United States Environmental Protection Agency

The NTA WebApp:

A web-based tool for rapid chemical identification from non-targeted analysis mass spectrometry data

Alex Chao

Jeffrey M. Minucci, Matthew W. Boyce, S. Thomas Purucker, Antony J. Williams, Jon R. Sobus

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA

Office of Research and Development
Center for Computational Toxicology & Exposure

ACS 2022
March 22, 2022

Exposomics: Addressing Health Issues

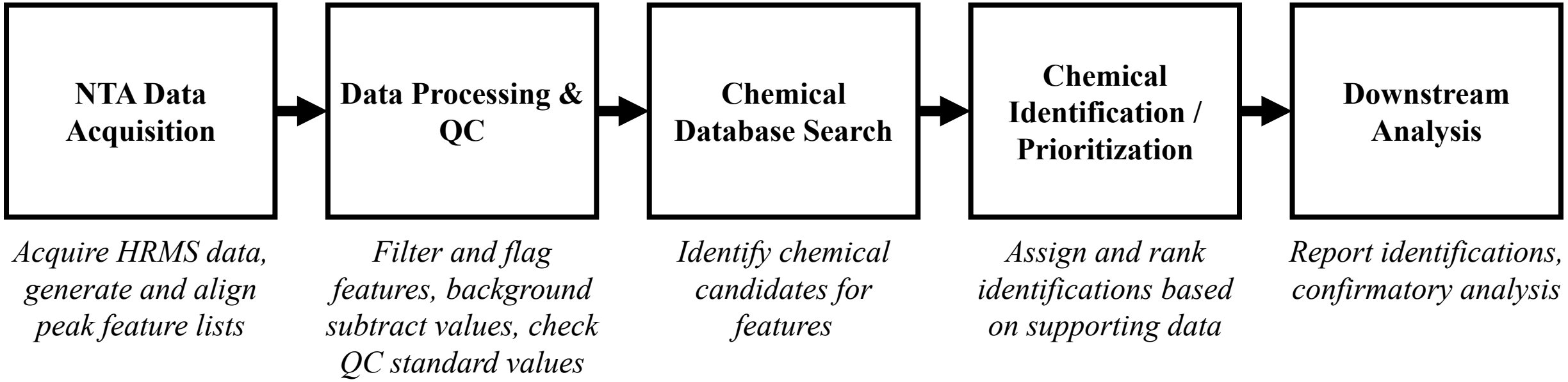
Chemical Monitoring Methods

What are the chemicals we are exposed to?

What are the amounts?

Do we need to be worried about it?

Development of an NTA Workflow



Development of a NTA Workflow

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

- **The Underlying Database: *DSSTox***
  - A highly curated database of environmentally relevant chemicals (>900k)
- **Generating Compatibility with HRMS Data: *MS-Ready Forms***
  - Mapping substance components into forms that would be observed by MS
- **Allowing Access into the Database: *CompTox Chemicals Dashboard***
  - A web-based conduit into DSSTox allowing for batch searching of MS data

EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research

Christopher M Grulke [1], Antony J Williams [1], Inthirany Thillanadarajah [2], Ann M Richard [1]

"MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

Andrew D McEachran [1 2], Kamel Mansouri [3 4 5], Chris Grulke [4], Emma L Schymanski [6], Christoph Ruttkies [7], Antony J Williams [8]

The CompTox Chemistry Dashboard: a community data resource for environmental chemistry

Antony J Williams [1], Christopher M Grulke [2], Jeff Edwards [2], Andrew D McEachran [3], Kamel Mansouri [2 3 4], Nancy C Baker [5], Grace Patlewicz [2], Imran Shah [2], John F Wambaugh [2], Richard S Judson [2], Ann M Richard [2]

Grulke, Christopher M., et al. Computational Toxicology 12 (2019): 100096.
McEachran, Andrew D., et al. Journal of cheminformatics 10.1 (2018): 1-16.
Williams, Antony J., et al. Journal of cheminformatics 9.1 (2017): 1-27.

https://comptox.epa.gov/dashboard/

Development of a NTA Workflow

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

- **Chemical ID via metadata: *Data Sources***
  - Count of chemical presence in publications, projects, data collections
- **Chemical ID via *in silico* predictions: *CFM-ID generated MS2 spectra***
  - Predicted MS2 spectra generated for all chemicals within DSSTox
- **Performance evaluations:**
  - Critical Assessment of Small Molecule Identification (**CASMI**) spectra
  - EPA's Non-Targeted Analysis Collaborative Trial (**ENTACT**) mixtures

Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard

Andrew D McEachran [1], Jon R Sobus [2], Antony J Williams [3]

Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns

Andrew D McEachran [1 2], Ilya Balabin [3], Tommy Cathey [4], Thomas R Transue [4], Hussein Al-Ghoul [5], Chris Grulke [6], Jon R Sobus [7], Antony J Williams [8]

In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples

Alex C
Tomm
Anton

Revisiting Five Years of CASMI Contests with EPA Identification Tools

Andrew D McEachran [1], Alex Chao [1], Hussein Al-Ghoul [1], Charles Lowe [2], Christopher Grulke [2], Jon R Sobus [2], Antony J Williams [2]

McEachran, Andrew D., et al. Analytical and bioanalytical chemistry 409.7 (2017): 1729-1735.
McEachran, Andrew D., et al. Scientific data 6.1 (2019): 1-9.
Chao, Alex, et al. Analytical and bioanalytical chemistry 412.6 (2020): 1303-1315.
McEachran, Andrew D., et al. Metabolites 10.6 (2020): 260.

Development of a NTA Workflow

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

- **Chemical ID via metadata: *Data Sources***
  - Count of chemical presence in publications, projects, data collections
- **Chemical ID via *in silico* predictions: *CFM-ID generated MS2 spectra***
  - Predicted MS2 spectra generated for all chemicals within DSSTox
- **Performance evaluations:**
  - Critical Assessment of Small Molecule Identification (**CASMI**) spectra
  - EPA's Non-Targeted Analysis Collaborative Trial (**ENTACT**) mixtures

DSSTox

Database API

Identifying known unknowns using the US EPA's Comptox Chemistry Dashboard

Andrew D McEachran [1], Jon R Sobus [2], Antony J Williams [3]

Linking in silico MS/MS spectra with chemistry data to improve identification of unknowns

Andrew D McEachran [1 2], Ilya Balabin [3], Tommy Cathey [4], Thomas R Transue [4], Hussein Al-Ghoul [5], Chris Grulke [6], Jon R Sobus [7], Antony J Williams [8]

In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples

Alex Chao [1], Hussein Al-Ghoul [1], Andrew D McEachran [1], Ilya Balabin [1], Tom Transue [1], Tommy Cathey [1], Ashley Thomas [1], Randolph R Singh [3 7], Elin M Ulrich [8], Antony J Williams [9], Jon R Sobus [10]

Revisiting Five Years of CASMI Contests with EPA Identification Tools

Andrew D McEachran [1], Alex Chao [1], Hussein Al-Ghoul [1], Charles Lowe [2], Christopher Grulke [2], Jon R Sobus [2], Antony J Williams [2]

McEachran, Andrew D., et al. Analytical and bioanalytical chemistry 409.7 (2017): 1729-1735.
McEachran, Andrew D., et al. Scientific data 6.1 (2019): 1-9.
Chao, Alex, et al. Analytical and bioanalytical chemistry 412.6 (2020): 1303-1315.
McEachran, Andrew D., et al. Metabolites 10.6 (2020): 260.

Development of a NTA Workflow

- **Assessing Chemical Hazard: *ToxCast***
  - A program to generate chemical toxicity data via high-throughput screening assays
- **Assessing Chemical Exposure: *ExpoCast***
  - A program to estimate chemical exposure potential via high-throughput models
- **Assessing Exposure via Consumer Products: *Chemical and Products Database (CPDat)***
  - A database mapping chemicals with consumer product usage to assess potential exposures

The ToxCast program for prioritizing toxicity testing of environmental chemicals

David J Dix [1], Keith A Houck, Matthew T Martin, Ann M Richard, R Woodrow Setzer, Robert J Kavlock

High-throughput models for exposure-based chemical prioritization in the ExpoCast project

John F Wambaugh [1], R Woodrow Setzer, David M Reif, Sumit Gangwal, Jade Mitchell-Blackwood, Jon A Arnot, Olivier Joliet, Alicia Frame, James Rabinowitz, Thomas B Knudsen, Richard S Judson, Peter Egeghy, Daniel Vallero, Elaine A Cohen Hubal

The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products

Kathie L Dionisio [1], Katherine Phillips [1], Paul S Price [1], Christopher M Grulke [2], Antony Williams [2], Derya Biryol [1 3], Tao Hong [4], Kristin K Isaacs [1]

Dix, David J., et al. Toxicological sciences 95.1 (2007): 5-12.
Wambaugh, John F., et al. Environmental science & technology 47.15 (2013): 8479-8488.
Dionisio, Kathie L., et al. Scientific data 5.1 (2018): 1-9.

Development of a NTA Workflow

- **Assessing Chemical Hazard: *ToxCast***
  - A program to generate chemical toxicity data via high-throughput screening assays
- **Assessing Chemical Exposure: *ExpoCast***
  - A program to estimate chemical exposure potential via high-throughput models
- **Assessing Exposure via Consumer Products: *Chemical and Products Database (CPDat)***
  - A database mapping chemicals with consumer product usage to assess potential exposures

DSSTox

Database API

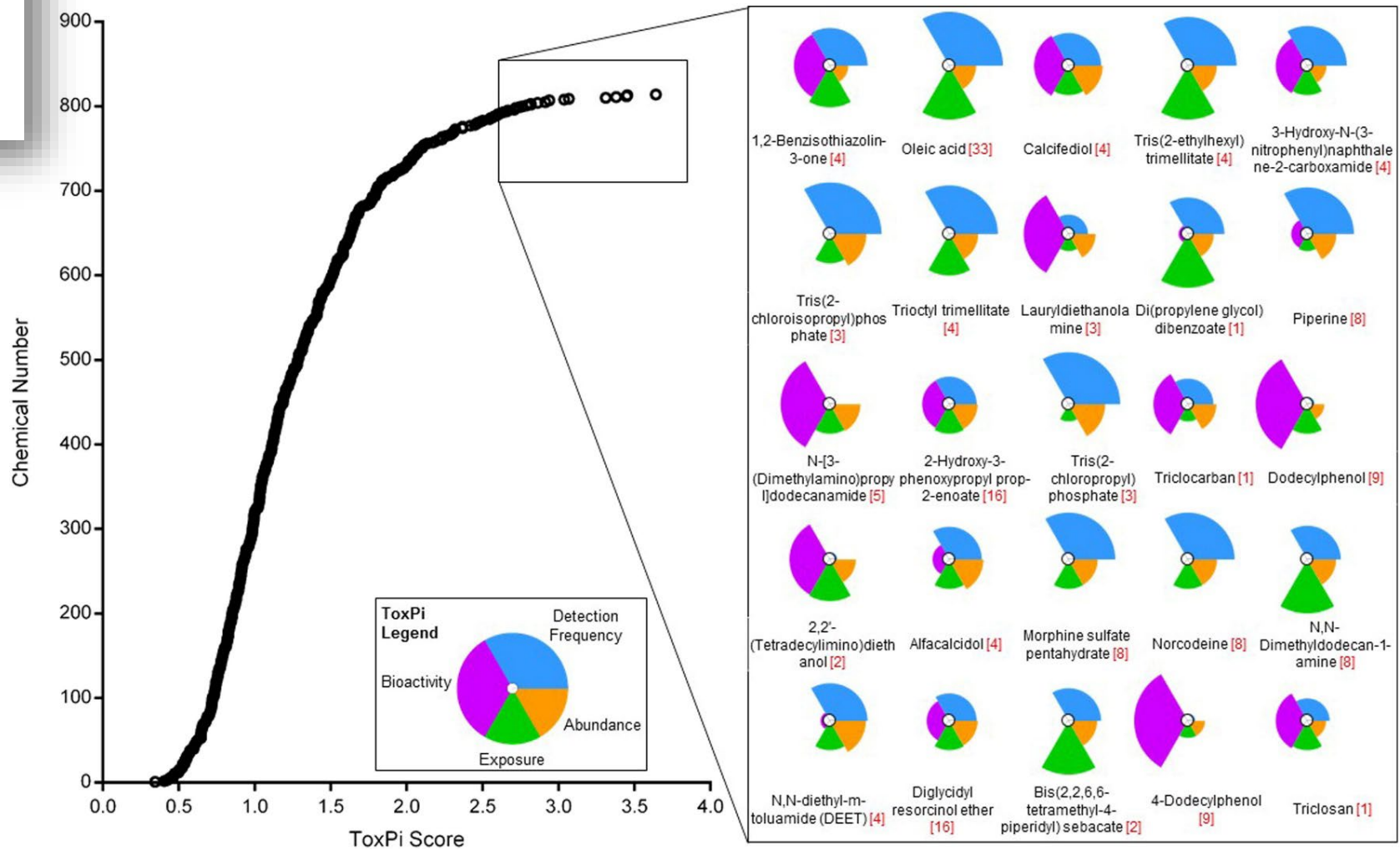
Dix, David J., et al. Toxicological sciences 95.1 (2007): 5-12.
Wambaugh, John F., et al. Environmental science & technology 47.15 (2013): 8479-8488.
Dionisio, Kathie L., et al. Scientific data 5.1 (2018): 1-9.

Chemical Risk Prioritization in House Dust

Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring

Julia E Rager [1], Mark J Strynar [2], Shuang Liang [1], Rebecca L McMahen [1], Ann M Richard [3], Christopher M Grulke [4], John F Wambaugh [3], Kristin K Isaacs [2], Richard Judson [3], Antony J Williams [3], Jon R Sobus [5]

- **Household dust as a metric of exposure**
  - Vacuum dust samples collected from 56 households for NTA
- ***ToxPi Score* to prioritize chemicals for follow-up**
  - ToxCast bioactivity
  - ExpoCast exposure estimates
  - Feature abundance
  - Sample detection frequency

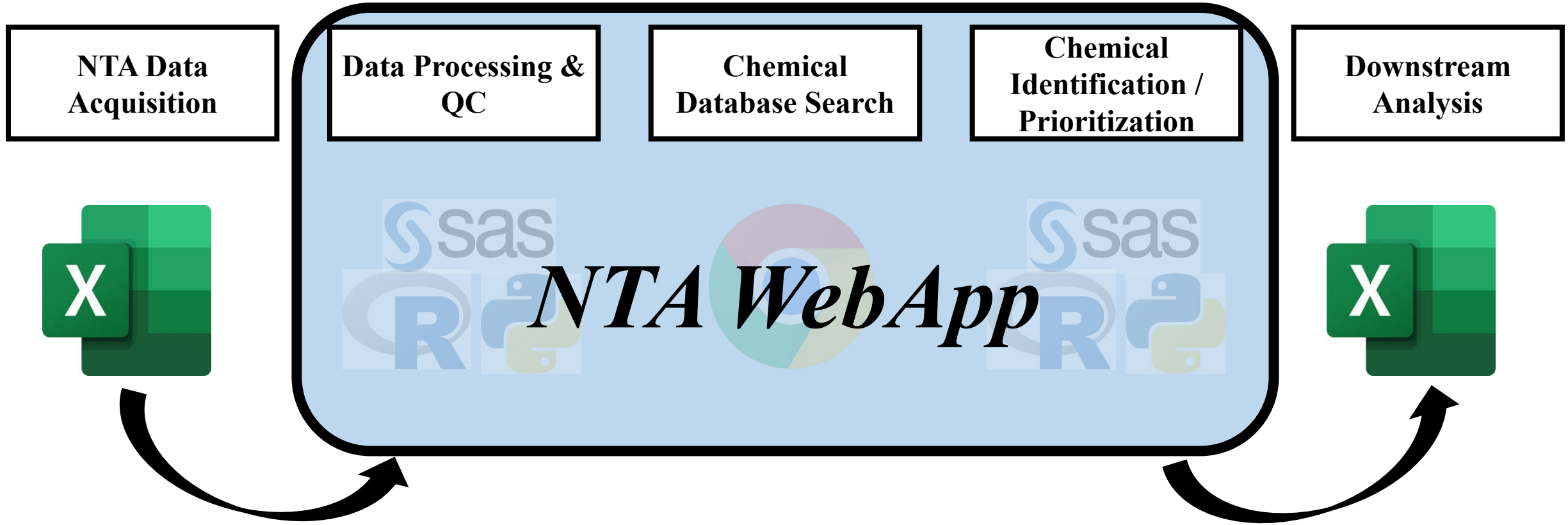


Rager, Julia E., et al. Environment international 88 (2016): 269-280.

Consolidation of Workflow into a WebApp

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

Consolidation of Workflow into a WebApp



- ***Standardization of processes*:** Single web-accessible point for processing NTA data
- ***Reduction of processing steps:*** Once input, data are carried through whole workflow
- ***Documentation of processes:*** Full workflow tracking for reproducibility and reporting (Input files, processing / search parameters, output results, QC results)

Data Processing Functionality

Perform Quality Control on NTA Data Set

Are features adducts of other features?

Are duplicate features present?

Are features reproducible across replicates?

Are features specific to samples?

Are initial annotations robust?

Are spiked tracers within allowable tolerances?

Flag and/or remove data points from the data set that do not meet QC criteria

Cleaning NTA Data for Reporting

- Determine median values for sample replicate groups
- Subtract blank median values for each feature from sample median values

In an NTA data set there may be **thousands** of features observed across **hundreds** of samples:

Millions of calculations required to clean and transform each data set

Chemical Database Search Functionality

Identify Potential Chemicals for Features

For each feature, search chemical database:

- Pull back all chemicals with a "matching" mass (mass within accuracy tolerance)
- Pull back associated chemical metadata

Distributed Structure Searchable Toxicity Database (DSSTox)

- WebApp has direct interface with DSSTox database

Prioritize Chemical Candidates for Features

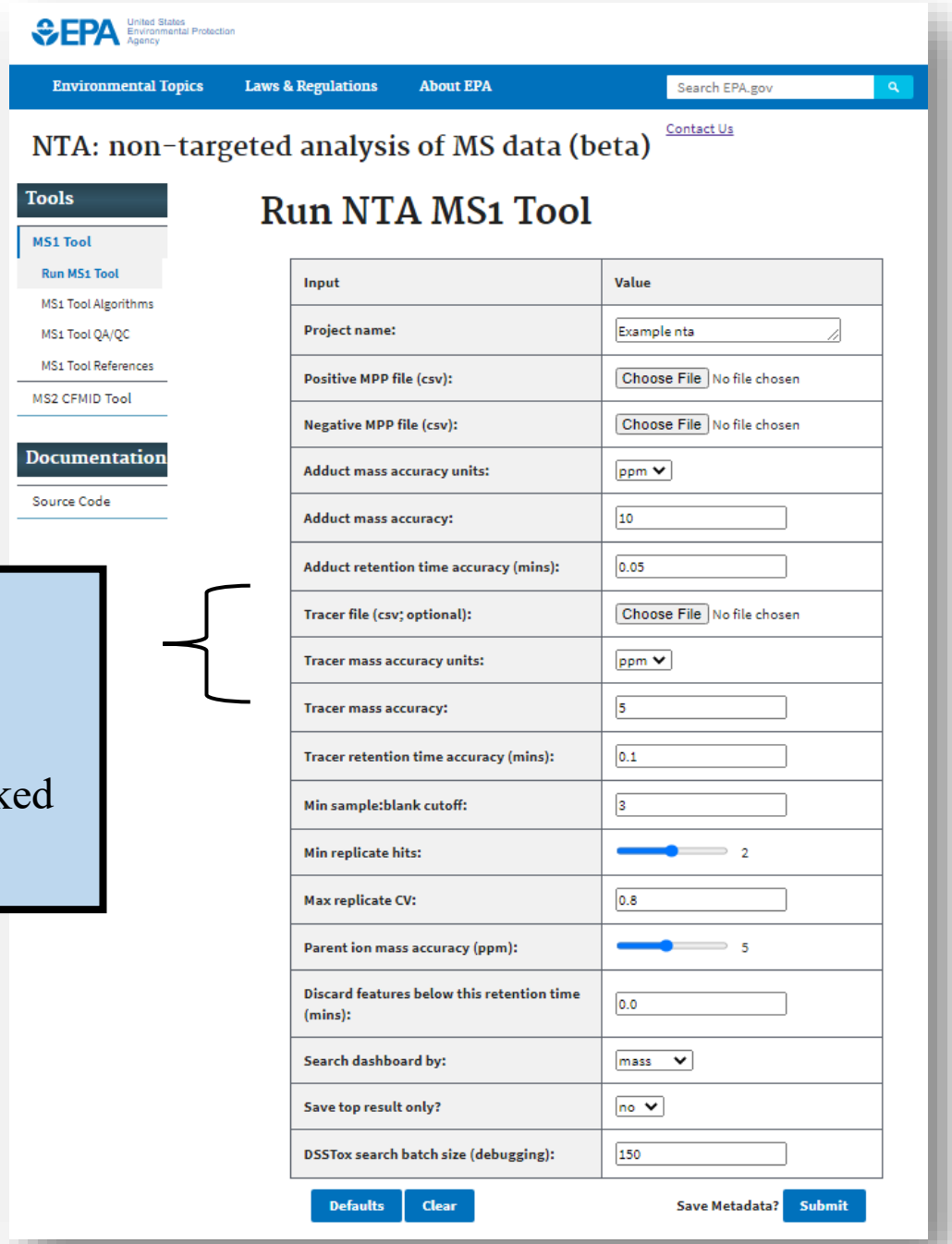
For each candidate, compare associated data to select the most likely candidates (e.g.):

- ToxCast assay data
- ExpoCast exposure estimates
- Data source counts
- MS2 *in silico* spectra scoring

DSSTox currently contains ~**906,000** chemicals to search through

A given feature may have **hundreds** of chemical candidates to compare in order to identify likely candidates

NTA WebApp: Input Page (MS1 data)



NTA Data Input Files:

- Peak-picked, aligned MS1 data (CSV)
- Matrix of feature mass, RT, and sample abundances

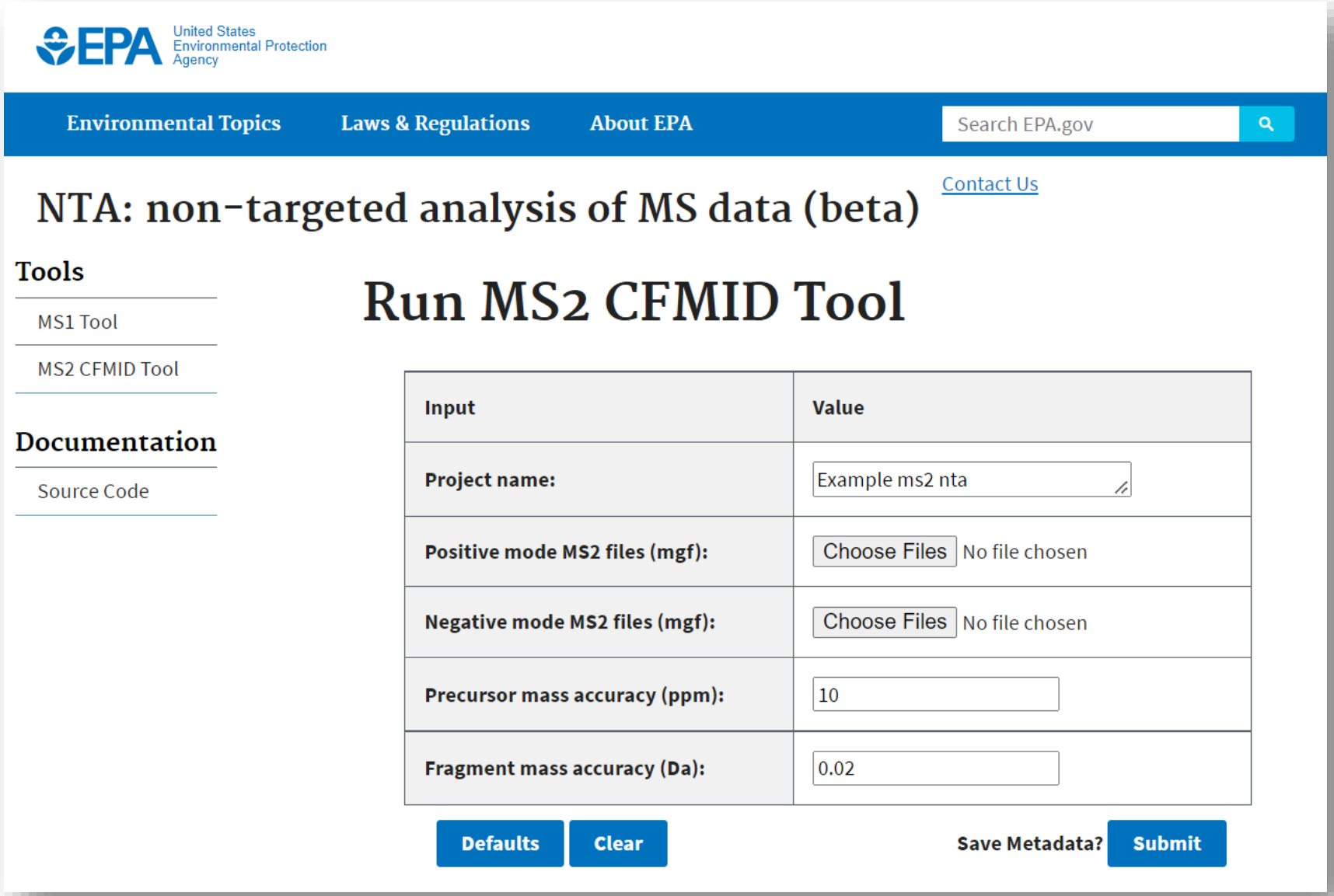
Tracer Input File

- Tracers: Isotopically labelled standards spiked into samples

Workflow Parameters:

- Data Processing
- Chemical retrieval

NTA WebApp: Input Page (MS2 data)



NTA Data Input Files:

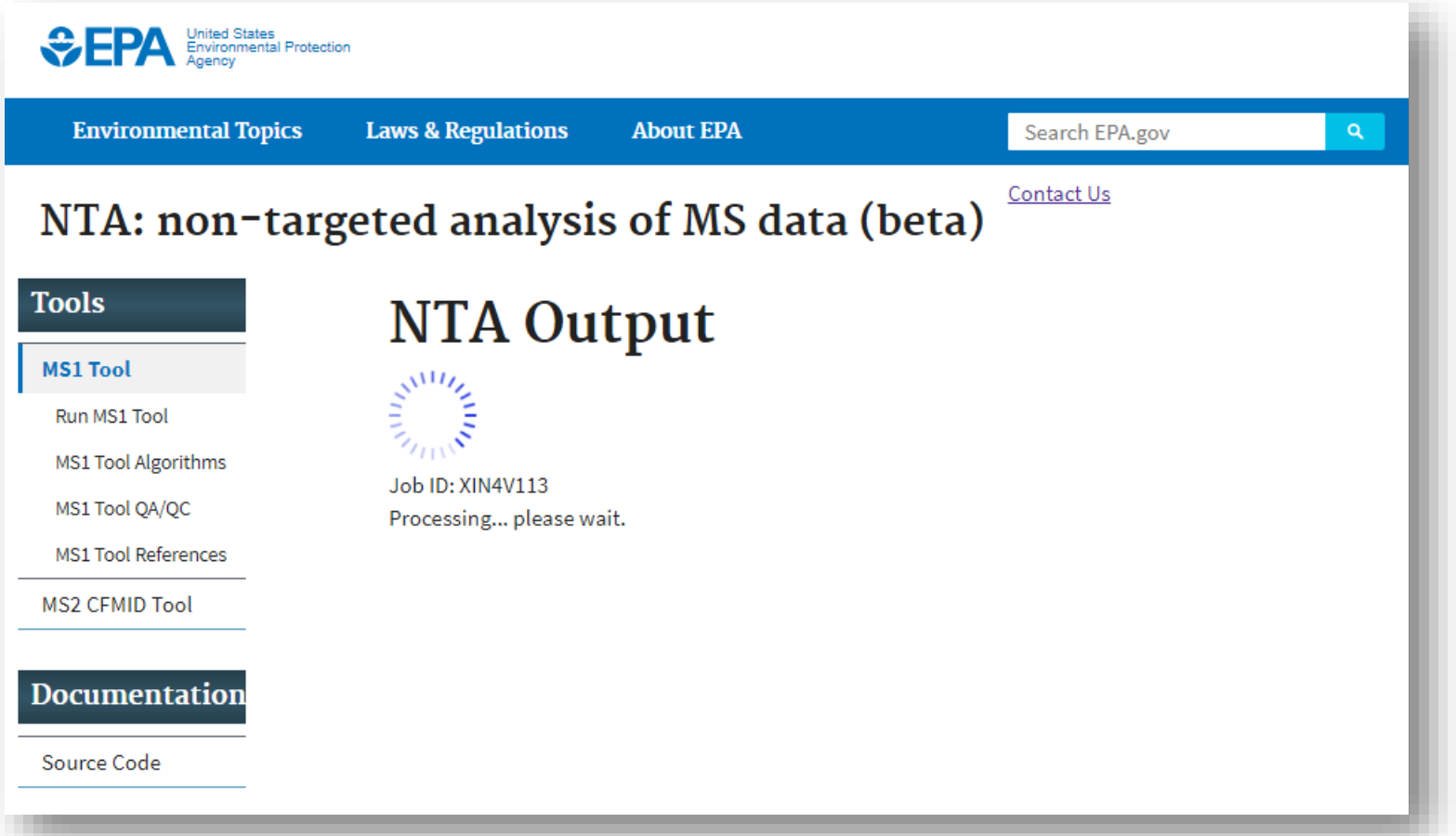
- Exported MS2 data (MGF format)
- Precursor mass, RT, fragment/intensity pairs

Workflow Parameters:

- Chemical retrieval
- Spectrum matching

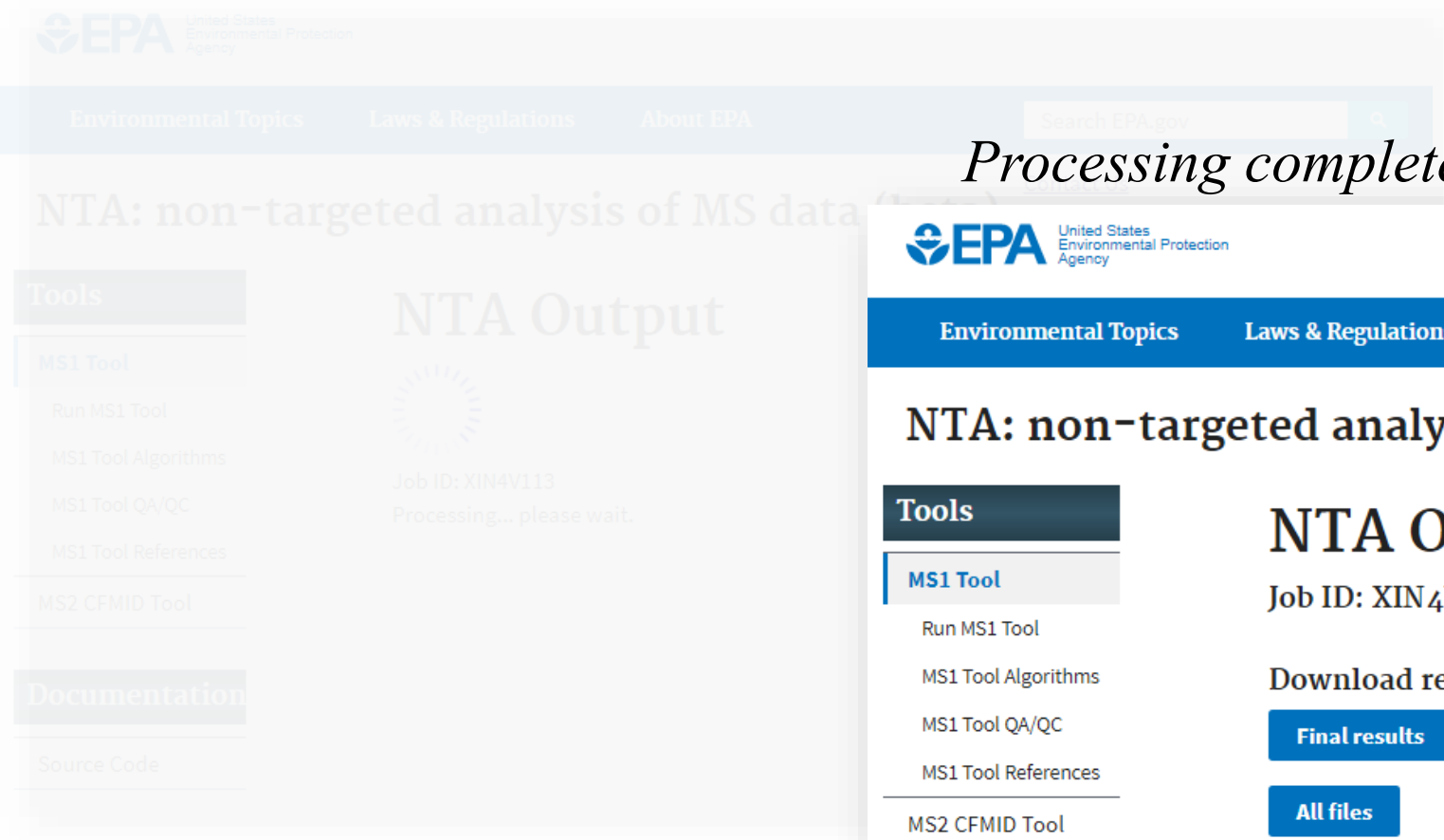
Submitting an NTA WebApp Job

In progress…

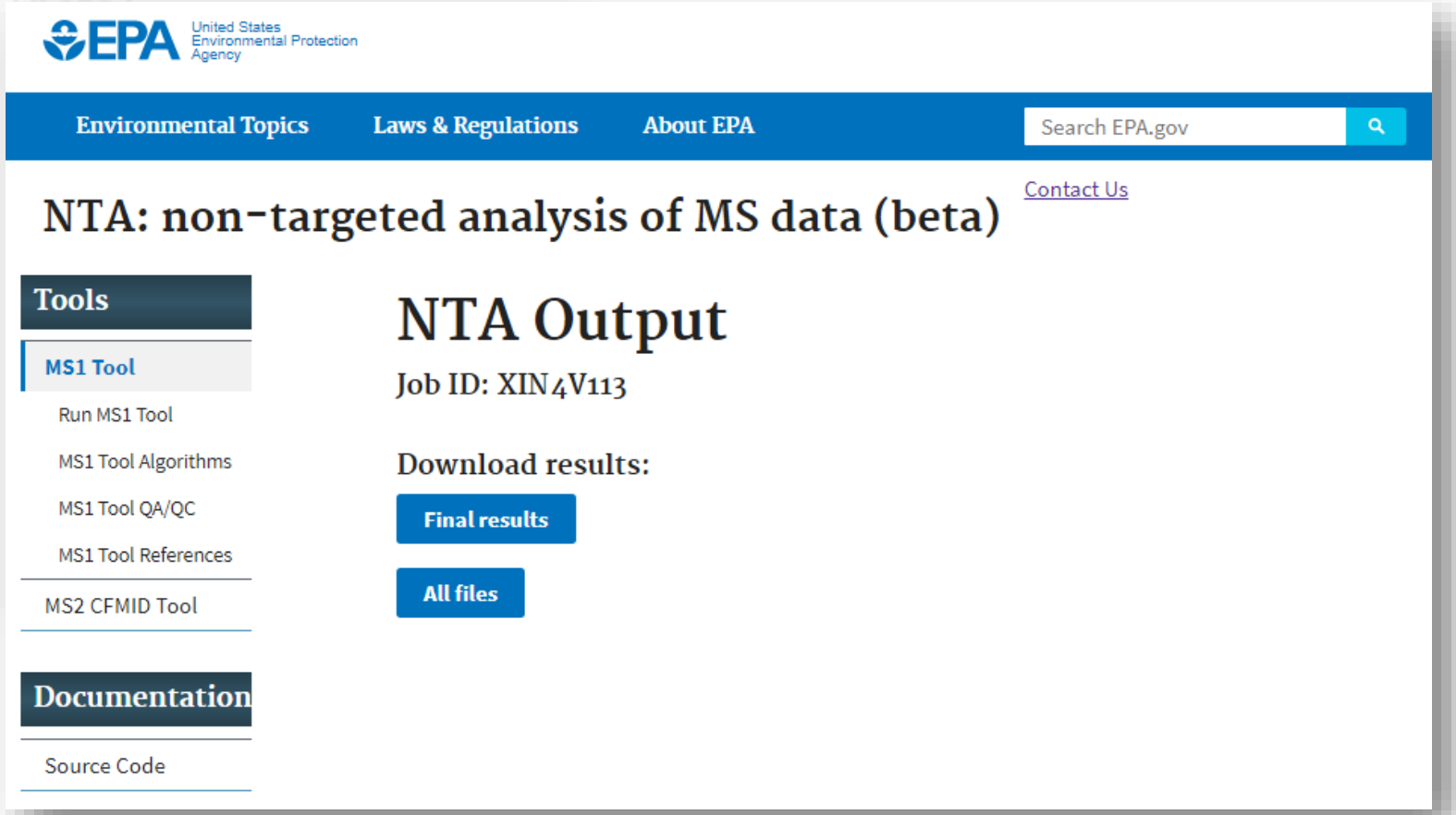


Submitting an NTA WebApp Job

In progress…



Processing complete and results ready for download



WebApp NTA Results Output Format

Feature Level Results

Feature ID	*Mass*	*Retention Time*	*Sample 1*	*Sample 2*	*Sample 3*
1	210.0876	6.904999	*Blank-subtracted median abundance values (QC filtered)*		
2	202.1223	7.808004			
3	670.5638	12.535			
4	706.5684	12.45099			
5	660.5236	12.16101			
6	616.4656	12.817			
7	278.147	9.584997			
8	216.1382	8.605996			
9	224.1037	7.854003			

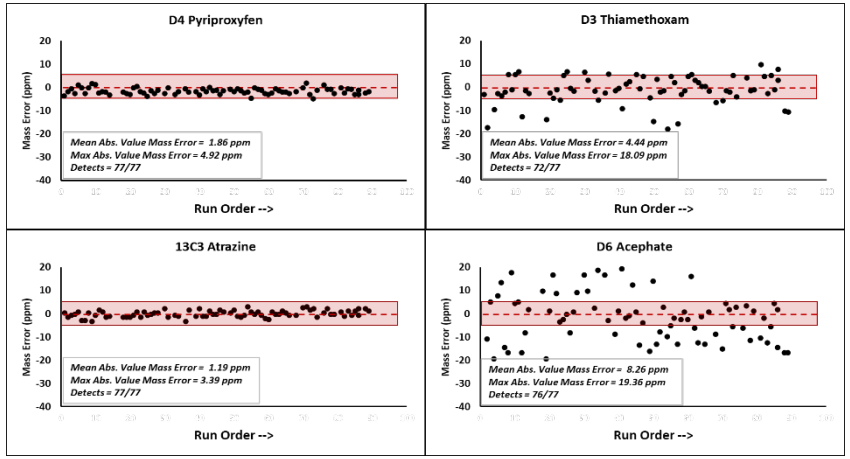
Chemical Level Results

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				
	Chemical Candidate 4				
2	Chemical Candidate 1				
	Chemical Candidate 2				
	Chemical Candidate 3				
	Chemical Candidate 4				
	Chemical Candidate 5				

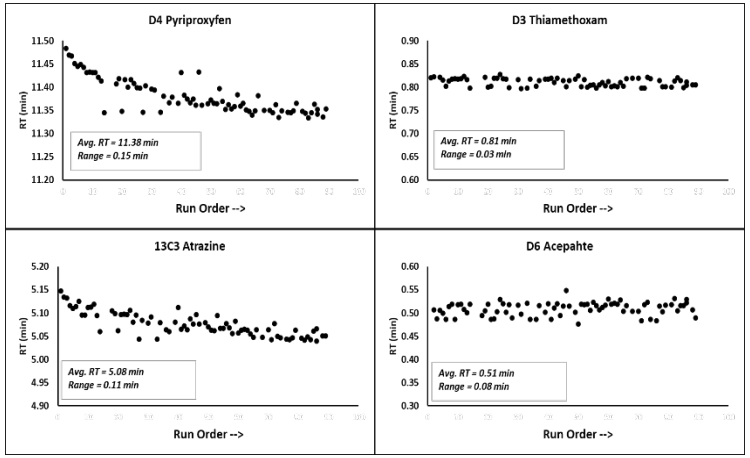
WebApp Tracer Results Output Format

Graphical Display of Tracer Results:

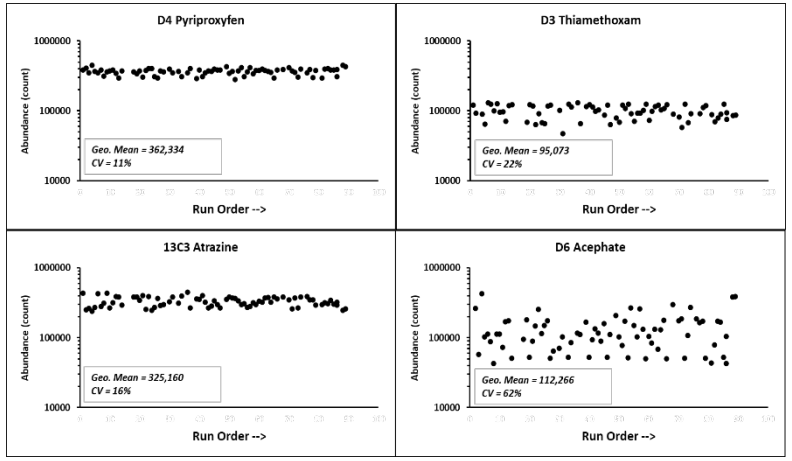
Mass Accuracy



Retention Time



Intensity

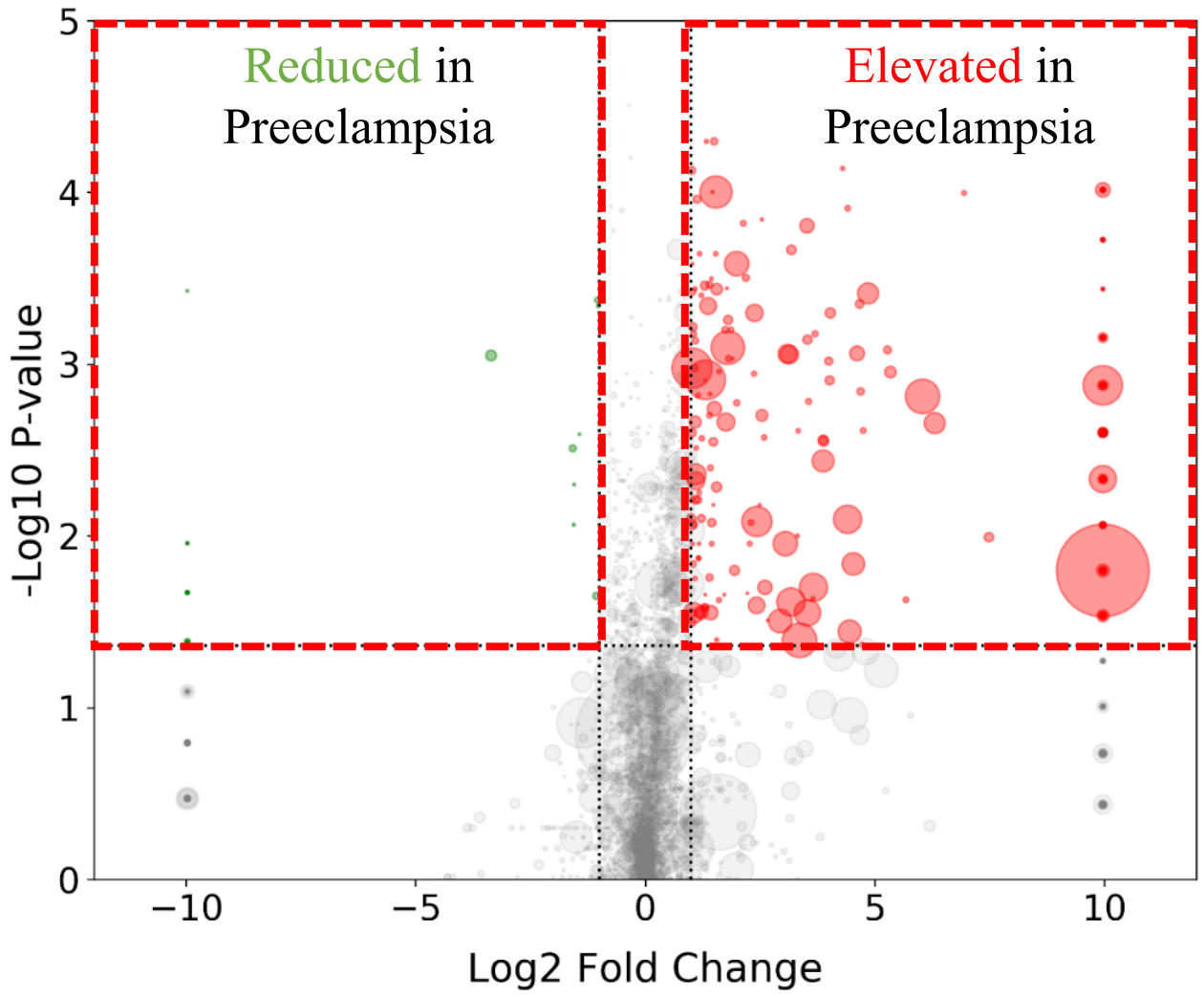


Spreadsheet of Tracer Results:

Tracer Compound	*Expected Mass*	*Expected Retention Time*	*Sample 1*	*Sample 2*	*Sample 3*	*Sample 4*	*Sample 5*	*Sample 6*	*Sample 7*	*Sample 8*	*Sample 9*
Compound 1	210.0876	6.90									
Compound 2	202.1223	7.81									
Compound 3	670.5638	12.54									
Compound 4	706.5684	12.45									

Observed retention time, mass, intensity values
Calculated mass/retention time error, intensity CV's

Use Case for Chemical Prioritization



NTA on placenta samples:

- Normotensive (n = 17) and preeclamptic (n = 18)
- **183 molecular features** found significantly different (~6000 potential candidates)
- Feature chemicals prioritized for targeted confirmatory work via:
  - Reference MS2 spectrum match
  - *In silico* MS2 spectrum match
  - Data source counts
  - Consumer product database presence (CPCat)
  - **46 chemicals** prioritized / acquired
- **25 chemicals** confirmed via targeted analyses

Chao, Alex, et al. (in revision).

What's Next?

- **Development of tools for improved NTA results**
  - Database incorporation of publicly available MS2 spectra
  - The Hazard Comparison Dashboard: database aggregation of publicly available toxicity data (*public version to be online shortly*)
  - Semi-quantitation methods: Generation of concentration estimates and uncertainty bounds from NTA data (*manuscript submitted*)
  - Compound method amenability predictions (*manuscript published*)
  - LC-MS retention time predictions (*manuscript published*)

Predicting compound amenability with liquid chromatography–mass spectrometry to improve non-targeted analysis

Charles N Lowe [1], Kristin K Isaacs [2], Andrew McEachran [3], Christopher M Grulke [2], Jon R Sobus [2], Elin M Ulrich [2], Ann Richard [2], Alex Chao [2], John Wambaugh [2], Antony J Williams [2]

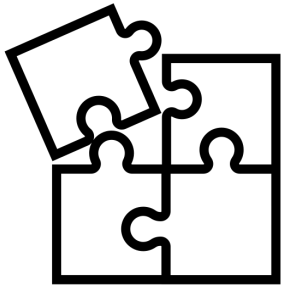
Development and Application of Liquid Chromatographic Retention Time Indices in HRMS-Based Suspect and Nontarget Screening

Reza Aalizadeh [1], Nikiforos A Alygizakis [1 2], Emma L Schymanski [3 4], Martin Krauss [5], Tobias Schulze [5], María Ibáñez [6], Andrew D McEachran [7], Alex Chao [7], Antony J Williams [7], Pablo Gago-Ferrero [8 9], Adrian Covaci [10], Christoph Moschet [11], Thomas M Young [11], Juliane Hollender [4 12], Jaroslav Slobodnik [2], Nikolaos S Thomaidis [1]

Lowe, Charles N., et al. Analytical and Bioanalytical Chemistry 413.30 (2021): 7495-7508.
Aalizadeh, Reza, et al. Analytical Chemistry 93.33 (2021): 11601-11611.

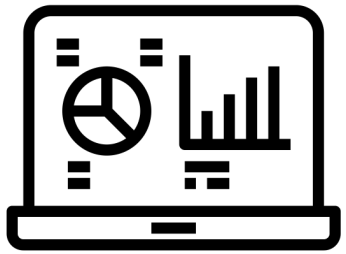
The NTA WebApp: Summary

- **The NTA WebApp is a synthesis of diverse work**
  - Multiple databases are integrated directly into workflow
  - More data = greater ability for chemical assignment and prioritization



- **The NTA WebApp is a step towards standardization**
  - Reproducible work through a tool containing entire workflow
  - Transparency and tracking of workflow

- **Upcoming developments**
  - New modules / databases in development for incorporation
  - *Manuscript in draft*: Public-facing version of WebApp to accompany publication (expected Fall 2022)



Acknowledgements

Hussein Al-Ghoul
Kathie Dionisio
Louis Groff
Jarod Grossman
Chris Grulke
Kristin Isaacs
Charles Lowe
James McCord
Andrew McEachran
Seth Newton
Allison Phillips
Katherine Phillips
Marie Russell
John Sloop
Elin Ulrich
John Wambaugh