United States
Environmental
Protection Agency

EPA's Non-Targeted Analysis WebApp: Combining Tentative Structure Identification With Risk Prioritization

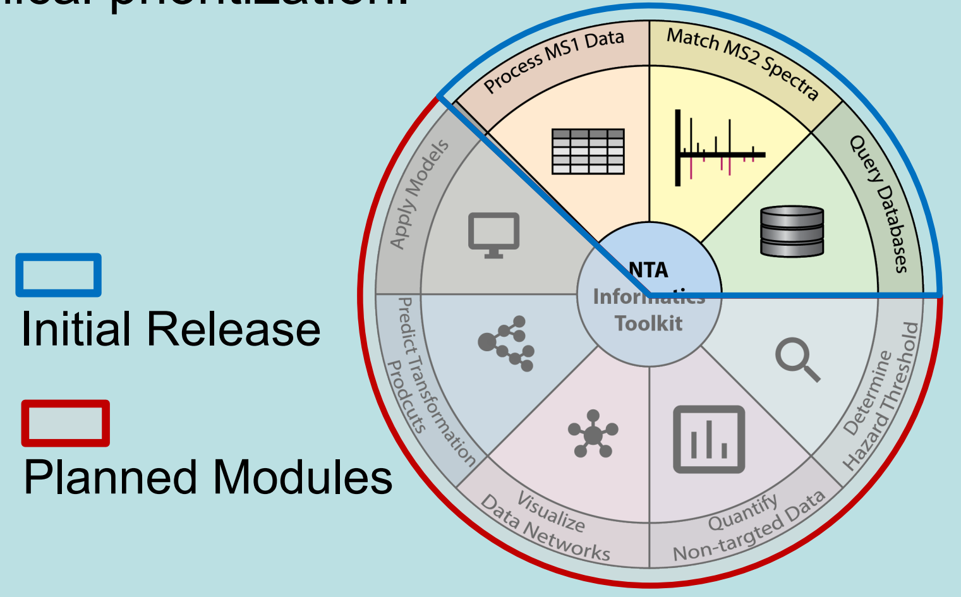
Matthew W. Boyce[1,a,b], Alex Chao[b], Jeffrey M. Minucci[c], S. Thomas Purucker[b], Antony J. Williams[b], Jon R. Sobus[b]
ORCID: [1]https://orcid.org/0000-0002-3794-1678
[a]Oak Ridge Associated University, Oak Ridge, TN 378731
[b]Center for Computational Toxicology and Exposure, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA
[c]Center for Public Health and Environmental Assessment, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

The EPA's NTA Informatics Toolkit

As applications for Non-targeted Analysis (NTA) continue to grow, so too does the demand for reproducible and transparent methods for handling non-targeted data. The EPA's NTA Informatics Toolkit is being developed as a modular resource to help meet this need and provide users a standardized method for interpreting non-targeted data through a web browser. The web application accepts either peak lists derived from MS^1 or MS^2 data and helps streamline analysis by:

(1) processing the data and performing quality control (QC) checks,
(2) identifying candidate structures using EPA's curated databases
(3) providing meta-data to aid in chemical prioritization.

The MS^1 and MS^2 data processing modules outlined in this presentation will be part of the initial release of the Toolkit. Future updates will aim to expand the capabilities of the Toolkit and transform it into a versatile platform for the analysis and interpretation of non-targeted data.

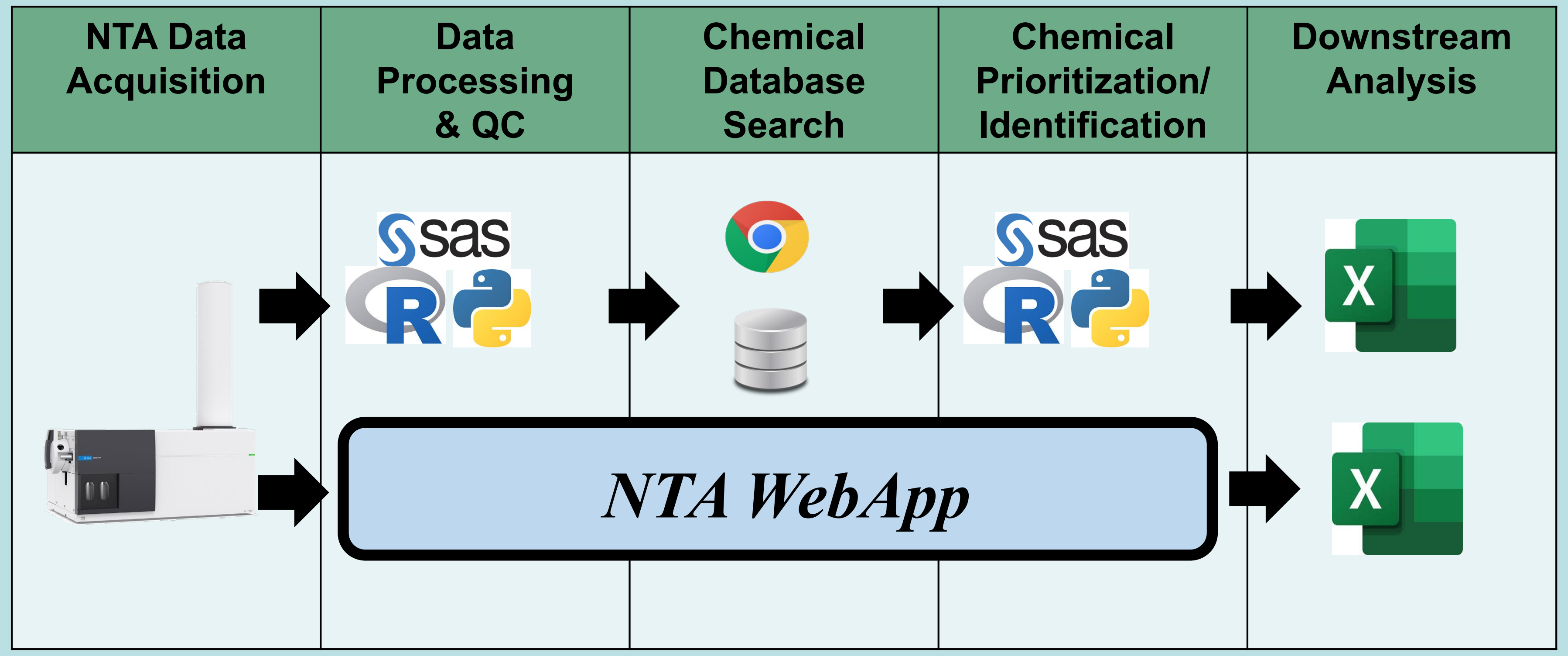


Streamline Non-Targeted Analysis

Typical non-targeted analysis workflows require multiple processing steps split between various software. The MS^1 and MS^2 modules outlined in this presentation aim to provide a single platform that:

- ▪***Standardizes analysis of non-targeted data*:** Single web-accessible point for processing NTA data
- ▪***Reduces the number of processing steps:*** Once submitted, data are carried through the workflow
- ▪***Documents all major processes:*** Full workflow tracking for reproducibility and reporting (Input files, processing / search parameters, output results, QC results)

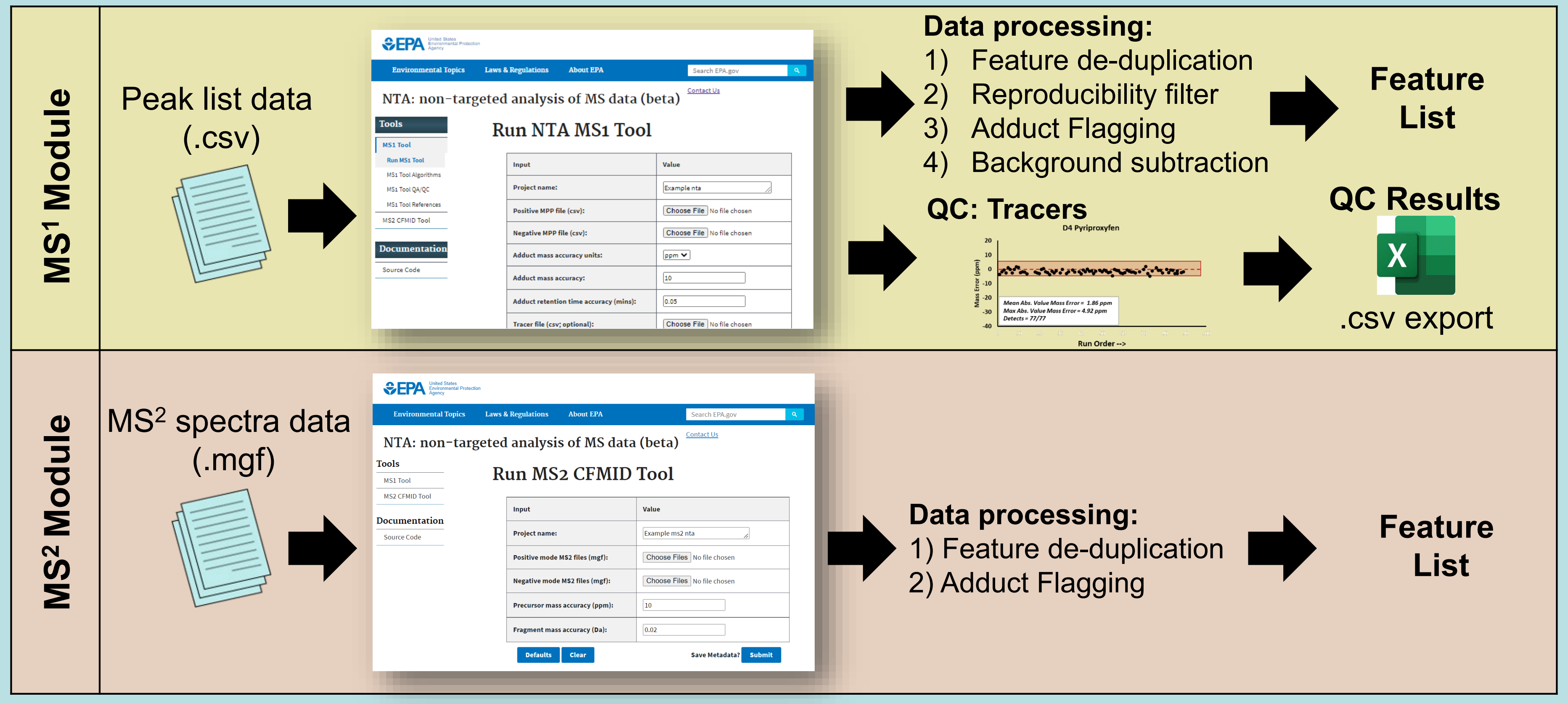
Comparison of a typical NTA workflow to the EPA's NTA WebApp



NTA Data Acquisition	Data Processing & QC	Chemical Database Search	Chemical Prioritization/ Identification	Downstream Analysis

Data Processing and QC

Users can easily upload data and start the analysis using two modules:



MS^1 Module

Peak list data (.csv)

EPA United States Environmental Protection Agency
Environmental Topics | Laws & Regulations | About EPA | Search EPA.gov
NTA: non-targeted analysis of MS data (beta) Contact Us
Tools: MS1 Tool, Run MS1 Tool, MS1 Tool Algorithms, MS1 Tool QA/QC, MS1 Tool References, MS2 CFMID Tool
Documentation: Source Code

Run NTA MS1 Tool

Input	Value
Project name:	Example nta
Positive MPP file (csv):	Choose File No file chosen
Negative MPP file (csv):	Choose File No file chosen
Adduct mass accuracy units:	ppm
Adduct mass accuracy:	10
Adduct retention time accuracy (mins):	0.05
Tracer file (csv; optional):	Choose File No file chosen

Data processing:
1) Feature de-duplication
2) Reproducibility filter
3) Adduct Flagging
4) Background subtraction

Feature List

QC: Tracers

QC Results

.csv export

MS^2 Module

MS^2 spectra data (.mgf)

EPA United States Environmental Protection Agency
Environmental Topics | Laws & Regulations | About EPA | Search EPA.gov
NTA: non-targeted analysis of MS data (beta) Contact Us
Tools: MS1 Tool, MS2 CFMID Tool
Documentation: Source Code

Run MS2 CFMID Tool

Input	Value
Project name:	Example ms2 nta
Positive mode MS2 files (mgf):	Choose Files No file chosen
Negative mode MS2 files (mgf):	Choose Files No file chosen
Precursor mass accuracy (ppm):	10
Fragment mass accuracy (Da):	0.02

Defaults | Clear | Save Metadata? | Submit

Data processing:
1) Feature de-duplication
2) Adduct Flagging

Feature List

MS^1 Module
Use case: Summarize and identify MS1 features using EPA's DSSTox database
QC (i.e., Tracers): .csv of mass and retention time of internal standards.
QC data are tracked as 'Tracers', which represent internal standards extracted from the peak list data and summarized as tabular and graphical representations. These data allow users to *trace* method performance and document data quality.

MS^2 Module
Use case: Extract unique features from MS^2 data and match against *in silico* spectra stored in EPA's CFMID database

Feature list
List of unique Features Identified in data and used to subsequent identification steps.

Feature Identification using Curated Databases

MS^1 Feature Identification - DSSTox Database

- Database of over 850,000 chemicals used to support EPA's computational and toxicological activities. (Grulke *et al.*, 2019. doi:10.1016/j.comtox.2019.100096)
- Each entry is curated by the EPA to ensure high-quality representations of chemical structures.
- The database can be queried using chemical formulae or monoisotopic masses from the Feature List prepares by MS^1.
- MS-Ready representations of each chemical preserve relational mappings between substance (e.g., chemical mixtures or salts) and chemical structural. (McEachran *et al., 2018. doi: 10.1186/s13321-018-0299-2)*

MS^2 Feature Identification - CFMID Database

- Reference library containing *in silico* spectra of compatible chemicals in the DSSTox database. (McEachran *et al.*, 2019, doi: 41597-019-0145-z)
- Spectra were generated using Competitive Fragmentation Modeling of Metabolite Identification (CFMID, v2.0). These predictions were prepared for ESI-positive and ESI-negative modes at 10, 20, and 40 eV collision energies.
- Similarity between experimental spectra and reference spectra are calculated using a composite dot-product algorithm for each combination of collision energy. Reference spectra are ranked by the sum of similarity scores

Chemical Prioritization

Chemical Prioritization of MS^1 data

Candidate structures identified by the MS^1 module can be prioritized using chemical metadata. Currently, the MS^1 module uses total number of cataloged vendors or suppliers (referred to as 'data sources') as the default method for ranking candidate structures. Users can develop their own rules for prioritization using metadata provided from three sources:

1) *DSSTox database* – combination of exposure and toxicity data
2) *Hazard Comparison Dashboard* – predicted or known toxicity data
3) *MS^2 Module Results* – spectra similarity calculated as part of the MS^2 module

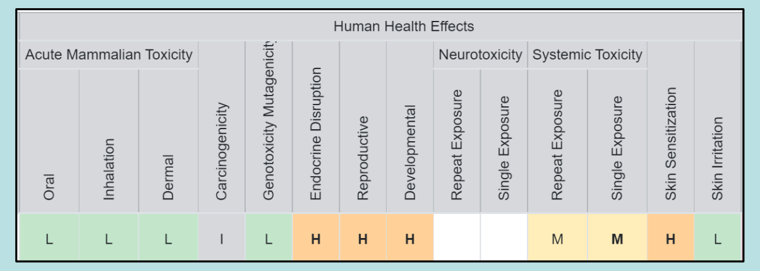
DSSTox database

Provides summary data for the EPA's ToxCast and ExpoCast data sets

- ToxCast toxicity
- ExpoCast exposure estimates
- Feature abundance
- Sample detection frequency

Hazard Comparison Dashboard

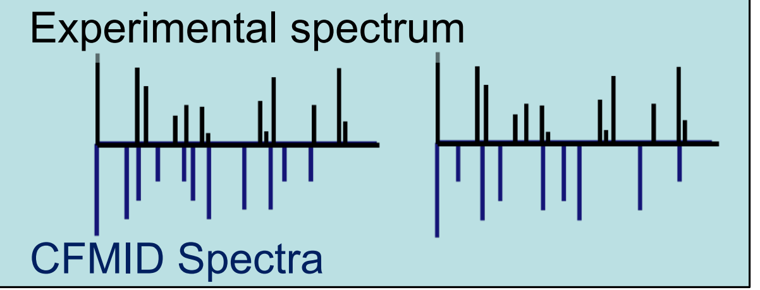
Provides toxicity data curated from literature or predicted by quantitative structural-activity relationship models.



Human Health Effects													
Acute Mammalian Toxicity								Neurotoxicity		Systemic Toxicity			
Oral	Inhalation	Dermal	Carcinogenicity	Genotoxicity Mutagenicity	Endocrine Disruption	Reproductive	Developmental	Repeat Exposure	Single Exposure	Repeat Exposure	Single Exposure	Skin Sensitization	Skin Irritation
L	L	L	I	L	H	H	H			M	M	H	L

MS^2 Module Results

Similarity between experimentally derived MS^2 spectra and CFMID spectra to prioritize structure



Data Output for Downstream Analysis

MS^1 Output

MS^1 data are provided in two formats: feature-level and chemical-level. Feature-level data provide summary statistics for each unique feature, while chemical-level data provide metadata for each tentative candidate identified for a feature. Data generated as part of the MS^2 module can be merged with the Chemical-level data for additional information to help guide feature identification.

Feature Level Results

Feature ID	*Mass*	*Retention Time*	*Samp1*	*Samp2*	*Samp3*
1	210.0876	6.904999	*Blank-subtracted median abundance values (Cleaned)*		
2	202.1223	7.808004			
3	670.5638	12.535			
4	706.5684	12.45099			

Chemical Level Results

Feature ID	*Chemical*	*MS-Ready Formula*	*Chem. Data 1*	*Chem. Data 2*	*Chem. Data 3*
1	Chemical Candidate 1	*MS-Ready Formula*	*Chemical-specific data and metadata values (ToxCast, ExpoCast, data sources, MS2 scores)*		
	Chemical Candidate 2				
	Chemical Candidate 3				
2	Chemical Candidate 1				
	Chemical Candidate 2				

What's Next for the Toolkit?

Development of the alpha version is set to complete by the end of 2022 and an accompanying manuscript will be released that details available features.

The modules outlined in this study represent the minimum functionality to aid in the analysis of $MS^1$and MS^2 data. Future updates will focus on adding additional modules that expand the capabilities of analyzing NTA data: *quantification of NTA data, implementation of experimental MS^2 databases, visualization of network maps, identification of metabolites*

Innovative Research for a Sustainable Future

This work does not reflect EPA policy.