Challenges in Mass Spectral Prediction

Uhlir Manuel

TBI - University of Vienna

February 14, 2024

Problem Statement

Vast number of Unquantifiable or Unknown Compounds:

81% of E.coli metabolites

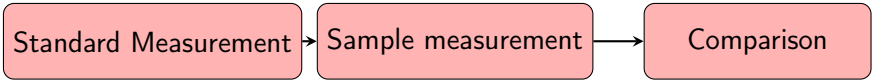
81% of Yeast metabolites

93% of Human metabolites

97% of Plant metabolites

Source: https://doi.org/10.1038/s41592-019-0344-8

How do you identify your compounds?



Massively limited by:

1. Availibility of the Standards
2. Sequential nature of Mass Spectrometry experiments
3. Cost

So what else can we do?

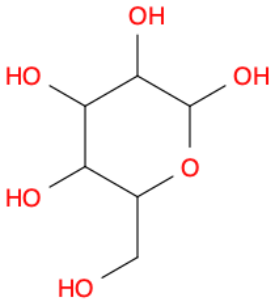
Match against databases

Problems

1. Availibility of spectra
2. Differences between machines
3. Trust into other peoples data or interpretation

Predicting Mass Spectra

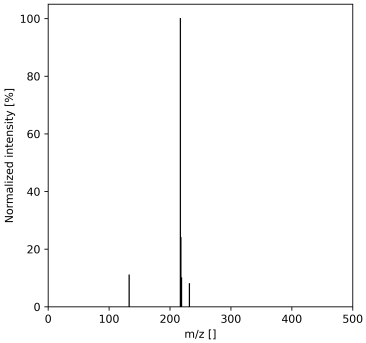
Molecular structure



Source: Openbabel

Predict

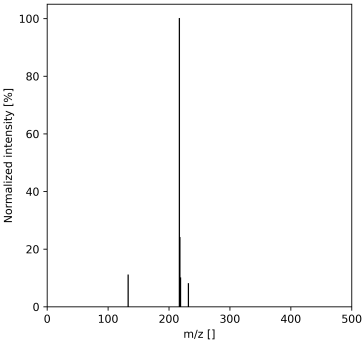
Tandem Mass Spectrum



Source: Chemspider, Matplotlib

Identify compounds

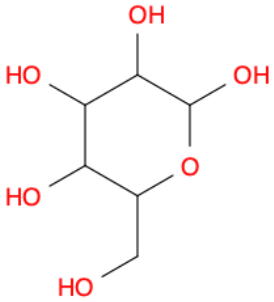
Tandem Mass Spectrum



Source: Chemspider, Matplotlib

Identify

Molecular structure

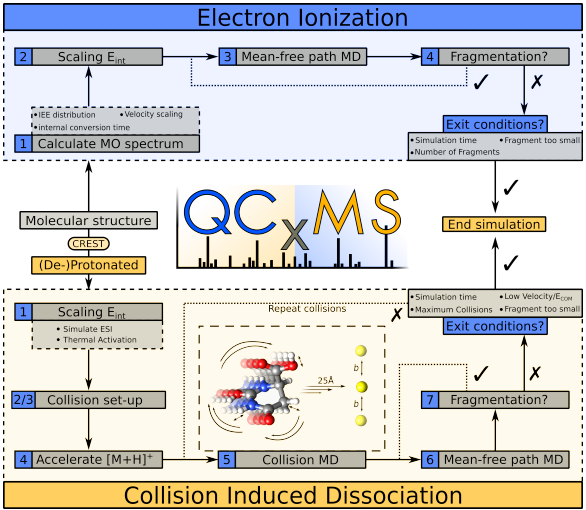


Source: Openbabel

Whats out there?

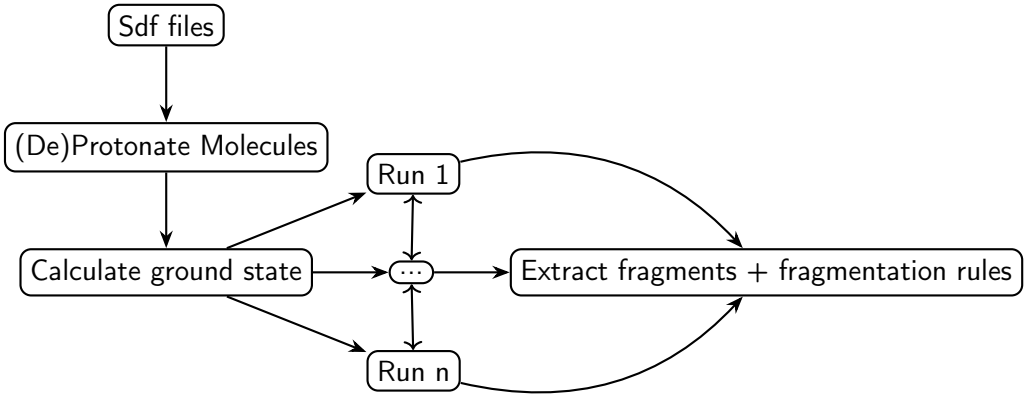
1. MetFrag
    1. Combinatorial expansion
    2. Filtering based on bond and fragment stabilities
    3. Additionally some well known rules are used (Inductive Cleavage)
2. CFM-ID (Competitive Fragmentation modelling)
    1. Machine learning model
    2. Rule based fragmentation for some classes
3. SIRIUS
    1. Build fragmentation tree
    2. Extracts a fingerprint
    3. Query CSI-FingerID with Fingerprint

Our approach



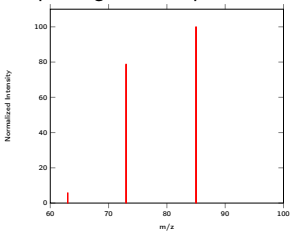
Source: QCxMS Documentation

Workflow

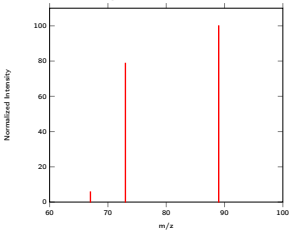

Sdf files
(De)Protonate Molecules
Run 1
Calculate ground state
...
Extract fragments + fragmentation rules
Run n

Comparing Mass Spectra

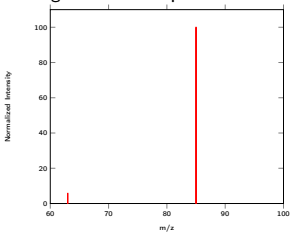
Unequal lengths of the spectra



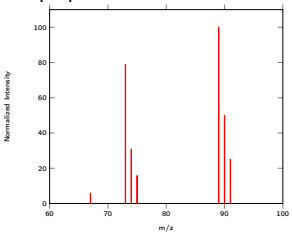
Shifts in the spectra



Missing or additional peaks



Isotopic patterns

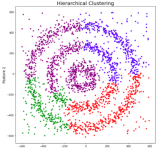


Aligning Mass Spectra - Algorithms

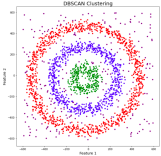
- Binning
  - Static - (Shift by 0.5)
  - Dynamic - Density based
  - Peak detection and width binning
- Clustering
  - k-means
  - Hierarchical
  - DBSCAN - Density Based Spatial Clustering of Applications with Noise
- Dynamic Time Warping

Clustering methods

- k-means
  - Number of clusters is determined by the user ×
- Hierarchical Clustering
  - Number of clusters is determined by the user ×



- DBSCAN
  - Density based
  - Number of clusters determined automatically ✓
  - Noisy data is ignored ✓
  - You still need to determine the parameters ϵ and n ×



Metrics

Intensity insensitiv

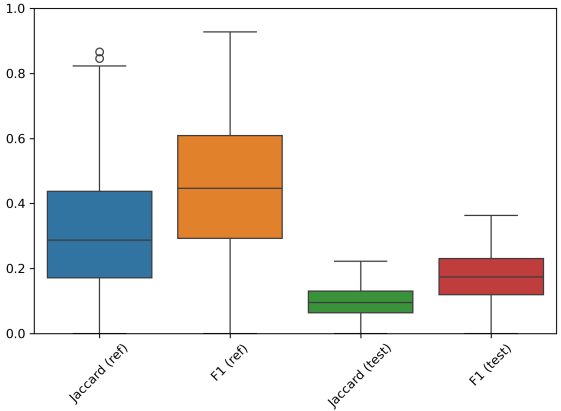
1. F1 Score
2. Jaccard index

Intensity sensitiv

1. Cosine similarity
2. Manhattan distance
3. Euclidian distance

Initial results

15 Substances (Sugars),
120 Reference Spectra from MoNA (MassBank of North America)



Next steps

1. Calibrate simulation parameters
2. Automatize choice of mode based on both pKa and availabe spectra
3. Aquire more mass spec data
4. Try different quantum chemical methods (e.g. DFT)s

Thank you for your attention!