Building a Non-Targeted Analysis Research Program at the U.S. EPA

Jon R. Sobus, Ph.D. & the EPA/ORD NTA Team

Center for Computational Toxicology and Exposure Research Triangle Park, NC
Current NTA Team

- Elin ‘Da Boss’ Ulrich
- Tony ‘Stark’ Williams
- ‘Dapper’ Charlie Lowe
- Scott ‘The Postman’ Clifton
- Jon ‘Nature Boy’ Sobus
- Alex ‘Can Do’ Chao
- Mark ‘Blue Steel’ Strynar
- James ‘Shake-n-Bake’ McCord
- Nelson ‘Prints’ Yeung
- Seth ‘Nice guy’ Newton
- Hannah ‘Dr. Cool’ Liberatore
- The Unflappable Ariel Wallace
- Tom ‘Mystery Man’ Purucker
- ‘Adventurin’ Jeff Minucci
Key Drivers for 21st Century Exposure Science

1) Understanding causes of disease

“…70-90% of disease risks are probably due to differences in environments”

2) Ensuring chemical safety

Although the risks of developing chronic diseases are attributed to both genetic and environmental factors, 70-90% of disease risks are probably due to differences in environments (1,2). Yet, epidemiologists increasingly use genome-wide association studies (GWAS) to investigate diseases, while relying on questionnaires to characterize "environmental exposures." This is because GWAS represent the only approach for exploring the totality of any risk factor genes. In this case, GWAS are associated with disease prevalence. Moreover, the value of cost-effective information is diminished when inaccurate and imprecise environmental data lead to biased inferences regarding gene-environment interactions (4). A more comprehensive and quantitative view of environmental exposure is needed if epidemiologists are to discover the major causes of chronic diseases. An obstacle to identifying the most important environmental exposures is the fragmentation of epidemiological research along lines defined by different factors. When epidemiologists investigate environmental risks, they tend to concentrate on a particular category of exposures involving air and water pollution, occupation, diet and obesity, stress and behavior, or types of infection. This slicing of the disease pie along parochial lines leads to scientific separation and confusion of the definition of "environmental exposures." In fact, all of these exposure categories can contribute to chronic diseases and should be investigated collectively rather than separately.

To develop a more coherent view of environmental exposure, it is important to recognize that toxic effects are mediated through chemicals that alter critical molecules, cells, and physiological processes inside the body. Thus, it would be reasonable to consider the "environment" as the body's internal chemical environment and "exposures" as the amounts of biologically active chemicals in the internal environment. Under this view, exposures are not restricted to chemicals (molecules) entering the body from air, water, or food, for example, but also include chemicals produced by inflammation, oxidative stress, lipid peroxidation, infections, gut flora, and other natural processes (5,6) (see the figure). This internal chemical environment continually fluctuates during life due to changes in external and internal sources, aging, infections, lifestyle, stress, psychological factors, and preexisting diseases. The term "exposome" refers to the totality of environmental exposures from conception onwards, and has been proposed to be a
High-Throughput Risk Characterization

- Many industrial & commercial chemicals are covered by the Toxic Substances Control Act (TSCA), which is administered by EPA.
- TSCA updated in June 2016 to allow risk-based evaluation of existing and new chemicals.
- Characterization of risk requires exposure and hazard data.
- EPA’s Office of Research and Development (ORD) is developing new approach methodologies (NAMs) for rapid risk characterization.
- NTA is a promising NAM, but requires careful evaluation and implementation.
NTA Research Produces Critical Data

All “…life-course environmental exposures (including lifestyle factors) from the prenatal period onwards…”

Figure adapted from: Rappaport SM. J Expo Sci Environ Epidemiol. 2011 Jan-Feb;21(1):5-9.
Our HRMS Tools of the Trade

- Agilent 6530B LC/Q-TOF
- Thermo GC/Q Exactive Hybrid Quad-Orbitrap
- Agilent 7250 GC/Q-TOF
- Coming Soon!!
- Agilent 6546 LC/Q-TOF
- Thermo LC/Orbitrap Fusion Tribrid
NTA Applications at EPA

- **Exposure surveillance**
  - What chemicals are in water, products, dust, blood, etc.?

- **Chemical prioritization**
  - What are relevant chemicals & mixtures?

- **Exposure forensics**
  - What are chemical signatures of exposure sources?

- **Biomarker discovery**
  - What chemicals are associated with health impairment?
Suspect Screening Analysis of Chemicals in Consumer Products


1 National Exposure Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States
2 Southwest Research Institute, San Antonio, Texas 78238, United States
3 Oak Ridge Institute for Science and Education (ORISE), Oak Ridge, Tennessee 37830, United States
4 National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States

19% of chemicals identified by NTA are on consumer product chemical lists.
Suspect screening and non-targeted analysis of drinking water using point-of-use filters

Seth R. Newton a,*, Rebecca L. McMahen a, b, Jon R. Sobus a, Kamel Mansouri b, c, 1, Antony J. Williams c, Andrew D. Meachran a, c, Mark J. Strynar a

a United States Environmental Protection Agency, National Exposure Research Laboratory, Research Triangle Park, NC 27711, United States
b Oak Ridge Institute for Science and Education Research Participant, 300 Terracon Drive, Research Triangle Park, NC 27709, United States
c United States Environmental Protection Agency, National Center for Computational Toxicology, Research Triangle Park, NC 27709, United States

Chemical Prioritization for Drinking Water

Top 20 Priority Compounds

Top 100 Priority Compounds

*Confirmed with standard
Exposure Forensics for Recycled Products

- Ubiquitous chemicals in articles (e.g., phthalates)
- Fragrances in recycled paper products
- Chemicals (variety of functions) that only occur in recycled tire products
- Fragrances in toys

Figure by C. Lowe and K. Isaacs
Biomarker Discovery for Placenta Samples

Different Environmental Exposures

The Placental Exposome (via LC-HRMS)

Altered Cell Signaling

29 in controls
508 in cases

Impaired Angiogenesis
Preeclampsia

Collaboration with J. Rager (UNC Chapel Hill) and J. Grossman (Agilent)
### NTA Best Practices

<table>
<thead>
<tr>
<th>Name</th>
<th>Example</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracers</td>
<td>Isotopically labeled standards: $^{13}$C$_3$-Atrazine, $^{13}$C$_4$, $^{15}$N$_2$-Fipronil</td>
<td>Allows tracking of chromatographic performance and mass accuracy</td>
</tr>
<tr>
<td>Replication</td>
<td>Triplicate injections of same sample vial</td>
<td>Removes risk of “one hit wonder”</td>
</tr>
<tr>
<td>Run order randomization</td>
<td>8, 3, 7, 4, 2, 1, 10, 5, 8, 6, 9, 2, 5, 4, 1, 9, 4, 7, 3, 8, 1, 6, 10, 9, 6, 7, 5, 3, 2, 10</td>
<td>Minimizes/averages out batch or sample order effects (e.g., carryover, temp &amp; instrument drift)</td>
</tr>
<tr>
<td>Pooled QC sample</td>
<td>Combine 5 mg/µL from each of 10 samples (total 50 mg/µL) prior to extract to create pooled QC</td>
<td>Separate confirmation of presence with different matrix, MS2 IDs</td>
</tr>
<tr>
<td>Blanks</td>
<td>Solvent, method, matrix, double blanks</td>
<td>Allows identification/subtraction/deletion of interferences introduced in lab processes</td>
</tr>
<tr>
<td>Multiple lines of evidence for ID</td>
<td>RT prediction/matching, spectra prediction/matching, data source ranking, functional/product uses, media occurrence</td>
<td>Improves confidence in identification when chemicals standards are unavailable</td>
</tr>
</tbody>
</table>
Agilent LC/Q-TOF Simplified Workflow

**Experimental Acquisition**
- Sample Extracts
  - LC/Q-TOF HRMS
    - MS^2 Acquisition
      - MS^2 .d Files
        - MS^2 .mgf Files
    - MS^1 Acquisition

**DB & Library Matching**
- Chemical Database
  - DB MS-Ready Structures
    - DB MS-Ready Formula & Monoisotopic Mass
      - Reference MS^2 Spectra
        - in silico MS^2 Spectra
  - Reference MS^2 Spectra

**Data Analysis**
- MS^1 Feature Table
  - Filtered Feature Table
  - Chemical Candidate Table
  - Aggregated Match Table
- MS^2 Reference Matches
  - MS^2 in silico Matches
Agilent LC/Q-TOF Simplified Workflow

**Experimental Acquisition**

- **Sample Extracts**
  - **LC/Q-TOF HRMS**
    - **MS² Acquisition**
      - **MS² .d Files**
    - **MS¹ Acquisition**
      - **MS² .mgf Files**

**DB & Library Matching**

- **Chemical Database**
  - **DB MS-Ready Structures**
    - **DB MS-Ready Formula & Monoisotopic Mass**
      - **Reference MS² Spectra**
        - **in silico MS² Spectra**
          - **MS² in silico Matches**

**Data Analysis**

- **MS¹ Feature Table**
  - **Filtered Feature Table**
    - **Chemical Candidate Table**
      - **Aggregated Match Table**
        - **MS² Reference Matches**
          - **MS² in silico Matches**
Experimental Acquisition

1,269 Substances in 10 Mixtures

Agilent 6530B Q-TOF

RP vs. HILIC
ESI vs. APCI

None 148
APCI- 19
ESI- 37
APCI+ 22
ESI+ 119

3 80 25
13 53
0 187 11
20 64
463
Agilent LC/Q-TOF Simplified Workflow

Experimental Acquisition
- Sample Extracts
  - LC/Q-TOF HRMS
    - MS\(^2\) Acquisition
      - MS\(^2\) .d Files
      - MS\(^2\) .mgf Files
    - MS\(^1\) Acquisition

DB & Library Matching
- Chemical Database
  - DB MS-Ready Structures
    - DB MS-Ready Formula & Monoisotopic Mass
      - Reference MS\(^2\) Spectra
        - in silico MS\(^2\) Spectra
      - MS\(^2\) .d Files
      - MS\(^2\) .mgf Files
  - MS\(^1\) Feature Table
    - Filtered Feature Table
    - Chemical Candidate Table
    - Aggregated Match Table
    - MS\(^2\) Reference Matches
      - MS\(^2\) in silico Matches

Data Analysis
Chemical Database = DSSTox
MS-Ready Structures

Spiked Substance: Tamoxifen
DTXSID1034187

Spiked Substance: Tamoxifen citrate
DTXSID8021301

Predicted Formula for Observed Molecular Feature: C_{26}H_{28}NO
Dashboard Search
1st: DTXSID1034187
2nd: DTXSID8021301

MS-Ready Processing
Dashboard Access

The CompTox Chemistry Dashboard can be used by mass spectrometists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formula searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as "MS ready structures" and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN, DTXTID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys. (UPDATED APRIL 2019)

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

Andrew D. McEachran,2,3, Kamei Mansouri,2,3, Chris Grulke, Emma L. Schymanski, Christoph Ruttkies and Antony J. Williams
Feature Removal:
1) Duplicate features
2) Non-reproducible features
3) Blank features (sample:blank)
4) Non-responsive features (dilutions)

Feature Flagging:
1) Multi-mode hits (+ and -)
2) Meas. precision (CV threshold)
3) Formula match (score ≥ threshold)
4) Negative mass defect
5) Halogenation
6) Has/is adduct
7) Has/is neutral loss
8) Has/is multimer

Dashboard Integration:
1) Data source & pub counts
2) Bioactivity & exposure levels
3) Presence on lists
4) Product & use categories
Agilent LC/Q-TOF Simplified Workflow

**Experimental Acquisition**
- Sample Extracts
  - LC/Q-TOF HRMS
    - MS\(^2\) Acquisition
    - MS\(^1\) Acquisition
      - MS\(^2\) .d Files
      - MS\(^2\) .mgf Files

**DB & Library Matching**
- Chemical Database
  - DB MS-Ready Structures
    - DB MS-Ready Formula & Monoisotopic Mass
      - Reference MS\(^2\) Spectra
      - in silico MS\(^2\) Spectra
        - MS\(^2\) Reference Matches
        - MS\(^2\) in silico Matches

**Data Analysis**
- MS\(^1\) Feature Table
  - Filtered Feature Table
  - Chemical Candidate Table
    - Aggregated Match Table
Generation of *in silico* Spectra

**CFM-ID v2.0**

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

**Authors:** Andrew D. McEachran, Ilya Balabin, Tommy Cathey, Thomas R. Transue, Hussein Al-Ghoul, Chris Grulke, Jon R. Sobus & Antony J. Williams

**Training Set:** Metlin MS2 spectra and structures

**Machine Learning**

**Fragmentation Prediction Model**

**DSSTox MS-Ready Structures** (~765,000)

**DSSTox MS2 spectra** (10, 20, 40v)


CFM-ID Database Matching

1. Query database by mass

- MGF file
- Exp MS2 Spectrum (Mass = 356.119)

2. Score in silico spectra

<table>
<thead>
<tr>
<th>Candidate</th>
<th>CFM-ID Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 1</td>
<td>0.5</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>0.2</td>
</tr>
<tr>
<td>Candidate 3</td>
<td>0.1</td>
</tr>
</tbody>
</table>
CFM-ID Database Matching (w/ Formula Information)

1. **Query database by mass**

   - MGF file
   - Exp MS2 Spectrum (Mass = 356.119)
   - CFM-ID Database

Retrieval candidate compounds within mass window

Filter candidates by formula

- **C_{19}H_{20}N_{2}O_{3}S** (Mass = 356.119)
- **C_{21}H_{21}ClO_{4}** (Mass = 356.118)

2. **Score in silico spectra**

   - **In silico MS2 Spectra** (CE 10, 20, 40)

   - **CFM-ID Scores**

<table>
<thead>
<tr>
<th>Candidate</th>
<th>in silico CE 10</th>
<th>in silico CE 20</th>
<th>in silico CE 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>0.2</td>
<td>0.1</td>
<td>0.02</td>
</tr>
<tr>
<td>Candidate 3</td>
<td>0.1</td>
<td>0.05</td>
<td>0.01</td>
</tr>
</tbody>
</table>
CFM-ID Database Matching (w/ Multiple CE_{experimental})

1. Query database by mass

- MGF file
- Exp MS2 CE10 Spectrum (Mass = 356.119)
- Exp MS2 CE20 Spectrum (Mass = 356.119)
- Exp MS2 CE40 Spectrum (Mass = 356.119)

CFM-ID Database

Retrieve candidate compounds within mass window

Candidate 1
C_{19}H_{20}N_{2}O_{3}S
(Mass = 356.119)

In silico MS2 Spectra
(CE 10, 20, 40)

Candidate 2
C_{19}H_{20}N_{2}O_{3}S
(Mass = 356.119)

In silico MS2 Spectra
(CE 10, 20, 40)

Candidate 3
C_{21}H_{21}ClO_{3}
(Mass = 356.118)

In silico MS2 Spectra
(CE 10, 20, 40)

2. Score in silico spectra

<table>
<thead>
<tr>
<th>CFM-ID Scores</th>
<th>in silico CE 10</th>
<th>in silico CE 20</th>
<th>in silico CE 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 1 CE_{exp} = 10</td>
<td>0.5</td>
<td>0.3</td>
<td>0.1</td>
</tr>
<tr>
<td>Candidate 1 CE_{exp} = 20</td>
<td>0.4</td>
<td>0.5</td>
<td>0.12</td>
</tr>
<tr>
<td>Candidate 1 CE_{exp} = 40</td>
<td>0.05</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
CFM-ID Scoring Approaches

Approach 1
Precursor 1
Experimental Spectrum at CE 10

Predicted Spectrum at CE 10
Candidate 1
Predicted Spectrum at CE 20
Predicted Spectrum at CE 40
Score = A

Approach 2
Experimental Spectrum at CE 10

Score = A+B+C

Approach 3
Experimental Spectrum at CE 20

Score = A+B+C

Experimental Spectrum at CE 40

Score = G+H+I

Score = A+B+C+D+E+F+G+H+I
EPA’S Non-Targeted Analysis Collaborative Trial

The Trial Mixtures:

10 Mixtures ranging from 95 to 365 compounds
(Total: 1,269 unique compounds)

“Pass” compounds = 377 with MS2 data

EPA Setup:

Agilent 1290 UPLC
Agilent 6530B Q-TOF with ESI source
Reference vs. *in silico* Library Coverage

PCDL \(\rightarrow\) Agilent reference MS\(^2\) library

“Pass” compounds (n=377) \(\rightarrow\) ENTACT chemicals observed with MS\(^2\) data

<table>
<thead>
<tr>
<th>MS2 Library</th>
<th>% of “Pass” Compounds Identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent PCDL</td>
<td>53%</td>
</tr>
<tr>
<td>CFM-ID Top Hit</td>
<td>50%</td>
</tr>
<tr>
<td>PCDL and/or CFM-ID Top Hit</td>
<td>73%</td>
</tr>
</tbody>
</table>

PCDL

88

111

77

CFM-ID

101

"Pass" Compounds

PCDL \(\rightarrow\) Agilent reference MS\(^2\) library

“Pass” compounds (n=377) \(\rightarrow\) ENTACT chemicals observed with MS\(^2\) data
NTA Workflows: Using CFM-ID Results as Filters

Score
Filter out candidates below score cutoff

Variability in score distribution

Rank
Filter out candidates above rank cutoff

Variability in number of candidate compounds

MS2 Spectrum 1
Score
Filter out candidates below score cutoff

Variability in score distribution

MS2 Spectrum 2
Score
Filter out candidates below score cutoff

Variability in score distribution

MS2 Spectrum 1
Rank
Filter out candidates above rank cutoff

Variability in number of candidate compounds

MS2 Spectrum 2
Rank
Filter out candidates above rank cutoff

Variability in number of candidate compounds

Filter by Top 20

n = 10

n = 500
## Normalizing CFM-ID Results Values

### Score Quotient
Normalize score to the highest candidate compound score

### Score Percentile
Normalize rank to the number of candidate compounds

<table>
<thead>
<tr>
<th>Rank</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
<th>Score Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
<td>20</td>
</tr>
</tbody>
</table>

Score Quotient = Score / Maximum Score
NTA Workflows: Using CFM-ID Normalized Results as Filters

**Score Quotient**
Filter out candidates below score quotient cutoff

*Score quotient cutoff = 0.5*
*Keep candidates scoring at least half of max score*

**Score Percentile**
Filter out candidates below percentile cutoff

*Score percentile cutoff = 0.5*
*Keep the top 50% of candidates*
## Applying Cut-off Filters to Data

<table>
<thead>
<tr>
<th>Candidate Compound</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Compound 2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Compound 3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
</tr>
<tr>
<td>Compound 4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Compound 5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

![Scatter plot showing data points representing various scores and quotients](chart.png)
Applying Cut-off Filters to Data

<table>
<thead>
<tr>
<th>Candidate Compound</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Compound 2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Compound 3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
</tr>
<tr>
<td>Compound 4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Compound 5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- True Compound
- Other Candidate Compounds

<table>
<thead>
<tr>
<th></th>
<th>True Positives</th>
<th>False Negatives</th>
<th>True Negatives</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score Quotient</td>
<td>0</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>False Positives</td>
<td>1</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>True Positives</td>
<td>0</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>False Negatives</td>
<td>1</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>True Negatives</td>
<td>0</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>False Positives</td>
<td>1</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>True Positives</td>
<td>0</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>False Negatives</td>
<td>1</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>True Negatives</td>
<td>0</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
<tr>
<td>False Positives</td>
<td>1</td>
<td>0.8</td>
<td>0.78</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Applying Cut-off Filters to Data

<table>
<thead>
<tr>
<th>Candidate Compound</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Compound 2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Compound 3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
</tr>
<tr>
<td>Compound 4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Compound 5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

True Compound

Other Candidate Compounds

<table>
<thead>
<tr>
<th></th>
<th>True Positives</th>
<th>False Negatives</th>
<th>True Negatives</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Score Quotient

Cut-off = 0
## Applying Cut-off Filters to Data

<table>
<thead>
<tr>
<th>Candidate Compound</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- **True Compound**: 1
- **False Negatives**: 0
- **True Negatives**: 2
- **False Positives**: 2

Score Quotient Cut-off = 0.5
Applying Cut-off Filters to Data

<table>
<thead>
<tr>
<th>Candidate Compound</th>
<th>CFM-ID Score</th>
<th>Maximum Score</th>
<th>Score Quotient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound 1</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>Compound 2</td>
<td>0.4</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Compound 3</td>
<td>0.39</td>
<td>0.5</td>
<td>0.78</td>
</tr>
<tr>
<td>Compound 4</td>
<td>0.1</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Compound 5</td>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Score Quotient: Cut-off = 0.9

True Compound
Other Candidate Compounds

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives</td>
<td>0</td>
</tr>
<tr>
<td>False Negatives</td>
<td>1</td>
</tr>
<tr>
<td>True Negatives</td>
<td>3</td>
</tr>
<tr>
<td>False Positives</td>
<td>1</td>
</tr>
</tbody>
</table>
Balancing Cut-offs

True Positive Rate (TPR) = \( \frac{TP}{TP + FN} \)

False Positive Rate (FPR) = \( \frac{FP}{FP + TN} \)

How many of the true compounds are we keeping?

How much of the junk are we getting rid of?
Quotient Vs. Percentile Cutoffs

Global ROC Curves (All ENTACT Mixtures)
Quotient Vs. Percentile Cutoffs

Global ROC Curves (All ENTACT Mixtures)

- True Compounds
- Other Candidate Compounds

Score Quotient

Score Quotient Cut-off
Quotient Vs. Percentile Cutoffs

Global ROC Curves (All ENTACT Mixtures)

True Compounds
Other Candidate Compounds

Score Quotient Cut-off

- True Compounds
- Other Candidate Compounds
Quotient Vs. Percentile Cutoffs

Global ROC Curves (All ENTACT Mixtures)

True Compounds
Other Candidate Compounds

Score Quotient Cut-off

- Score Quotient

True Compounds
Other Candidate Compounds
Quotient Vs. Percentile Cutoffs

Cut-off Values for Global TPR = 0.9

<table>
<thead>
<tr>
<th></th>
<th>Cut-off value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quotient (by formula)</td>
<td>0.18</td>
</tr>
<tr>
<td>Percentile (by formula)</td>
<td>38</td>
</tr>
<tr>
<td>Quotient (by mass)</td>
<td>0.13</td>
</tr>
<tr>
<td>Percentile (by mass)</td>
<td>32</td>
</tr>
</tbody>
</table>

Apply to individual ENTACT mixtures
CFM-ID Cut-off Filtering: Individual ENTACT Mixtures
EPA/ORD NTA activities:

• Focused on applications
  – qualitative (to date) → semi-quantitative (soon)
  – must support HT exposure prediction & risk evaluation

• R&D required to support applications
  – Experimental + cheminformatic + computational efforts = Viable NTA program

• Growing capacity with new instrumentation
  – Requires flexible workflows
    ▪ Work smarter, not harder
    ▪ Don’t reinvent the wheel
    ▪ Build once, use many (A. Williams)
Contributing Researchers

**EPA ORD**
Hussein Al-Ghoul*
Alex Chao*
Jarod Grossman*
Kristin Isaacs
Sarah Laughlin*
Charles Lowe
James McCord
Jeff Minucci
Seth Newton
Katherine Phillips
Tom Purucker
Randolph Singh*
Mark Strynar
Elin Ulrich

**EPA ORD (cont.)**
Chris Grulke
Kamel Mansouri*
Andrew McEachran*
Ann Richard
John Wambaugh
Antony Williams

**Agilent**
Jarod Grossman
Andrew McEachran

**GDIT**
Ilya Balabin
Tom Transue
Tommy Cathey

* = ORISE/ORAU

This work was supported, in part, by ORD’s Pathfinder Innovation Program (PIP) and an ORD EMVL award.
Questions?

sobus.jon@epa.gov

The views expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.