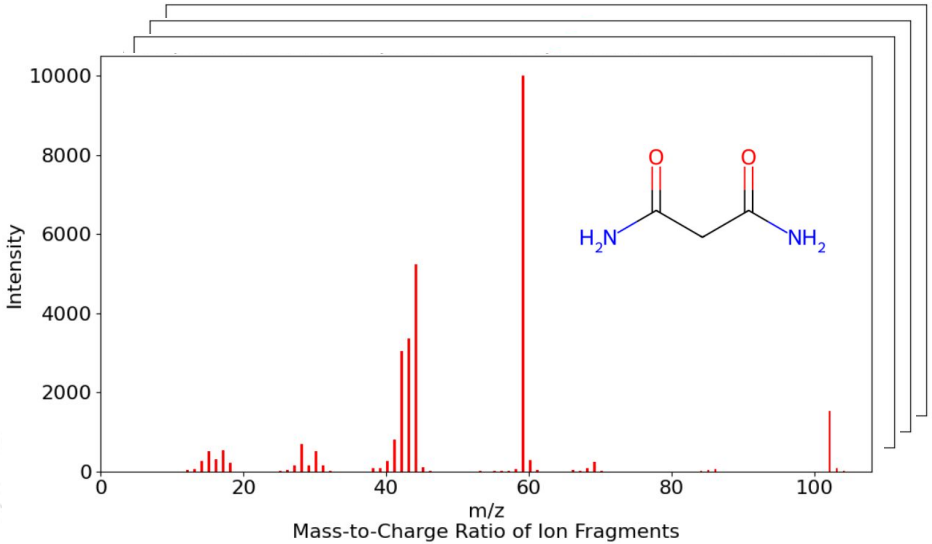
Solution: machine learning

We can develop an algorithm that learns from existing data to predict mass spectra that can be used for inference of electron impact ionization fragmentation patterns.

Prediction of ionization mass spectra

webbook.nist.gov:

Electron ionization mass spectra
for 6500+ compounds



pubchem.ncbi.nlm.nih.gov:

SMILES (Simplified molecular-input line-entry system)

SMILES is a way to encode a structural information of a molecule using a linear string of characters)

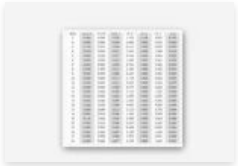
compound_name	SMILES
(+)-p-Bromo-.alpha.-phenethylamine	CC(C1=CC=C(C=C1)Br)N
(1-Hydroxy-2,2,2-trichloroethyl)formamide	C(=O)NC(C(Cl)(Cl)Cl)O
(1R,2R)-(+)-1-Phenylpropylene oxide	CC1C(O1)C2=CC=CC=C2
(2-Methoxyphenyl)acetonitrile	COC1=CC=CC=C1CC#N
(2-Thienylthio)acetone	CC(=O)CSC1=CC=CS1
...	...

Prediction of ionization mass spectra

quantemol

Input features

SMILES



Numeric features

Molecular features:

- molecular mass
- total N atoms
- total N bonds

Bond features:

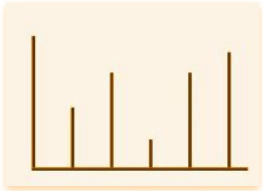
- bond types (single,double,..)
- bond chirality (e,z,unknown,none)
- bonds in rings
- conjugated bonds

Atomic features:

- atom types (C, O, N, S ...)
- atom valence
- number of neighbours
- atom hybridization
- atom aromaticity

Machine learning model training

Target

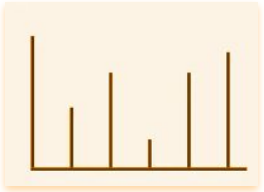


Mass spectrum vectors
y=(y1, y2 , ... yn)

Model evaluation metric

Evaluation metric: cosine similarity between normalized real and predicted mass spectra

$$\text{cosine\_similarity}(y_{\text{real}}, y_{\text{pred}}) = \frac{\sum_{i=0}^{m_{\max}-1} y_{\text{real}}[i] \cdot y_{\text{pred}}[i]}{\sqrt{\sum_{i=0}^{m_{\max}-1} (y_{\text{real}}[i])^2} \cdot \sqrt{\sum_{i=0}^{m_{\max}-1} (y_{\text{pred}}[i])^2}}$$



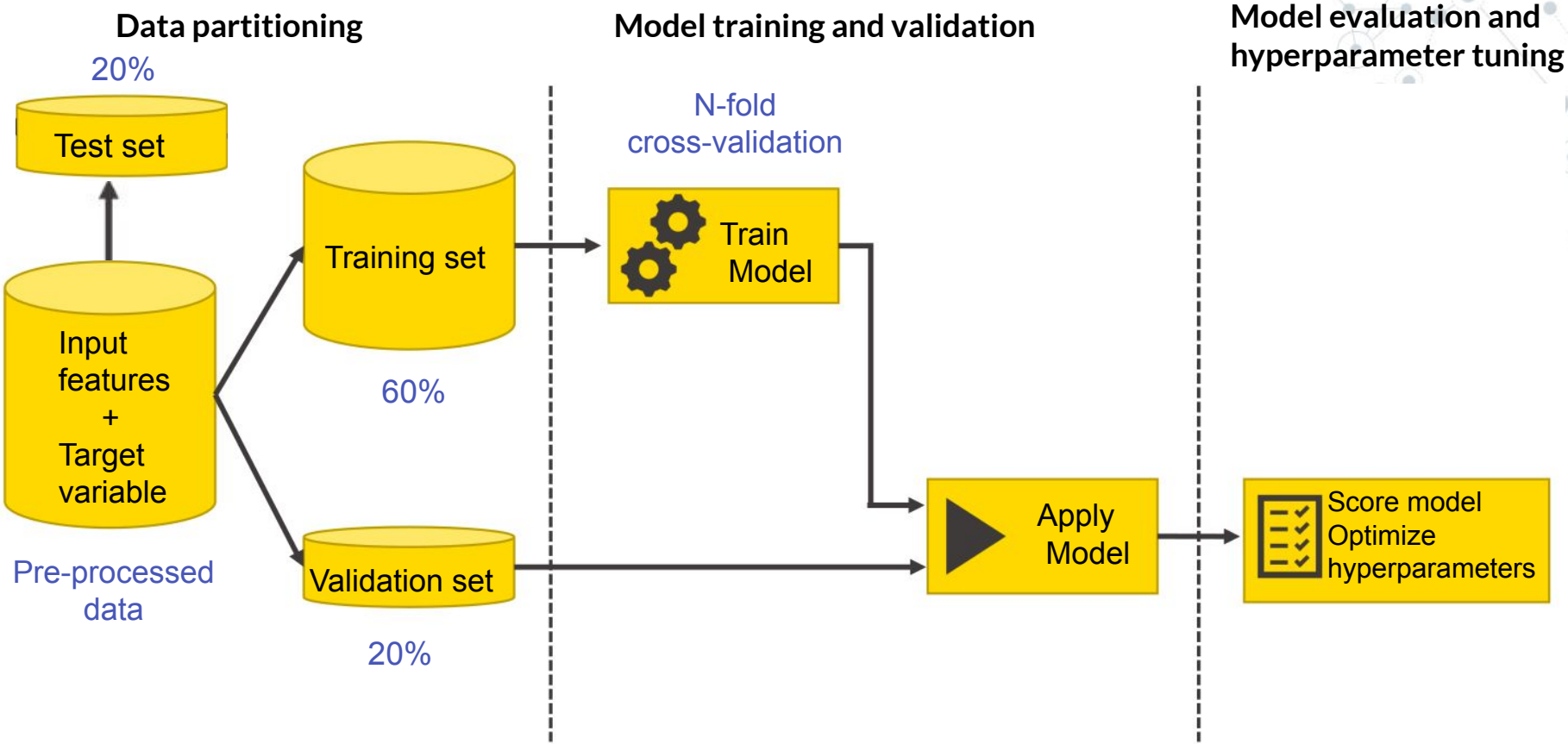
m_{max} - max m/z

y_{real} - real intensity vector

y_{pred} - predicted intensity vector

Mass spectrum vectors $y=(y_0, y_1, \ldots y_{max-1})$

Machine learning model selection workflow

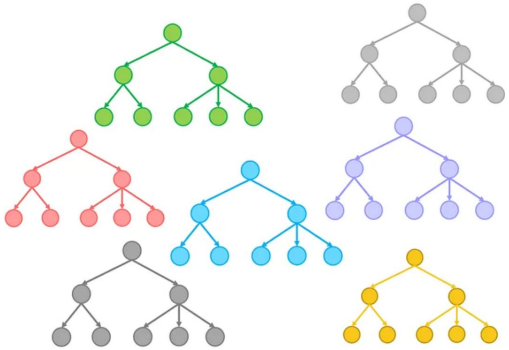


Final model architecture

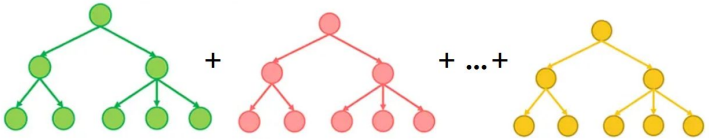
quantemol

Best performing individual algorithms:

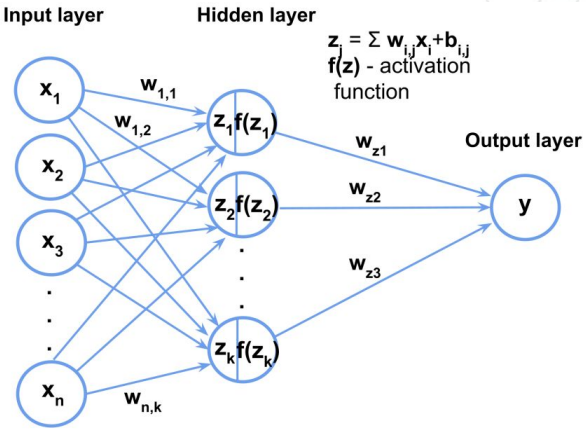
Random Forest



Extreme Gradient Boosting (XGBoost)



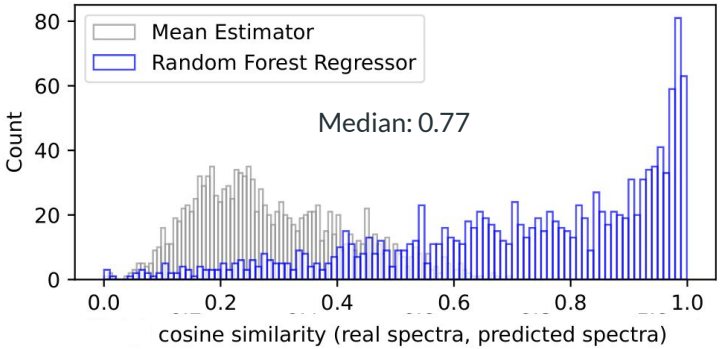
Multilayer perceptron

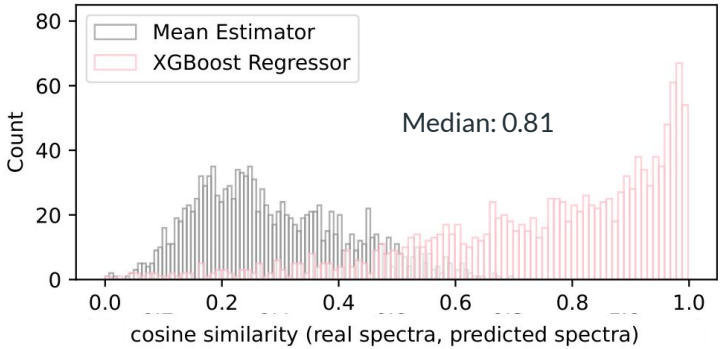


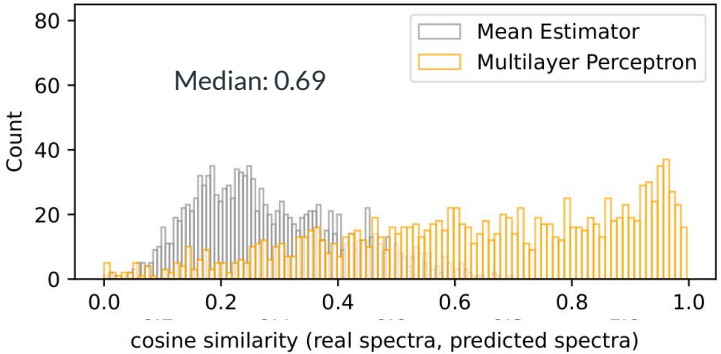
Final algorithm: a voting regressor combining different machine learning models:

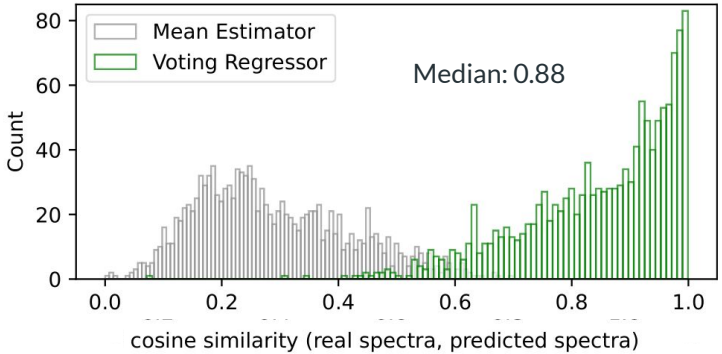
$$f(\mathbf{X}) = \omega_1 \mathbf{Random\_Forest} + \omega_2 \mathbf{XGBoost} + \omega_3 \mathbf{MLP}$$

Prediction of ionization mass spectra







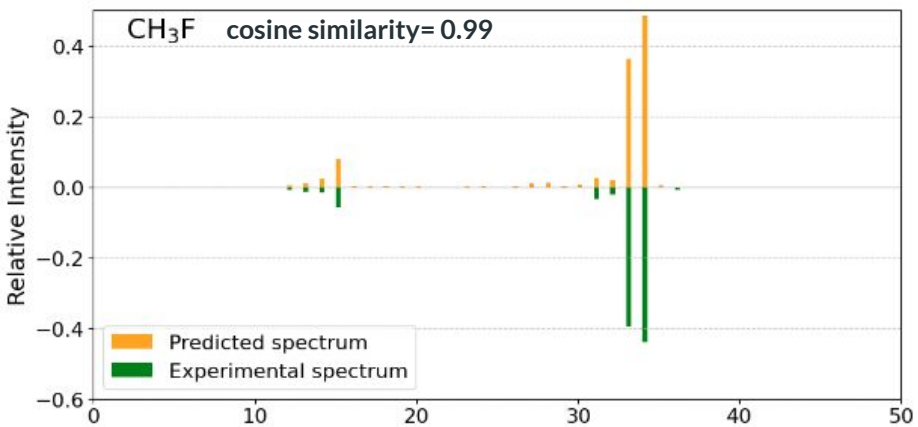


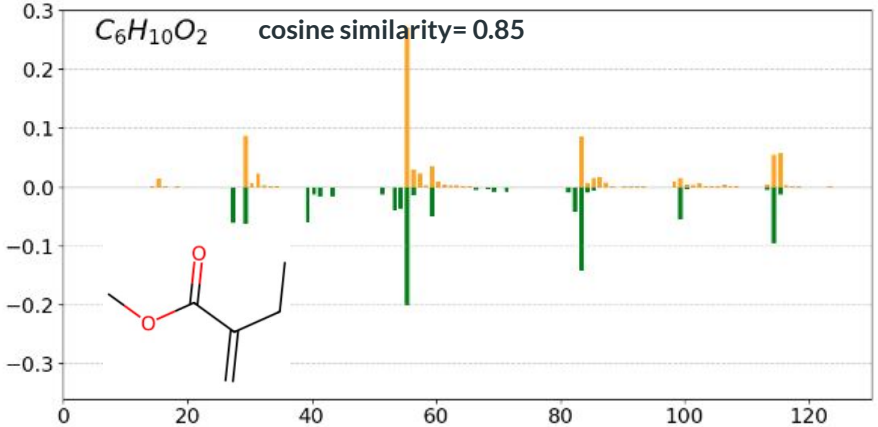
The mean estimator predicts the average mass spectrum across the training set for all test cases.

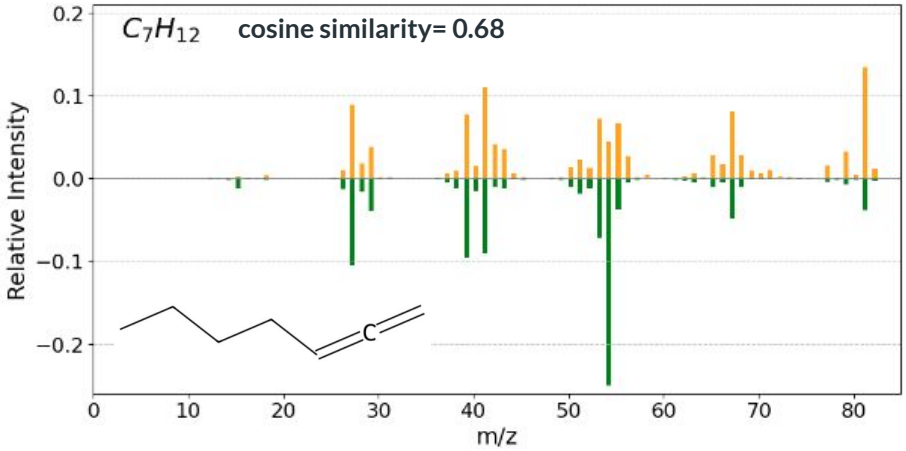
For ~70% of compounds in the test set cosine similarity is greater than 0.8

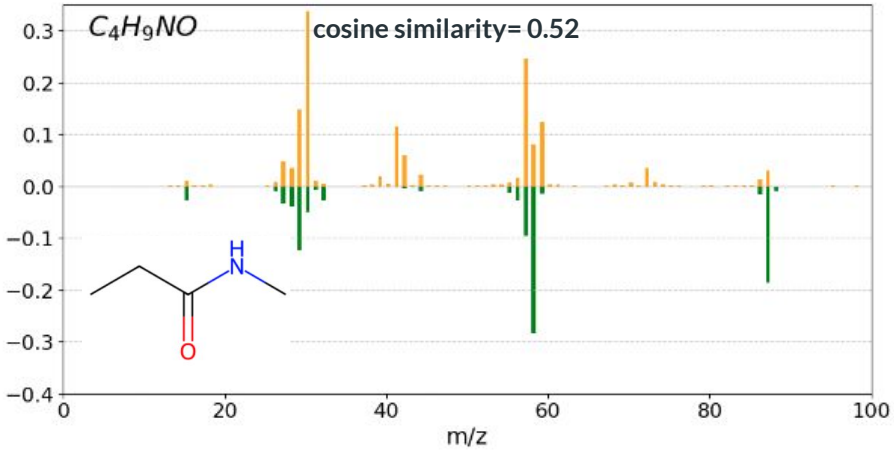
Prediction of ionization mass spectra

quantemol









Mass spectra estimator page

https://www.quantemoldb.com/ms/mass-spectrum-estimator/

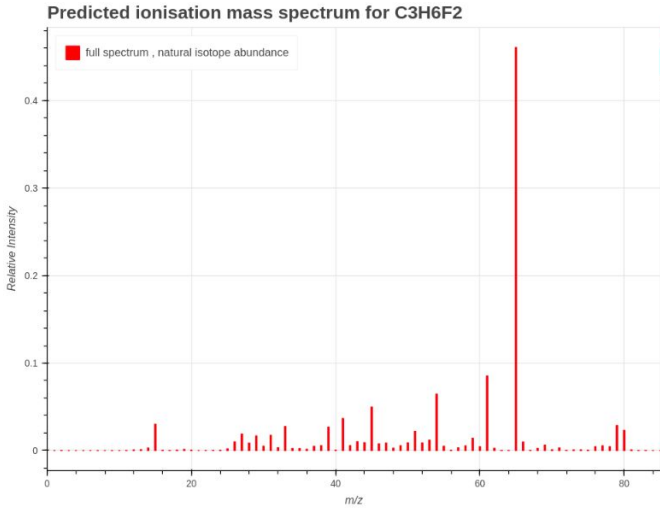
Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94.
2. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1994.
3. Linstrom, P. J.; Mallard, W. G. The NIST Chemistry WebBook: A chemical data resource on the internet. J. Chem. Eng. Data 2001,

Please enter the chemical formula of a species for which you want to obtain a mass spectrum. P

- Elements are case-sensitive, e.g. CHF3, C3H8, etc.
- Currently, the mass spectrum estimator accepts open-shell species with 1 unpaired electron.

Chemical Formula: RESET

SUBMIT



Fragment	Fraction
C2F2H3+	0.515
C3FH6+	0.096
C2H2F+	0.072
C3H5+	0.056
H3C+	0.041
C3F2H5+	0.034
C3H3+	0.032
C3F2H6+	0.031
C2H3+	0.03
FC+	0.026
C3FH4+	0.025
FC2+	0.021
H2C2+	0.02

Conclusions

- Machine learning offers an inexpensive alternative to experiments and first principles calculations, providing reasonable estimations.

- We have developed a machine learning–based algorithm for the fast prediction of mass spectra. Predicted mass spectra can be used for inference of partial electron impact ionization cross sections