Identifying markers of exposure using a combination of in silico *predictive tools and non-targeted analysis*

Matthew W. Boyce[a,b], Kristin A. Farvala[c], Jon Sobus[b], Antony Williams[b], Alex Chao[b], John F. Wambaugh[b], Lucina Lizarraga[d], Grace Patlewicz[b]

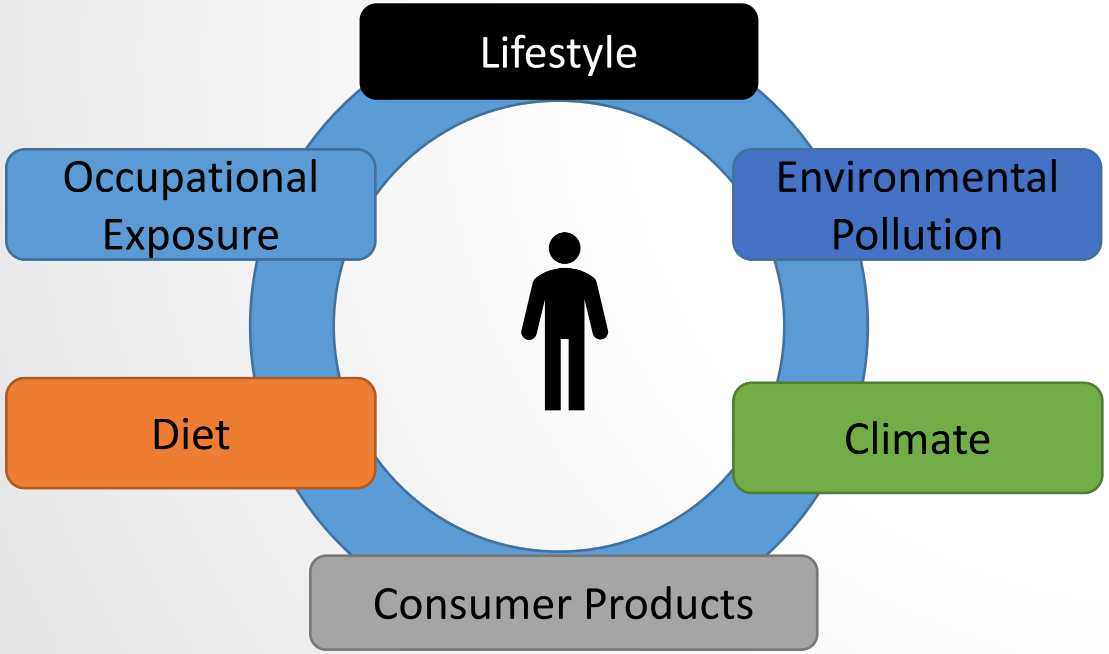
[a]Oak Ridge Associated University, Oak Ridge, TN 37830, USA
[b]Center for Computational Toxicology and Exposure, US EPA, RTP, NC 27711, USA
[c]Forensics and Specialty Analysis, Southwest Research Institute, San Antonio, TX 78228, USA
[d]National Center for Environmental Assessment, US EPA, Cincinnati, OH 45268, USA

This presentation does not reflect EPA policy.

Understanding the Exposome

"…the *exposome* encompasses life-course environmental exposures (including lifestyle factors), from the prenatal period onwards"

-Christopher Wild



Why study the Exposome?

~10% of diseases can be attributed to genetics, while the remaining stem from environmental sources

- Find associations between chemicals and disease
- Determine health risk, susceptibility, or disease progression

Understanding the health risk of an exposure requires understanding the metabolic fate of the substance

A need for metabolic data

Most compounds lack metabolic data which limits our ability to accurately assess health risk

- Read-Across can be used to bridge data-gaps for risk assessment; however, selection of appropriate analogues should account for metabolic similarities
- *In silico* tools can provide metabolic predictions, but their accuracy is hard to assess

Analytical challenges to measuring metabolites

- Metabolites are measured within complex mixtures and require additional computation methods to differentiate relevant metabolites from the remaining matrix
- Metabolites are often orders of magnitude lower in abundance than endogenous compounds
- There are a lack of spectral databases or standards to confirm identifications

Coupling Non-targeted analysis and *in silico* predictions

Non-targeted analysis (NTA): A tool suited for metabolomics

A methodology that uses high-resolution mass spectrometry (HRMS) to analyze many distinct features within a complex sample. Suited for analysis without *a priori* knowledge and can be used for identification or semi-quantification.

Using *in silico* predictions to guide NTA

Predicting metabolic structures

- Prediction software provide discrete structures to reference against HR-MS spectra and serve as a ***Suspect-Screening list***
- Aggregating results from multiple prediction software provides a thorough breadth of predictions to improve coverage

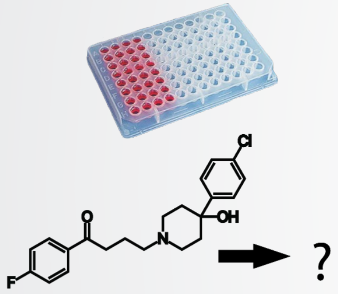
Generating a MS Spectra Database

- Converts structures predicted from *in* silico tools into MS^2 fragmentation spectra for structure identification
- Overcomes the limitation of having little to no available reference spectra for novel or poorly studied compounds

Guiding NTA with *in silico* predictions

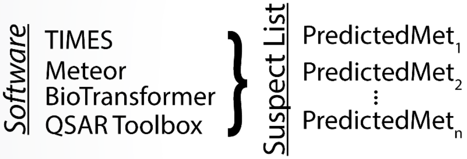
Sample Preparation → *In Silico* Data Generation → Data Acquisition → Data Processing → Data Analysis

In Vitro Assay



In Silico data

Aggregate Metabolite Predictions



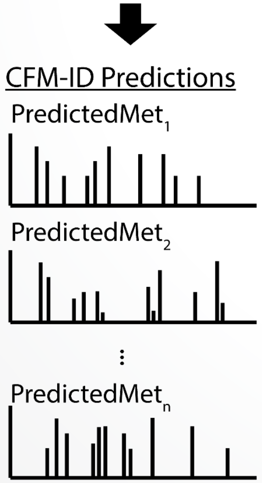
Software

TIMES
Meteor
BioTransformer
QSAR Toolbox

Suspect List

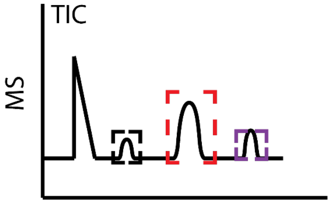
PredictedMet$_1$
PredictedMet$_2$
⋮
PredictedMet$_n$

CFM-ID Predictions

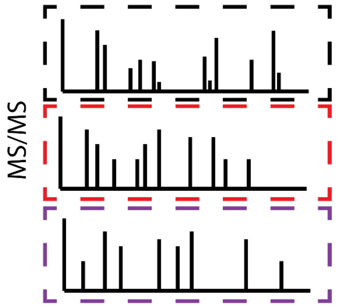


Non-Targeted Analysis

Parent ion selection



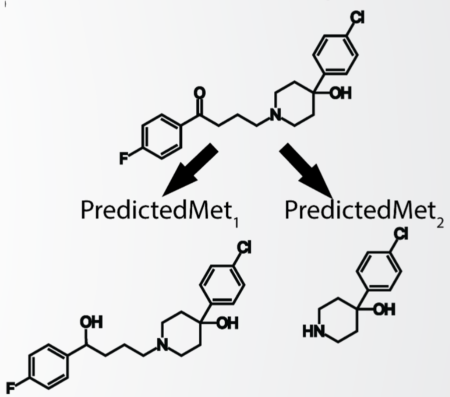
Parent ion fragmentation



Feature Selection & Data Cleaning

1) Peak Selection
2) Feature Identification
3) Data Cleaning

Metabolite Assignment



Starting Compounds

Sample Preparation

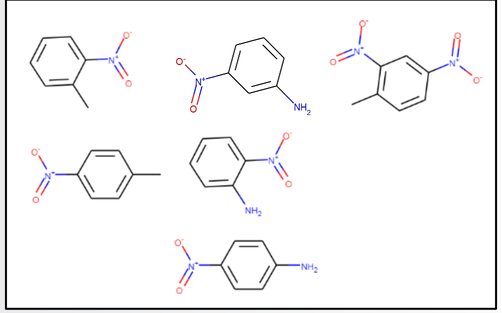
In Silico Data Generation

Data Acquisition

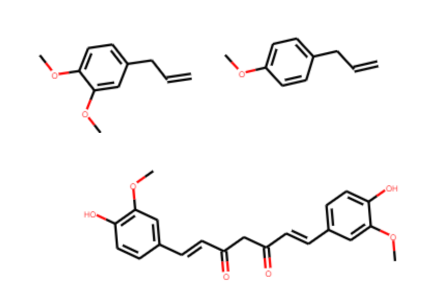
Data Processing

Data Analysis

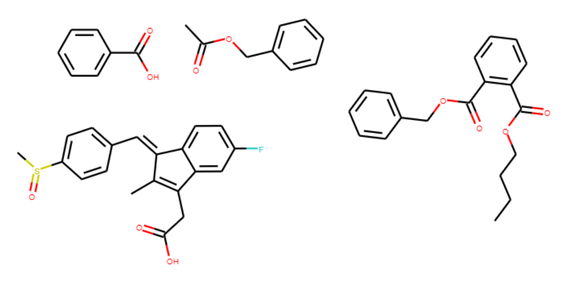
Nitrobenzenes



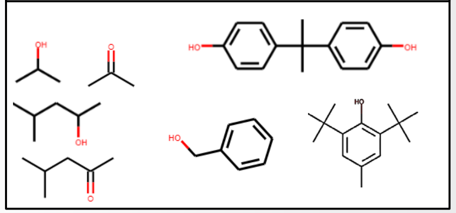
Anisoles



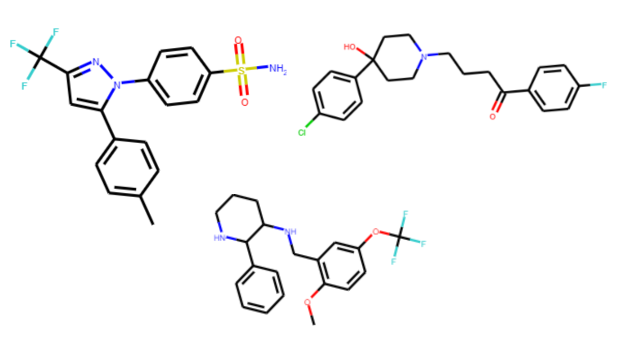
Benzoyl/Carboxylic acids



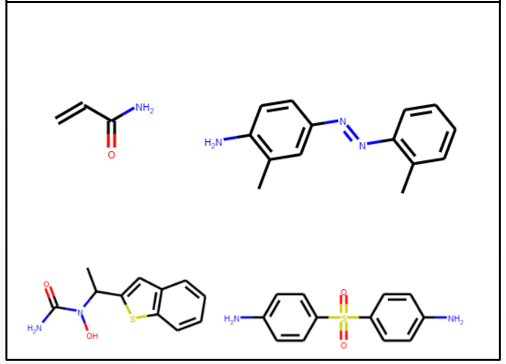
Ketones/Alcohols



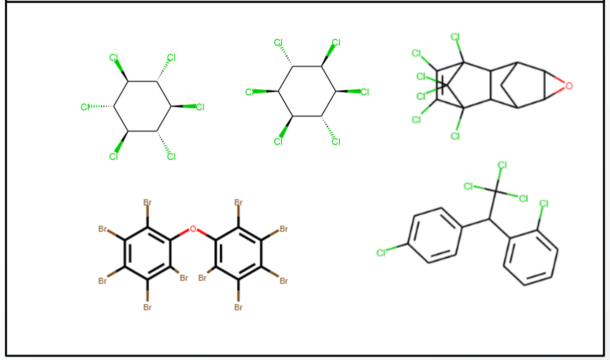
Organofluorines



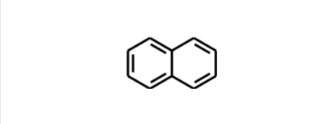
Amide/Aniline



Organohalide ring structures



Napthalene



In vitro assay

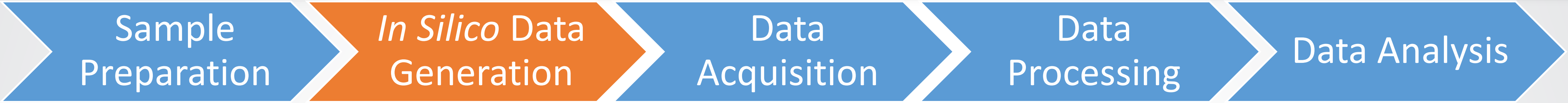
Sample Preparation
In Silico Data Generation
Data Acquisition
Data Processing
Data Analysis

Metabolite Generation

- Starting compounds metabolized via pooled primary human hepatocytes (10 donors)
    - Three time points: 0, 1, 4h
    - Three sample treatments: Supernatant (post lysis), B-glucuronidase treated, cell pellet

- Standards/Controls
    - Vehicle blank – DMSO
        - Used as blank for MS analysis
    - Standard control – Cell free solution with compound
        - Used to identify retention time window and mass error

Compiling a suspect screening list



Sample Preparation

In Silico Data Generation

Data Acquisition

Data Processing

Data Analysis

Known Metabolites

- Pulled 438 metabolites from 49 papers
- Markush structures were enumerated
- Metabolites registered into EPA's DSSTox chemical registration system to generate specific identifiers (DTXSID/DTXCIDs) to facilitate subsequent data analysis

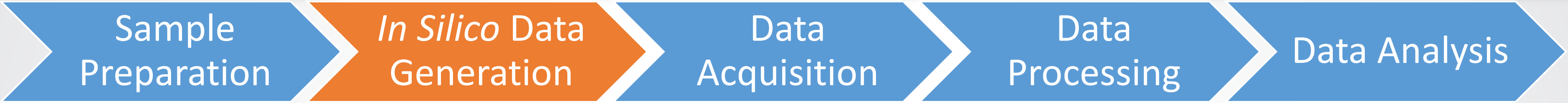
Predicted Metabolites

- Compiled predicted structures from:
  - TIMES
  - BioTransformer
  - QSAR Toolbox
  - Meteor Nexus
- 1,666 predictions in total

Suspect Screening List

- 490 unique molecular formulae for MS^1 formula assignment
- Used to guide MS^2 analysis and generate CFM-ID predictions

Generating database of *in silico* MS^2 spectra



Sample Preparation

In Silico Data Generation

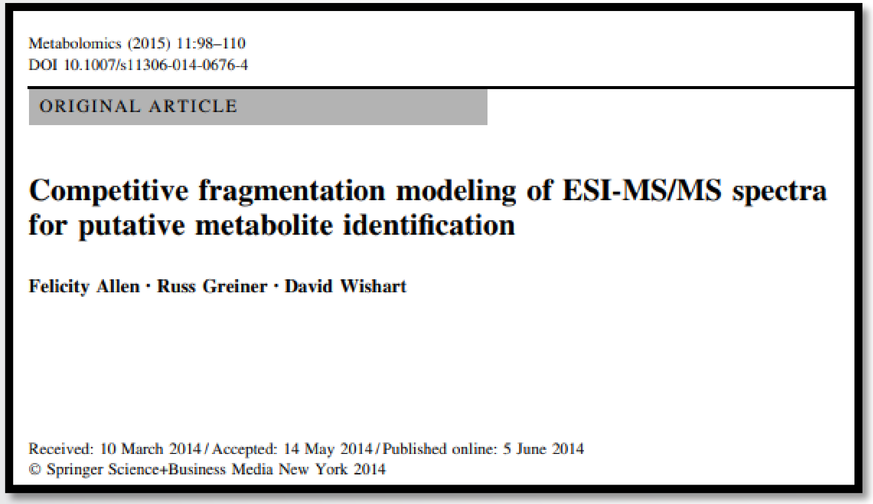
Data Acquisition

Data Processing

Data Analysis

Fragmentation spectra were generated for each predicted metabolite

Competitive Fragmentation Modeling-ID (CFM-ID)



Metabolomics (2015) 11:98–110
DOI 10.1007/s11306-014-0676-4

ORIGINAL ARTICLE

Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification

Felicity Allen · Russ Greiner · David Wishart

Received: 10 March 2014 / Accepted: 14 May 2014 / Published online: 5 June 2014

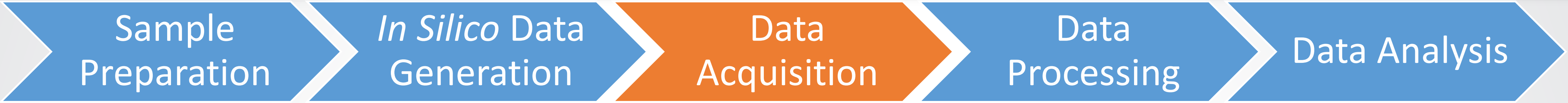
Spectra were generated using CFM-ID

- Reference spectra were generated at three collision energies (CE)
- Data were stored in database to query against for comparisons
- Validated against CASMI datasets for HRMS identification

  DOI: 10.3390/metabo10060260
- Implemented into EPA's CompTox Dashboard

  DOI: 10.1038/s41597-019-0145-z

MS^1 and MS^2 analysis



Sample Preparation

In Silico Data Generation

Data Acquisition

Data Processing

Data Analysis

LC-qTOF was used to collect high resolution MS^1 and MS^2 data

MS^1

- ESI+ and ESI-
- Range 100 – 1700 m/z
- Used to collect features for identification

MS^2

- Data-dependent acquisition (using suspect screening list)
- 1 replicate per treatment per time point
- Used to identify a feature’s probable structure

Preliminary analysis of MS^1 data to select samples for further analysis

- Candidate metabolites identified for 17 of 33 compounds
- Parent peaks present for 12 of 33 compounds
- Compounds with identified metabolites are be carried forward

Data processing steps

Sample Preparation

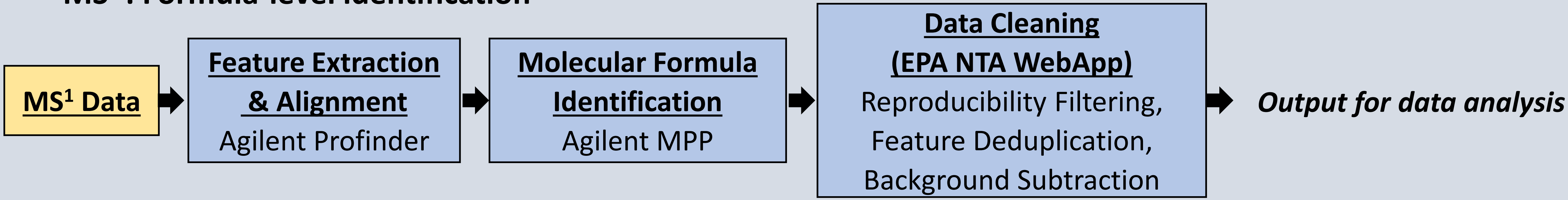
In Silico Data Generation

Data Acquisition

Data Processing

Data Analysis

MS^1 : Formula-level identification



MS^1 Data

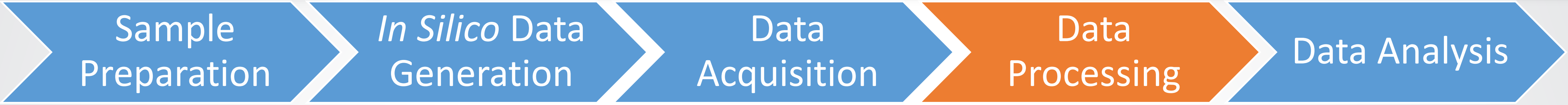
Feature Extraction & Alignment
Agilent Profinder

Molecular Formula Identification
Agilent MPP

Data Cleaning (EPA NTA WebApp)
Reproducibility Filtering,
Feature Deduplication,
Background Subtraction

Output for data analysis

Data processing steps



Sample Preparation

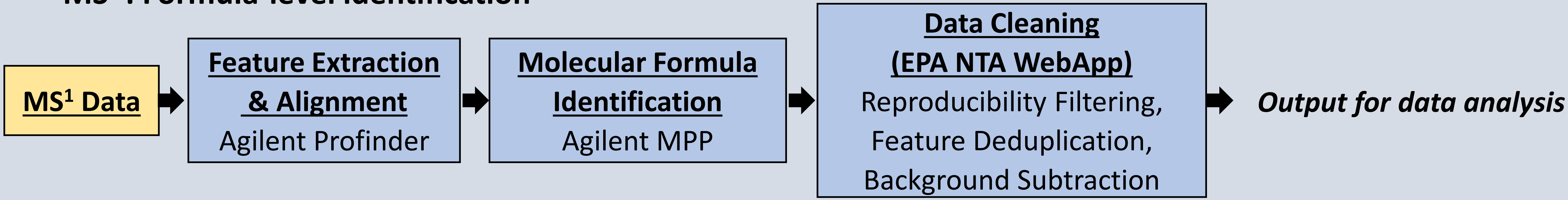
In Silico Data Generation

Data Acquisition

Data Processing

Data Analysis

MS^1 : Formula-level identification



MS^1 Data

Feature Extraction & Alignment
Agilent Profinder

Molecular Formula Identification
Agilent MPP

Data Cleaning (EPA NTA WebApp)
Reproducibility Filtering, Feature Deduplication, Background Subtraction

Output for data analysis

Output of MS^1 processing: Annotated features

Suspect-Screening matches

- Identified using suspect list
- Molecular formula with suspected structural assignments

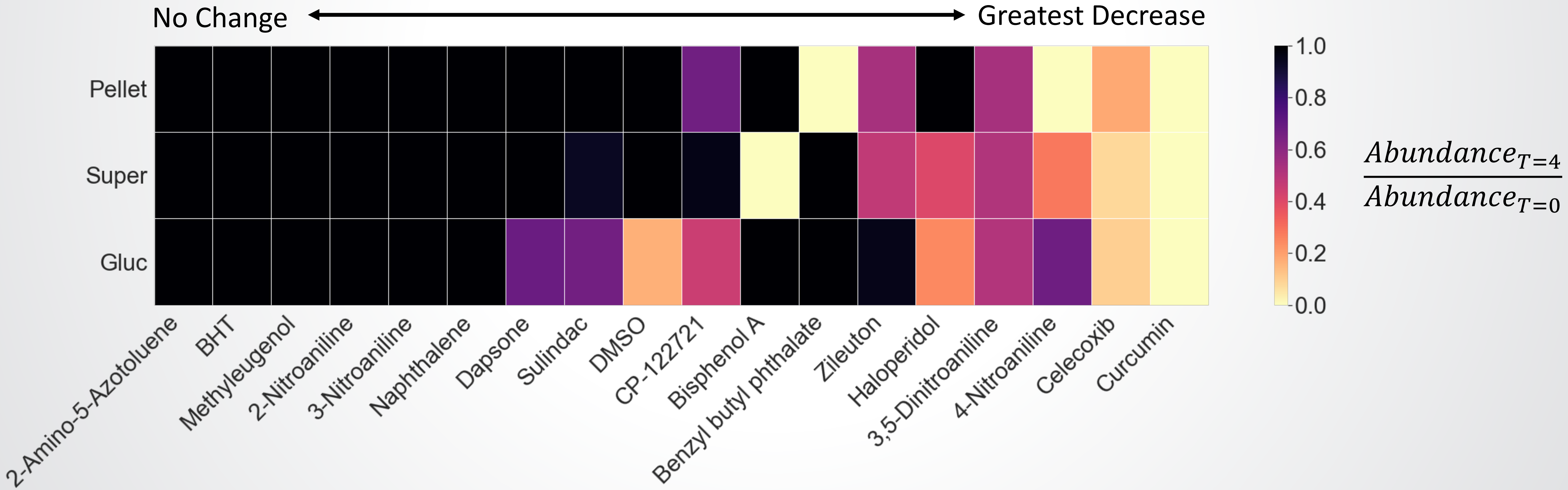
Features without suspect matches

- Formula proposed using Agilent's Molecular-Formula generator
- Formulae with no known structural assignments

Which parents are being metabolized?

Sample Preparation → *In Silico* Data Generation → Data Acquisition → Data Processing → **Data Analysis**

Relative change in parent signal over 4h



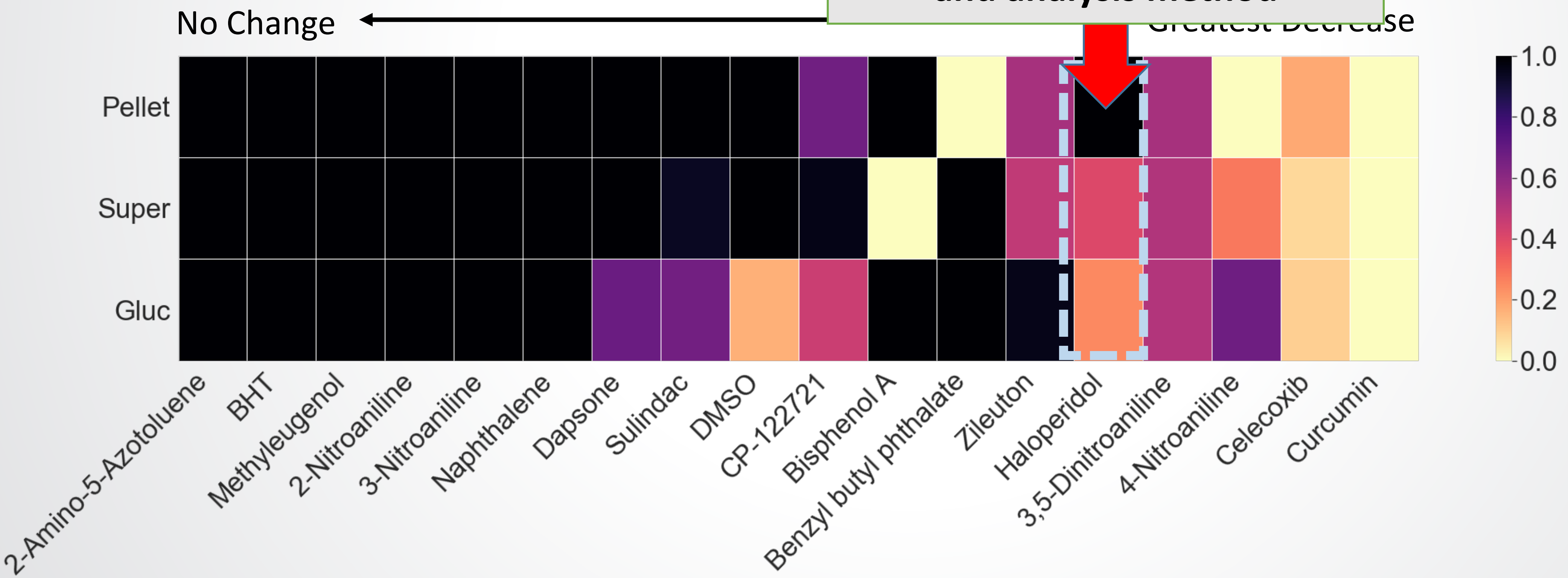
$$\frac{Abundance_{T=4}}{Abundance_{T=0}}$$

Which parents are being metabolized?

Sample Preparation
In Silico Data Generation
Data Acquisition
Data Processing
Data Analysis

Relative change in pare

Used to develop processing and analysis method



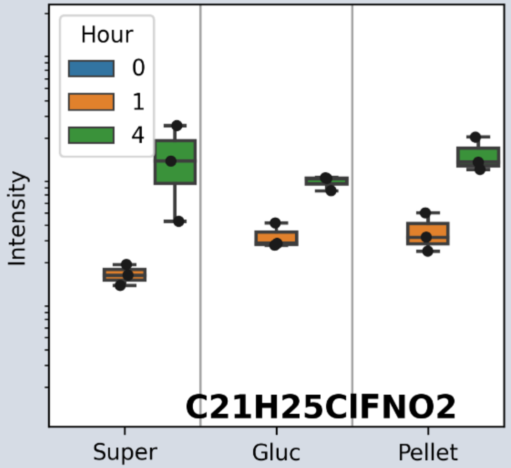
$$\frac{Abundance_{T=4}}{Abundance_{T=0}}$$

Identifying relevant features

Sample Preparation → *In Silico* Data Generation → Data Acquisition → Data Processing → Data Analysis

MS[1] Analysis Workflow

1) Broad feature filtering



Criteria for selecting features:

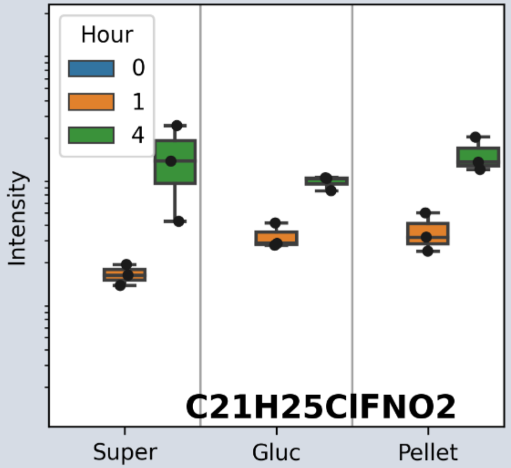
1. Increases over time
2. Appears in a minimum of two time points

Identifying relevant features

Sample Preparation → *In Silico* Data Generation → Data Acquisition → Data Processing → Data Analysis

MS^1 Analysis Workflow

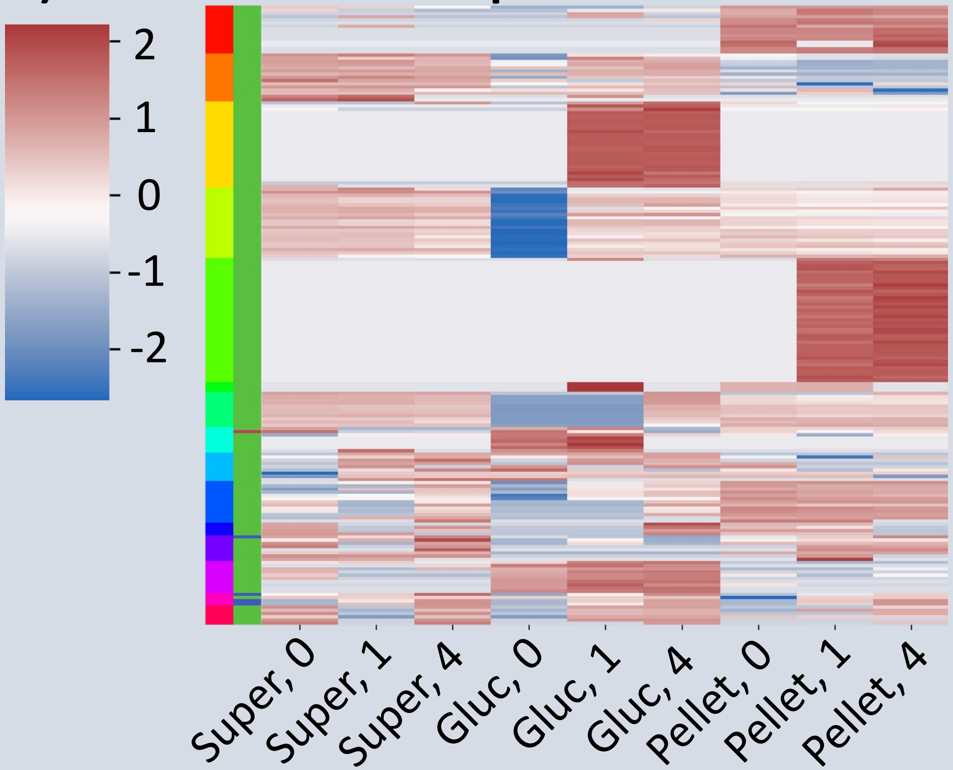
1) Broad feature filtering



Criteria for selecting features:

1. Increases over time
2. Appears in a minimum of two time points

2) Cluster similar compounds

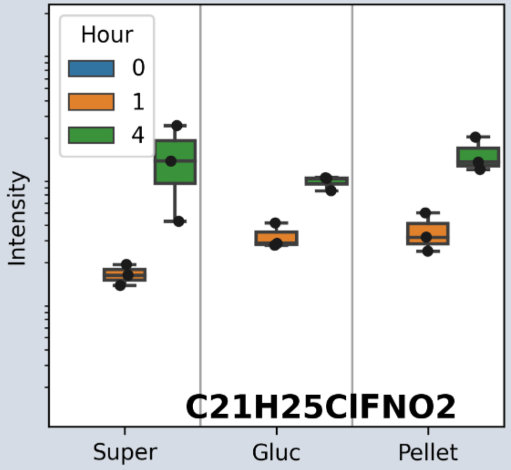


Identifying relevant features

Sample Preparation
In Silico Data Generation
Data Acquisition
Data Processing
Data Analysis

MS^1 Analysis Workflow

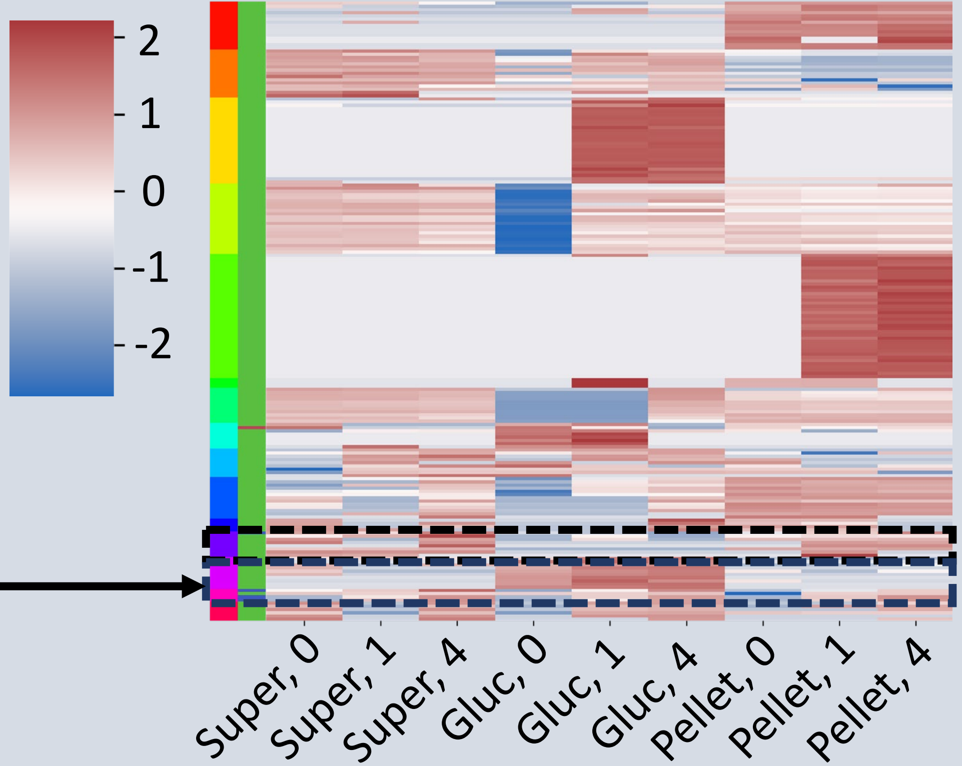
1) Broad feature filtering



Criteria for selecting features:

1. Increases over time
2. Appears in a minimum of two time points

2) Cluster similar compounds



Clusters containing features annotated of known metabolites

Identifying relevant features

Sample Preparation

In Silico Data Generation

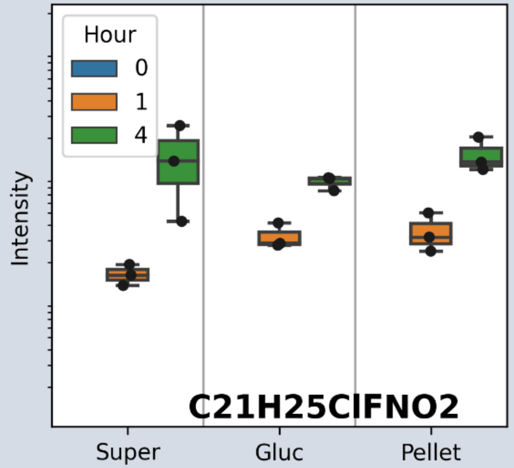
Data Acquisition

Data Processing

Data Analysis

MS^1 Analysis Workflow

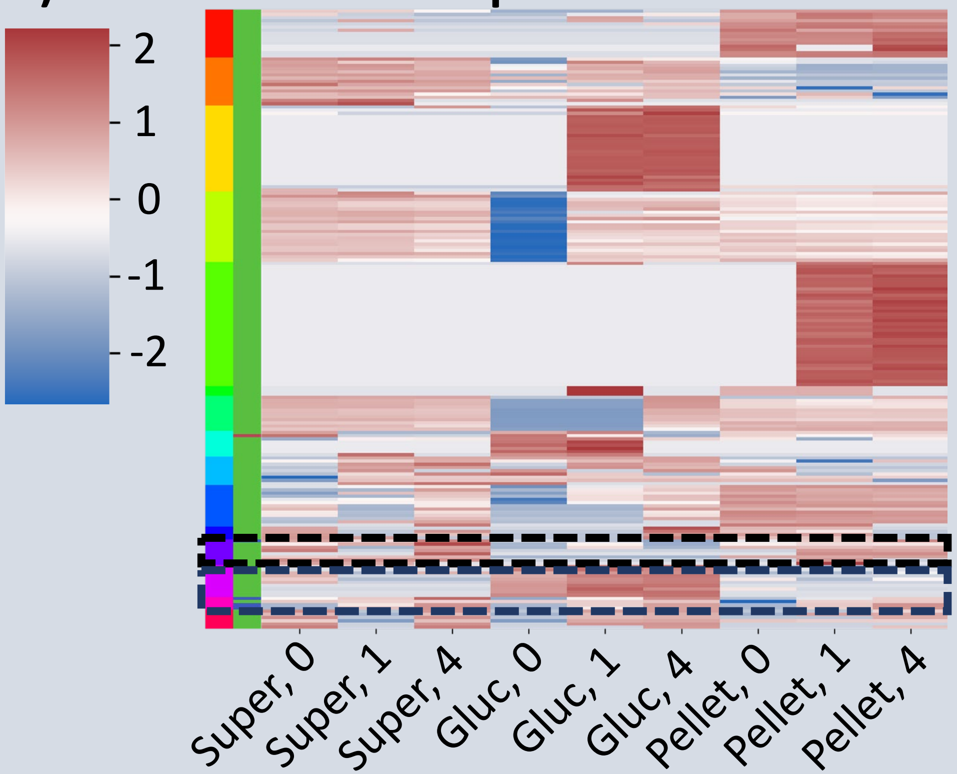
1) Broad feature filtering



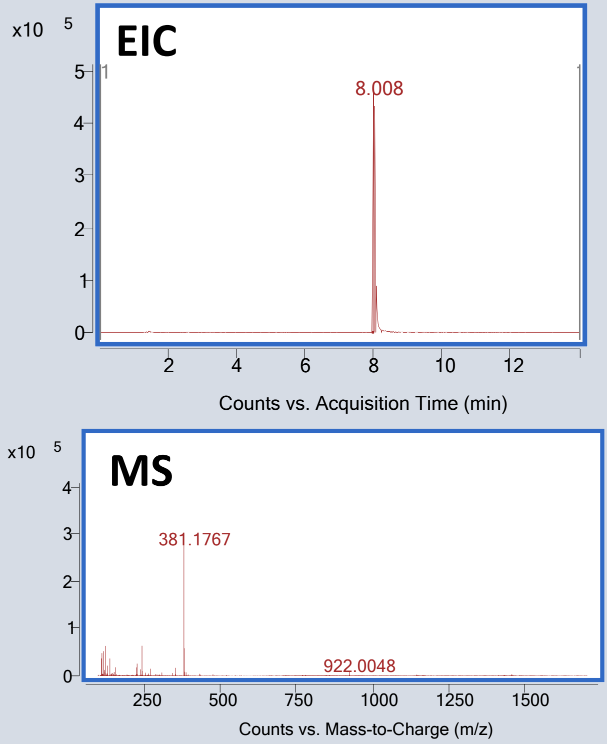
Criteria for selecting features:

1. Increases over time
2. Appears in a minimum of two time points

2) Cluster similar compounds



3) Manual Review



Assigning structure to features

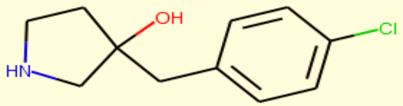
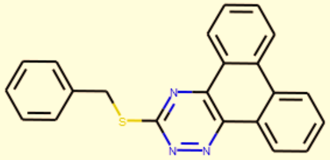
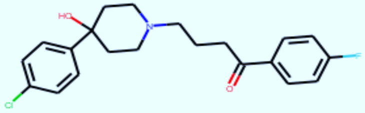
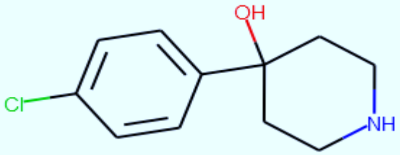
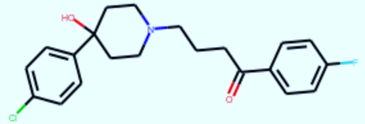
Sample Preparation → *In Silico* Data Generation → Data Acquisition → Data Processing → **Data Analysis**

MS^2 Analysis Workflow

CFM-ID Comparisons

- MS^2 data were matched against the CFM-ID database and scored based on similarity to the predicted spectra at each CE
- Predictions were ranked based on the sum of the similarity values, and normalized as a 'Q-Score' (ranging from 0 – 1)

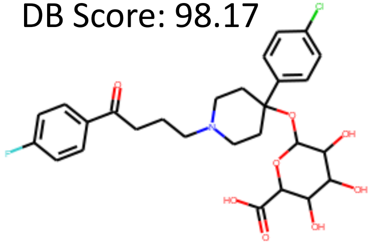
Match Prioritization

	C11H14ClNO 8 matches	C22H15N3S 1 match	C21H23ClFNO2 2 matches
Top Match	Q-Score: 1.0  	Q-Score: 1.0  	Q-Score: 1.0 
Expected Match	Q-Score = 0.90  	None	Q-Score: 1.0 

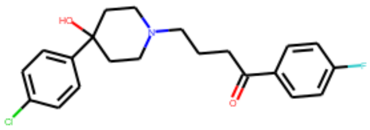
Metabolite identifications

CFM-ID Match (level 2b):

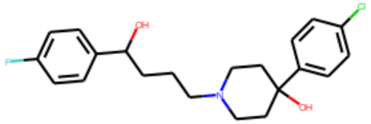
Q-Score: 1.0
DB Score: 98.17



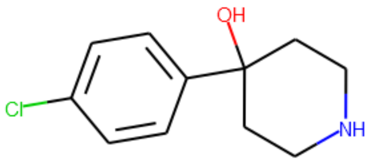
Q-Score: 1.0
DB Score: 85.04



Q-Score: 1.0
DB Score: 98.43

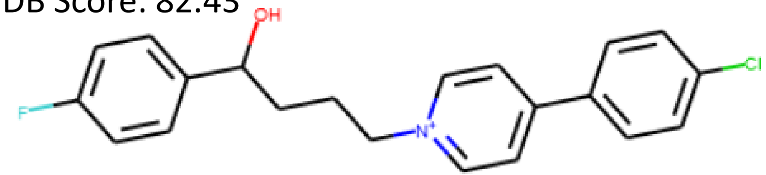


Q-Score: 0.9
DB Score: 95.83

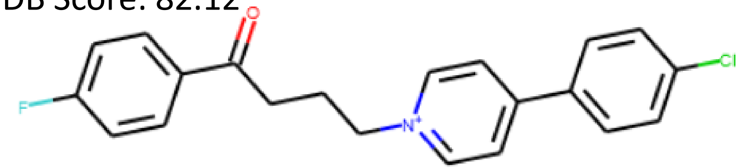


Suspect Match (level 3):

DB Score: 82.43



DB Score: 82.12



Predicted Formula (level 5):

- $C_{9}H_{13}ClN_{2}O_{6}$ (278.0816)
- $C_{9}H_{11}NO_{3}$ (181.0738)
- $C_{6}H_{10}O_{3}$ (130.0632)
- $C_{24}H_{30}N_{2}O_{11}$ (522.1851)

Conclusions and Next Steps

We have developed a NTA workflow for characterizing metabolic profiles of target compounds:

- *In silico* tools to develop a suspect screening list and MS^2 spectra database
- Agilent software and the NTA WebApp to process/clean the data
- Statistical analysis to find relevant features for identification

We are working through the remaining data and are interested in using the results to:

- Benchmark the performance of the *in silico* metabolite prediction software
- Derive kinetics relationships for parent compounds and their metabolites
- Expanding this method for the characterization of data-poor compounds to assist in risk assessment

Acknowledgements

Environmental Protection Agency

- Grace Patlewicz
- John Wambaugh
- Lucina Lizarraga
- Jon Sobus
- Alex Chao
- Tony Williams
- Chris Grulke
- Ann Richards
- Brian Meyer
- Vicente Samano
- Nancy Baker
- Daniel Chang

Southwest Research Institute

- Kristin Farvala

Thermo Scientific

- Jessica A. Bonzo

Where to reach me

Boyce.matthew@epa.gov

linkedin.com/in/mwboyce16/

orcid.org/0000-0002-3794-1678