EPA United States Environmental Protection Agency

Cross-Laboratory Comparison of Non-Targeted Analysis Method Performance using Complex Synthetic Mixtures

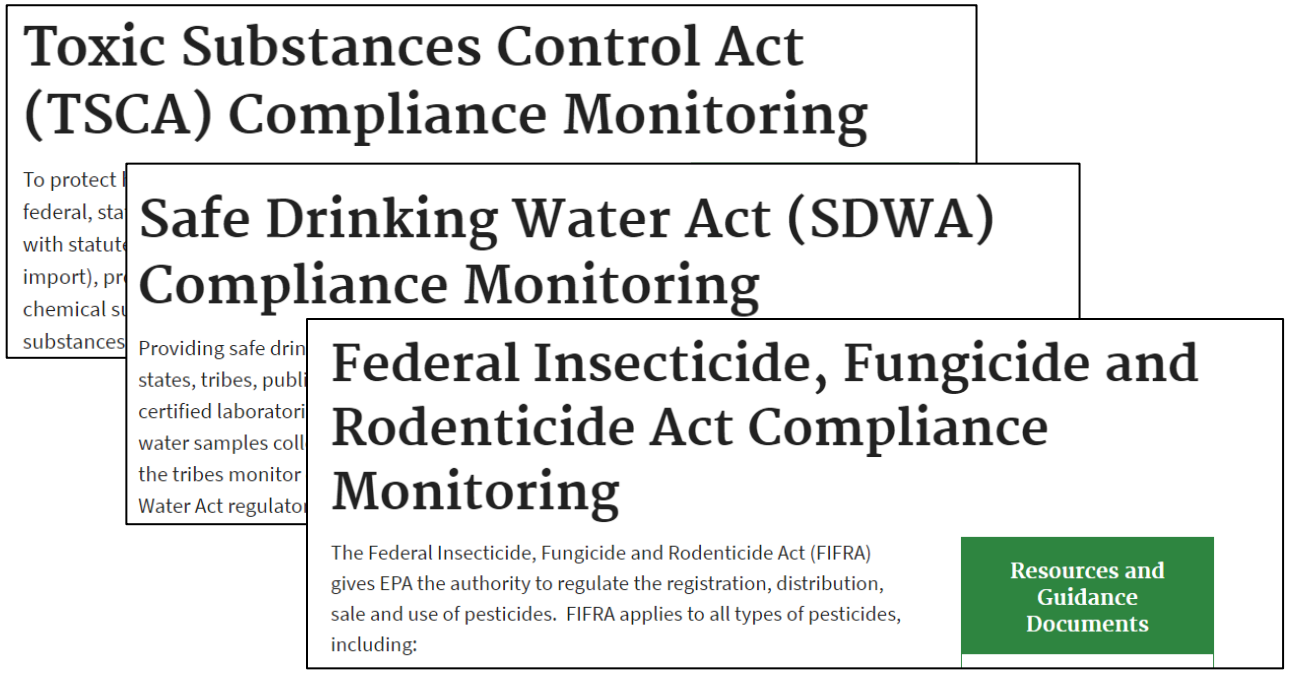
Jon Sobus[1], Katherine Phillips[1], Jarod Grossman[2], Alex Chao[1], Antony Williams[1], Chris Grulke[1], Ann Richard[1], Andrew McEachran[2], Elin Ulrich[1]

[1] Center for Computational Toxicology and Exposure
[2] ORAU/ORISE Participant

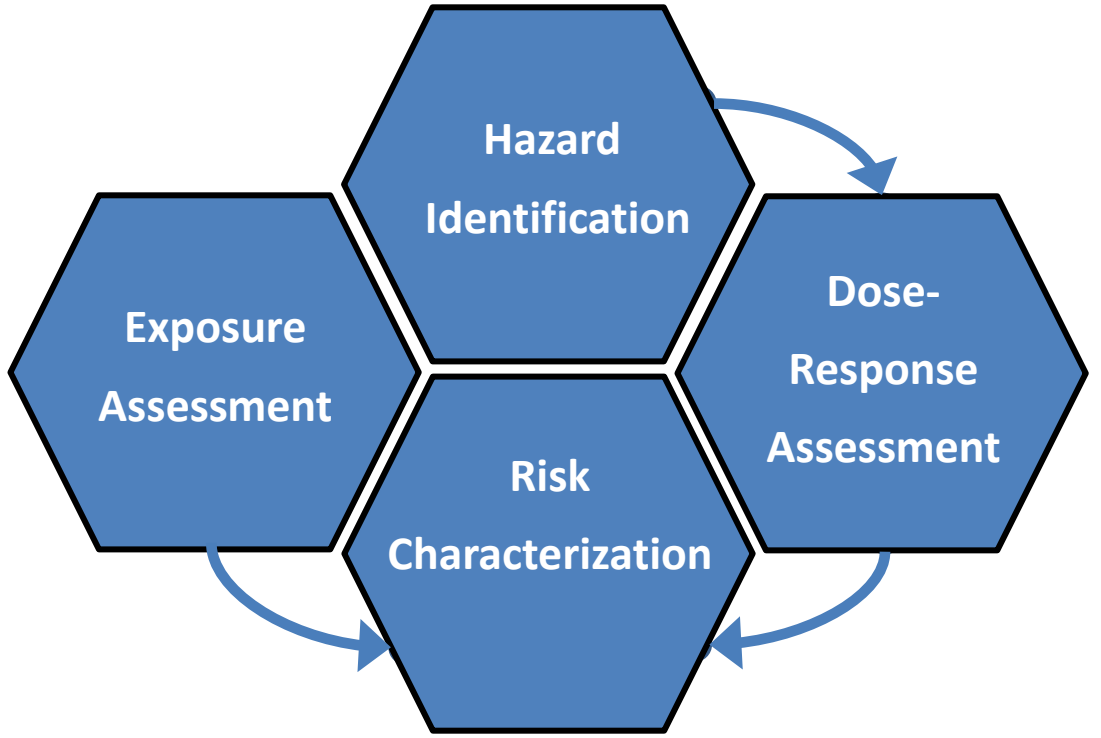
Office of Research and Development

Why Does EPA Need Measurement Data?

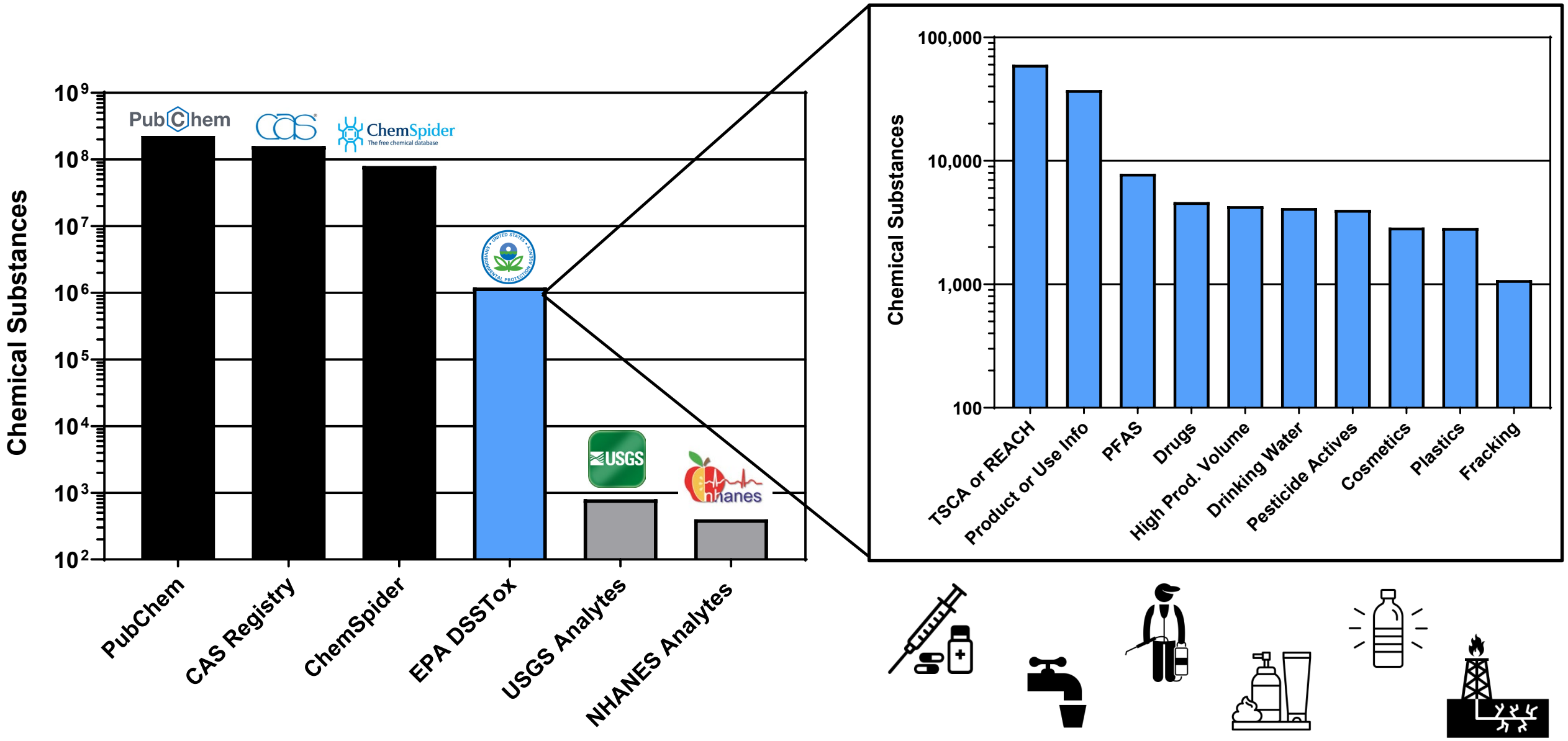
- **Measurement data needed to ensure chemical safety**
  - Characterize risk
  - Regulate use & disposal
  - Manage human & ecological exposures
  - Ensure compliance under federal statutes



Chemical Monitoring Needs



Data Disparity: Have vs. Need



Challenges

- High-quality exposure data are unavailable for most chemicals
- Measurement data traditionally generated using “targeted” methods
- Targeted analytical methods:
  - Require *a priori* knowledge of chemicals of interest
  - Produce data for few selected analytes (10s-100s)
  - Require standards for method development & compound quantitation
  - Are blind to emerging contaminants
  - Can’t keep pace with the needs of 21st century chemical safety evaluations

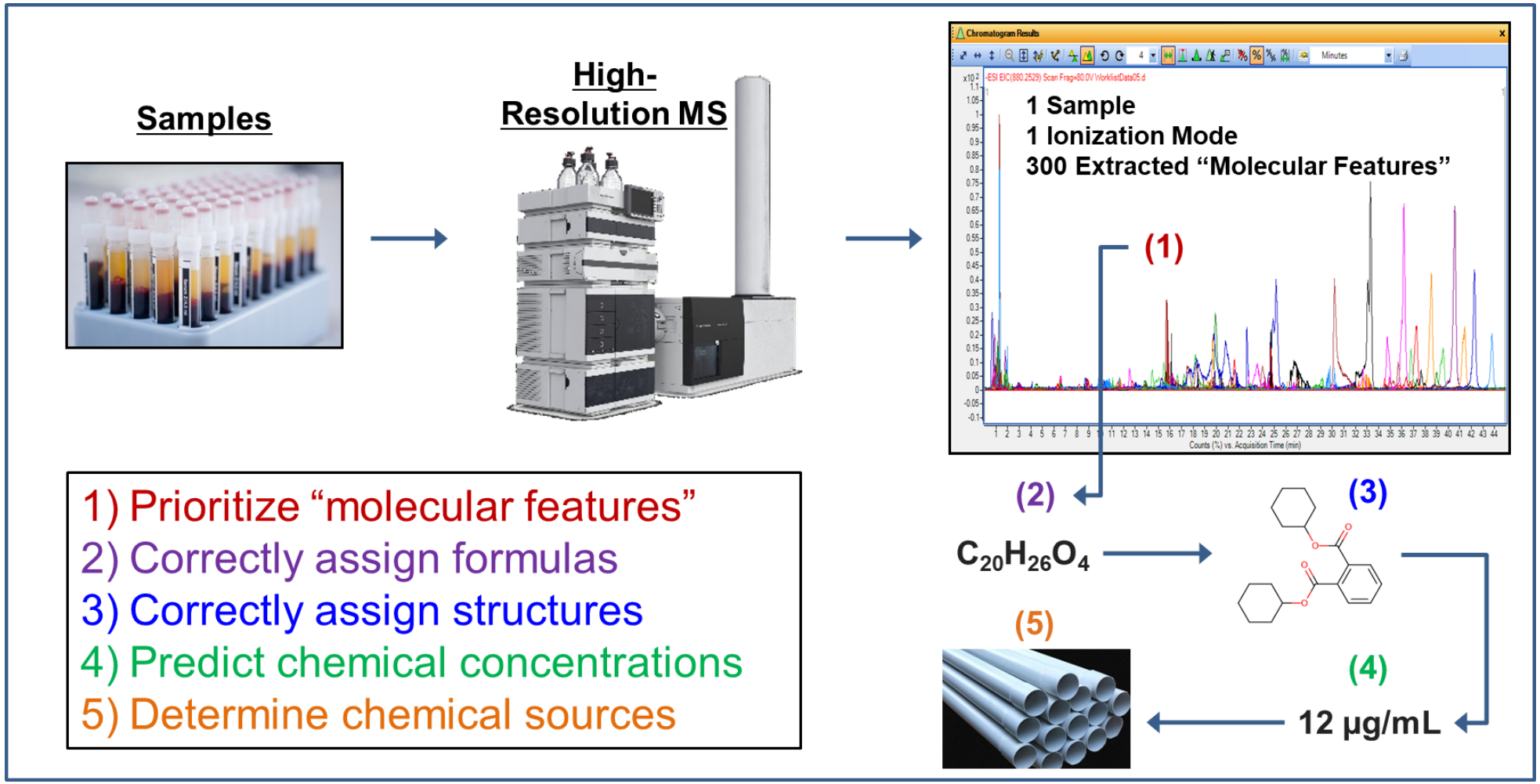
What's So Great About NTA?

Rapidly screen for "knowns"

Discover "unknowns"

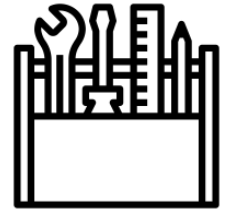
Uncover historical exposures

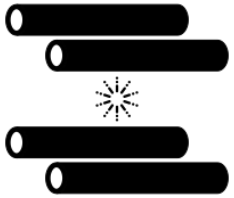
Generate source fingerprints…

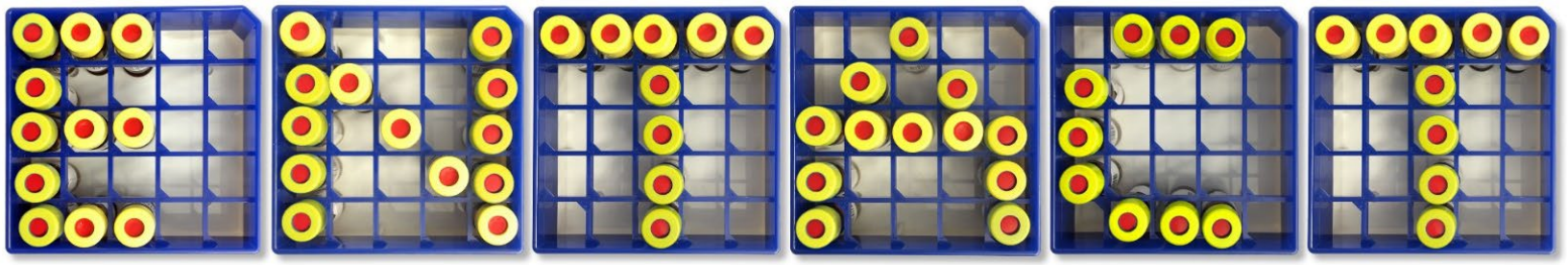


Science Questions for Research Community

- How variable are tools and results from lab to lab?
- Are some methods/workflows better than others?
- How does sample complexity affect performance?
- What chemical space does a given method cover?
- How sensitive are specific instruments/methods?







EPA's Non-Targeted Analysis Collaborative Trial

EPA United States Environmental Protection Agency

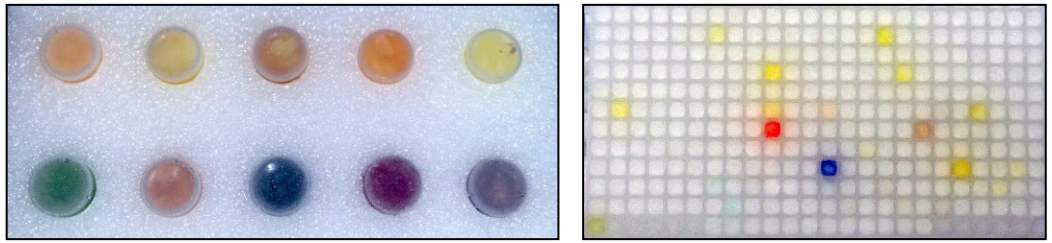
ENTACT Part 1

Chemicals from ToxCast Library

~1200 ToxCast Chemicals (highest quality)

10 Mixtures (100-400 chemicals each)

*Multi-Well Plates**



~25 Collaborators & 6 Contractors*:

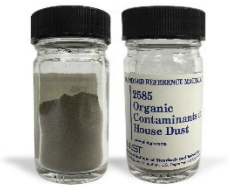
1st: Blinded analysis

2nd: Unveiling of chemicals

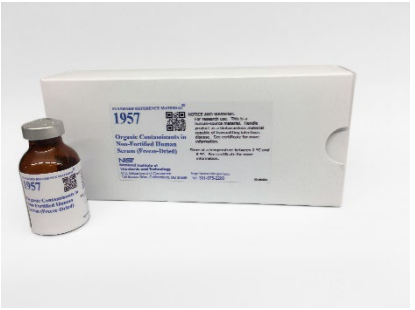
3rd: Unblinded evaluation

ENTACT Part 2

Reference & Fortified House Dust

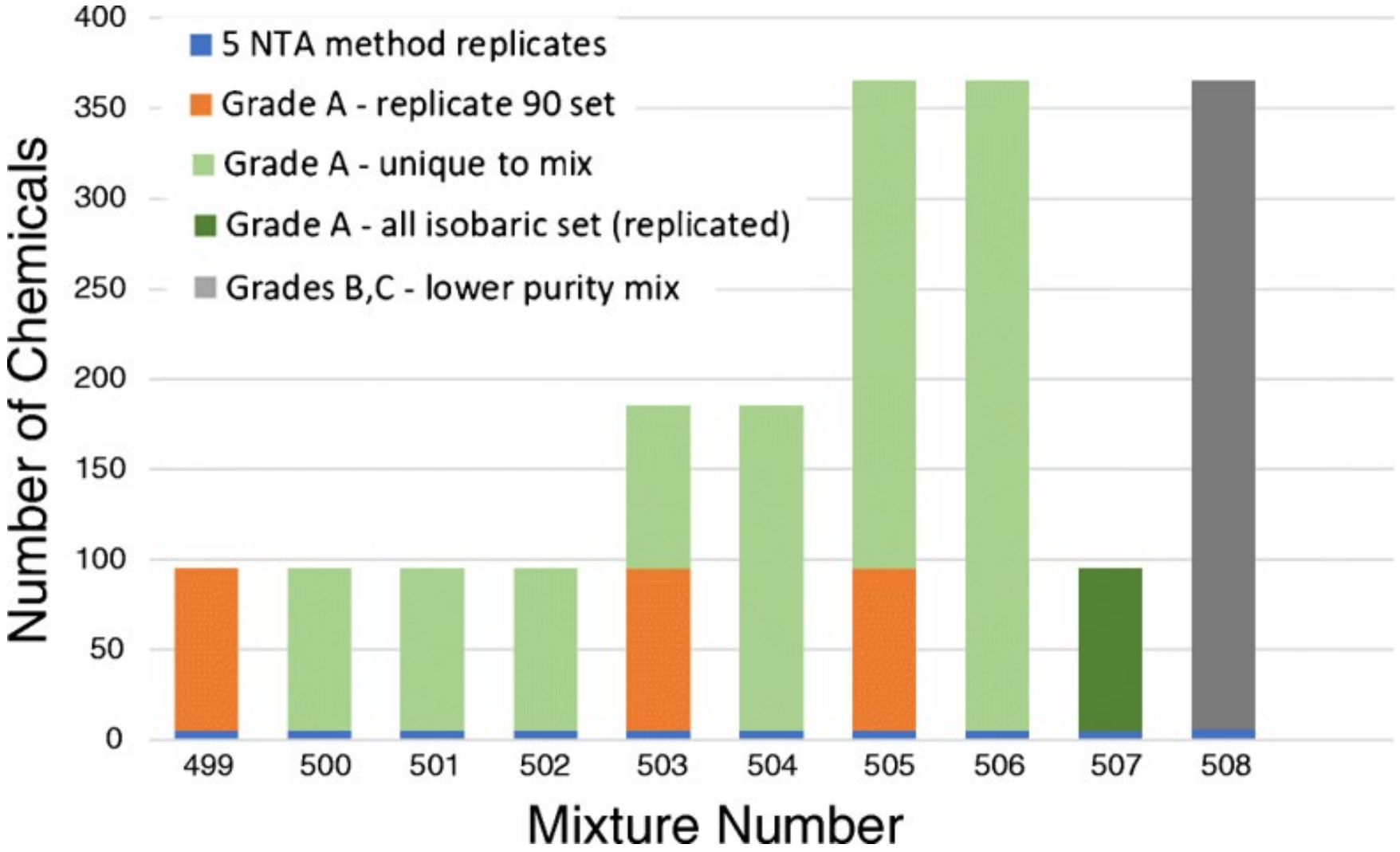


Reference & Fortified Human Serum



Reference & Fortified Silicone Wristbands

Design of ENTACT Mixtures



Replication in substance spikes offers a unique means to assess NTA method reproducibility!

Ulrich et al. 2019. doi: 10.1007/s00216-018-1435-6

Who is Working on ENTACT?

Contractors:

University of Alberta
Emory University
Duke University
Oregon State University
San Diego State University
Pacific Northwest National Laboratory

19 Blind submissions

15 Unblinded submissions

Vendors:

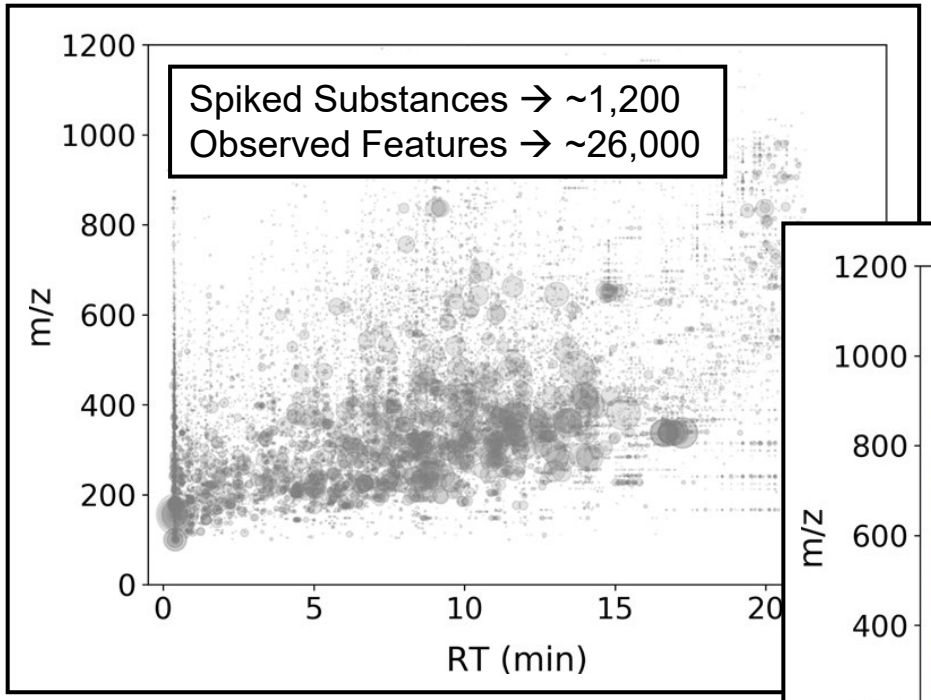
The Scripps Research Institute
Agilent Technologies
Thermo Scientific
Leco
AB SCIEX
Waters

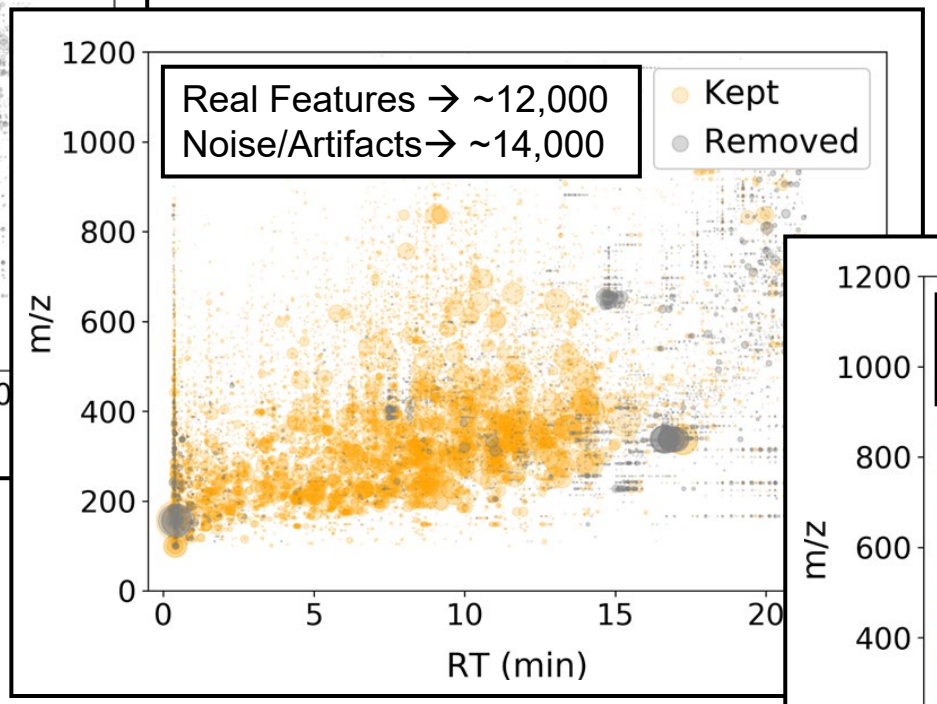
General Participants:

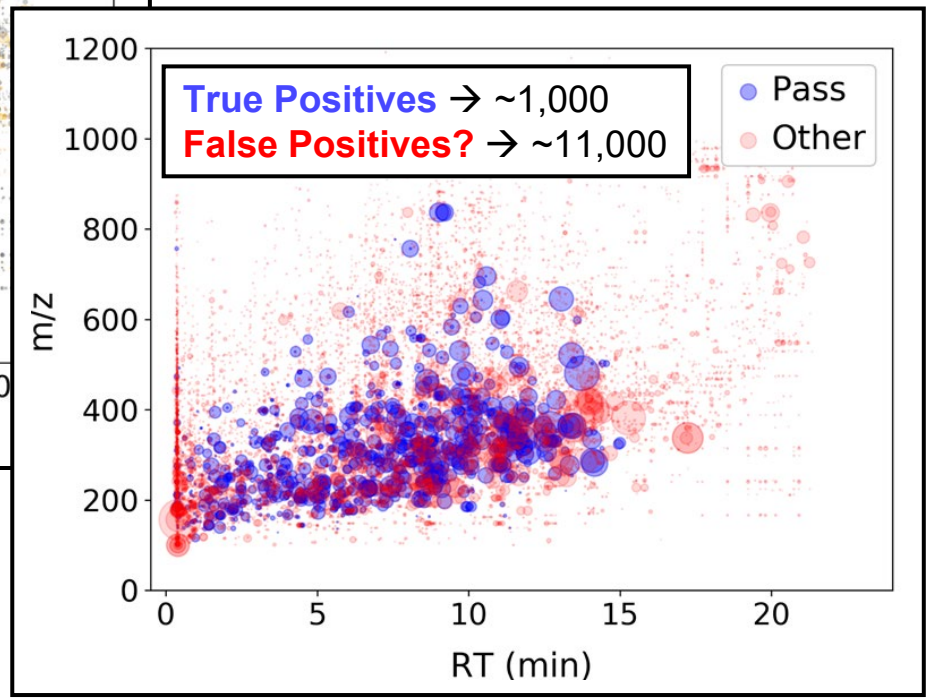
Université du Luxembourg
University of Florida
Wisconsin State Laboratory of Hygiene, University of Wisconsin-Madison
NIST
NC State University
Recetox
Cornell University
UW Tacoma
Mount Sinai
University of Birmingham
DTSC Department of Toxic Substances Control
eawag aquatic research
United States Environmental Protection Agency
FIU

EPA Lab Results for ENTACT Mixtures

LC-QTOF HRMS
(ESI+ and ESI-)



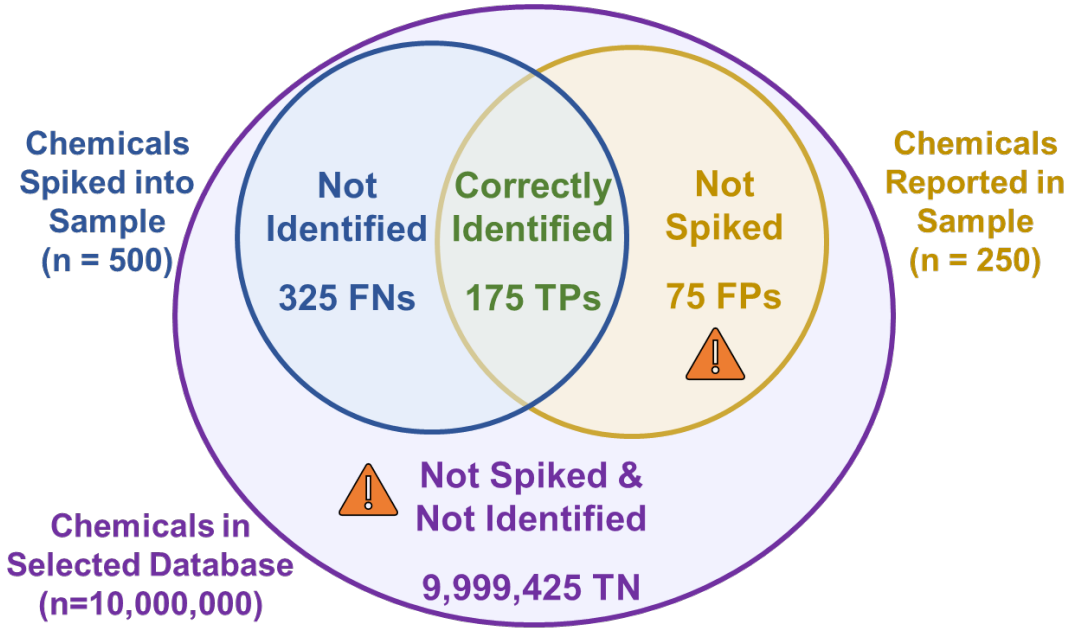




Substance Identified?	Substance Spiked? Yes	Substance Spiked? No
Yes	True Positives (≤ 65%)	False Positives?
No	False Negatives (≥ 35%)	True Negatives?

Sobus et al. 2019. doi: 10.1007/s00216-018-1526-4

Evaluation Tools Must Be Used With Caution



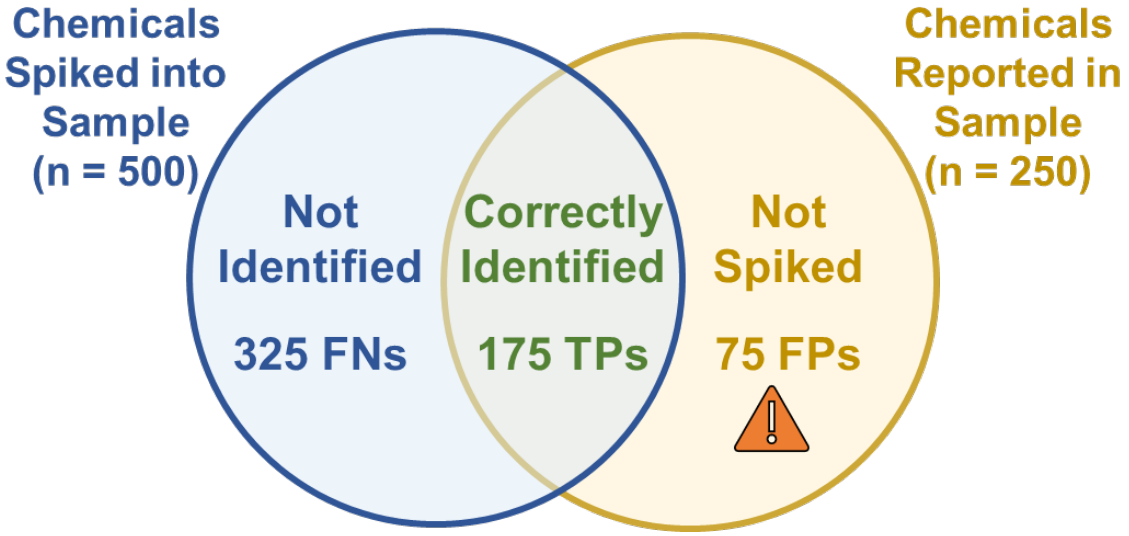
Boundary n = 10,000,000		Chemical is… spiked into sample	Chemical is… not spiked into sample		
Chemical is…	reported in sample	TP 175	FP 75	Precision 0.70	FDR 0.30
	not reported in sample	FN 325	TN 9,999,425		
		TPR 0.35	FPR 0.00001	F_1 0.47	Accuracy 0.99996
		FNR 0.65	TNR 0.99999	MCC 0.49	

Fisher et al. 2022. doi: 10.1007/s00216-022-04203-3

A hypothetical example

- How do we differentiate FPs from unintentional TPs?
- How do we appropriately handle TNs?

A (Slightly) Simpler Scenario



Boundary n = 575		Chemical is… spiked into sample	Chemical is… not spiked into sample	
Chemical is…	reported in sample	TP 175	FP 75	Precision 0.70 ⚠ FDR 0.30
	not reported in sample	FN 325		
		TPR 0.35		F_1 0.47 ⚠
		FNR 0.65		

Fisher et al. 2022. doi: 10.1007/s00216-022-04203-3

- Still have challenges with FP interpretation
- What about lower-level “hits”? (e.g., Level 4)
- What about ID reproducibility?

Processing ENTACT Data Submissions

- Individual methods treated separately (if appropriate)
- One candidate mass/formula/compound per feature
- Confidence level revised as needed (with consensus)
- Matching to spiked substances by mass, formula & structure
- "**Observed**" if structure or formula (no spiked isomers) match
- "**Correctly Identified**" if structure match
- "**Reproducible**" if correctly ID'd >50% of the time
  - "Eligible" compounds spiked >1 time and identified ≥1 time

Forward vs. Reverse Evaluation

Forward

Spiked Compound	Observed?	Correctly ID'd	Reproducibly ID'd
1	Yes	Yes	Yes
2	Yes	Yes	No
3	No	No	--
4	No	No	--
5	Yes	Yes	Yes
6	Yes	No	--
7	Yes	Yes	No
8	No	No	--
9	Yes	Yes	--
10	No	No	--
…	…	…	…
100	Yes	Yes	Yes

Reverse

Reported Compound	Spiked?
1	No
2	Yes
3	Yes
4	Yes
5	No
6	No
7	Yes
8	No
9	Yes
10	Yes
…	…
125	No

Outlining Utilized Performance Metrics

- **Observability Rate** = # Observed / # Spiked
  - If it was spiked, could your instrument detect it?
- **True Positive Rate** = # Correctly Identified / # Spiked
  - If it was spiked, could your workflow correctly ID it?
- **Correct ID Rate** = # Correctly Identified / # Observed
  - If it was observed, could your workflow correctly ID it?
- **Reproducibility Rate** = # Reproducible / # Eligible
  - If it was correctly ID'd once, was it correctly ID'd most of the time?

Forward Evaluation Metrics

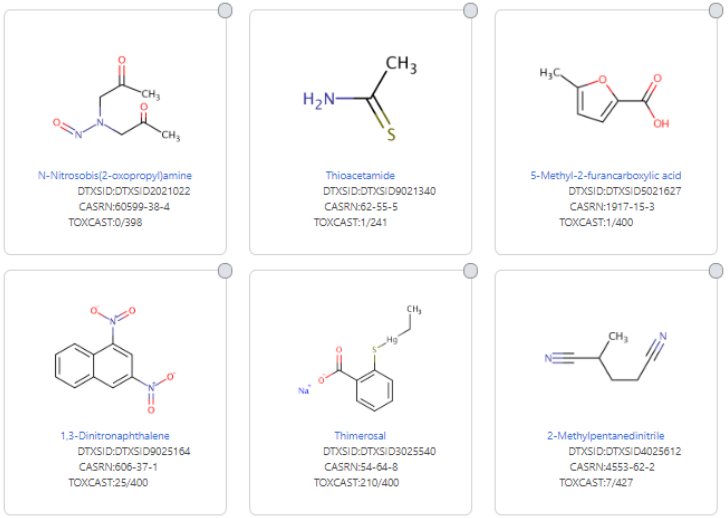
- **Reporting Rate** = # Reported / # Spiked
  - What is the ratio of reported to spiked compounds?
- **Correct Reporting Rate** = # Correctly Identified / # Reported
  - If it was reported, was it a correctly identified spiked compound?

Reverse Evaluation Metrics

Method Comparison: “Observed” Compounds

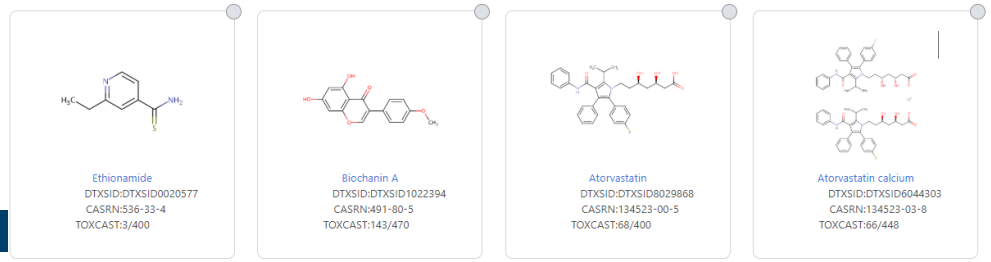
7 Labs, 12 Methods

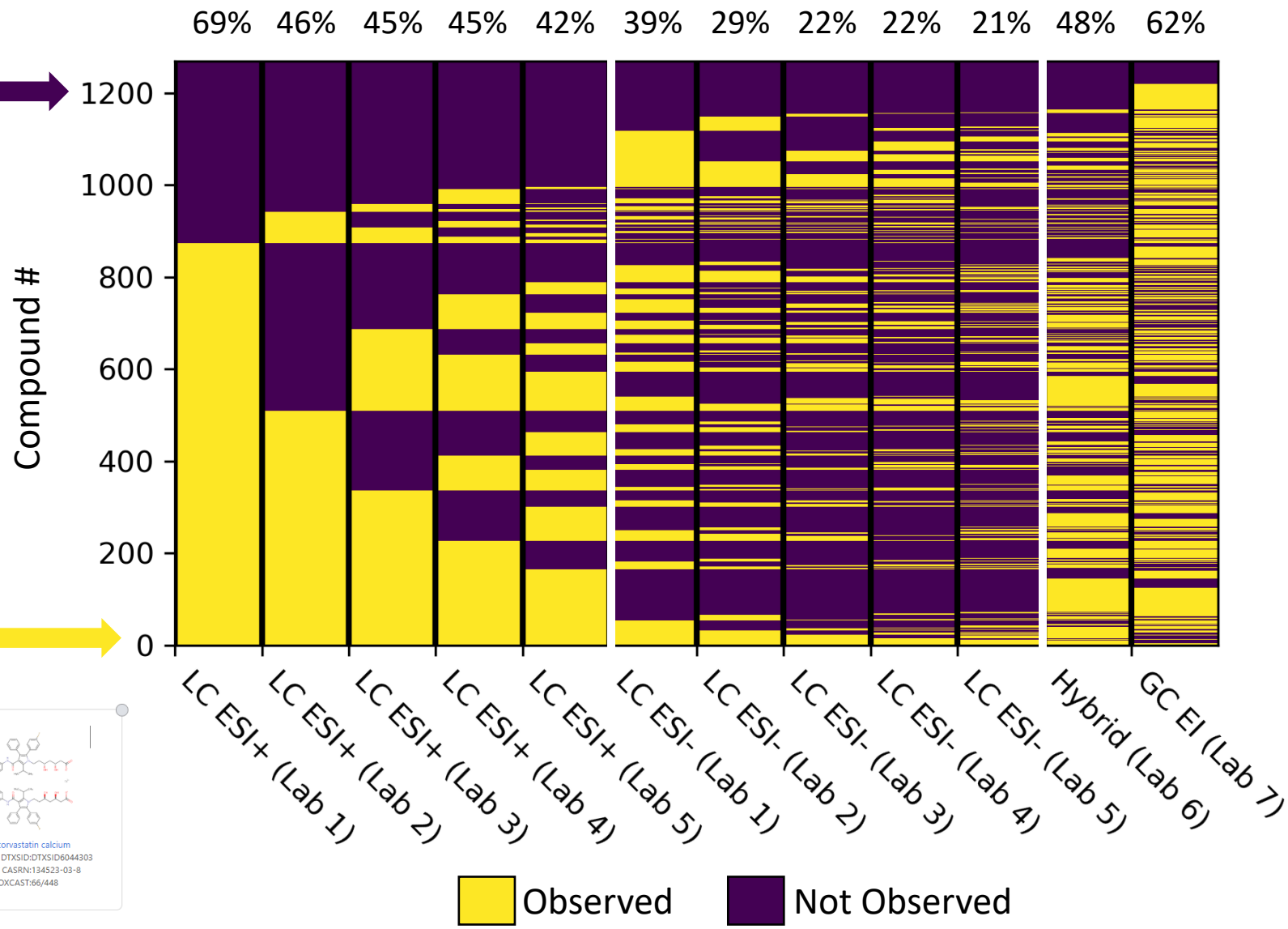
~5% Not Observed by Any Method



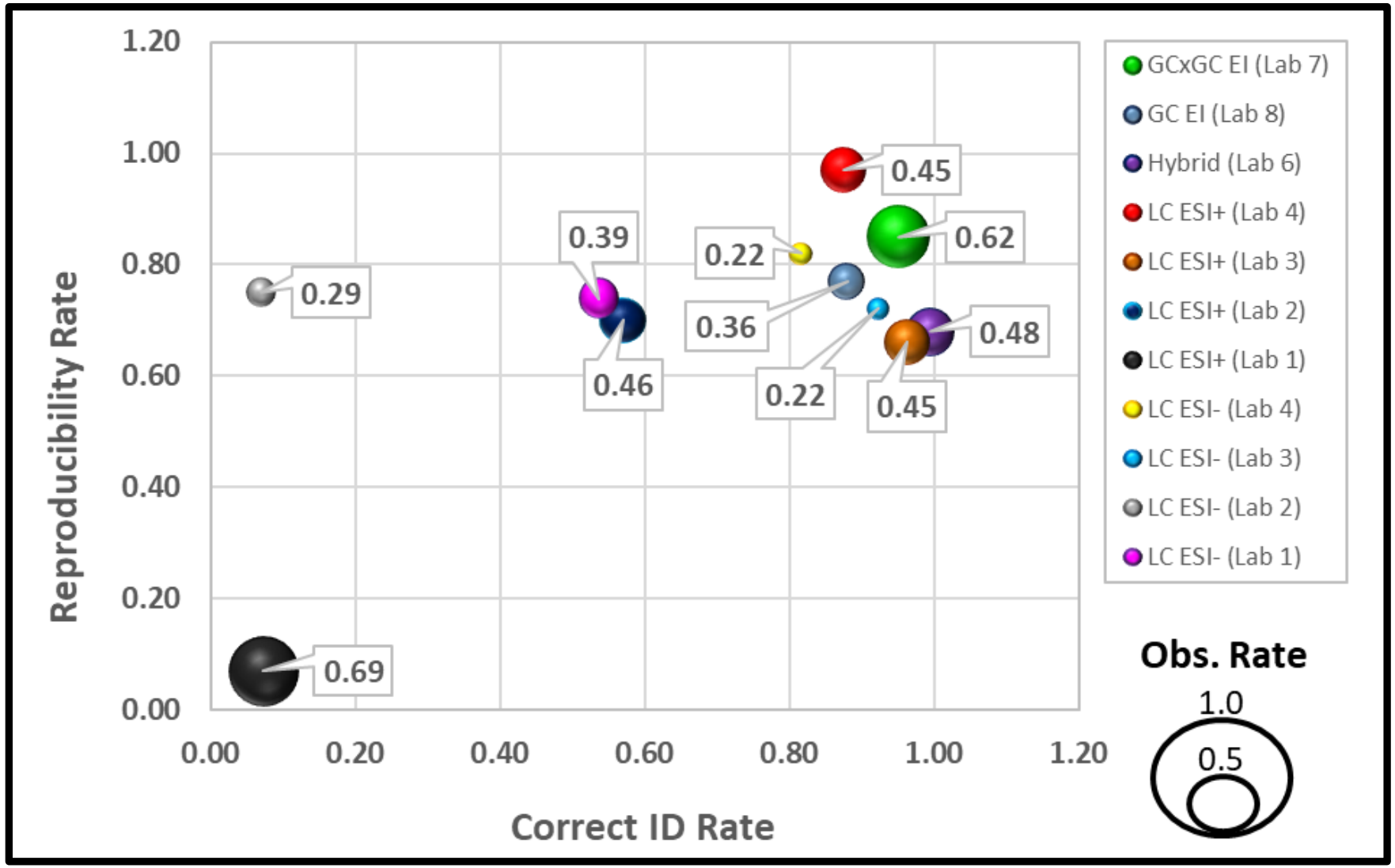
...

<1% Observed by All 12 Methods





Method Comparison: 3 Forward Metrics



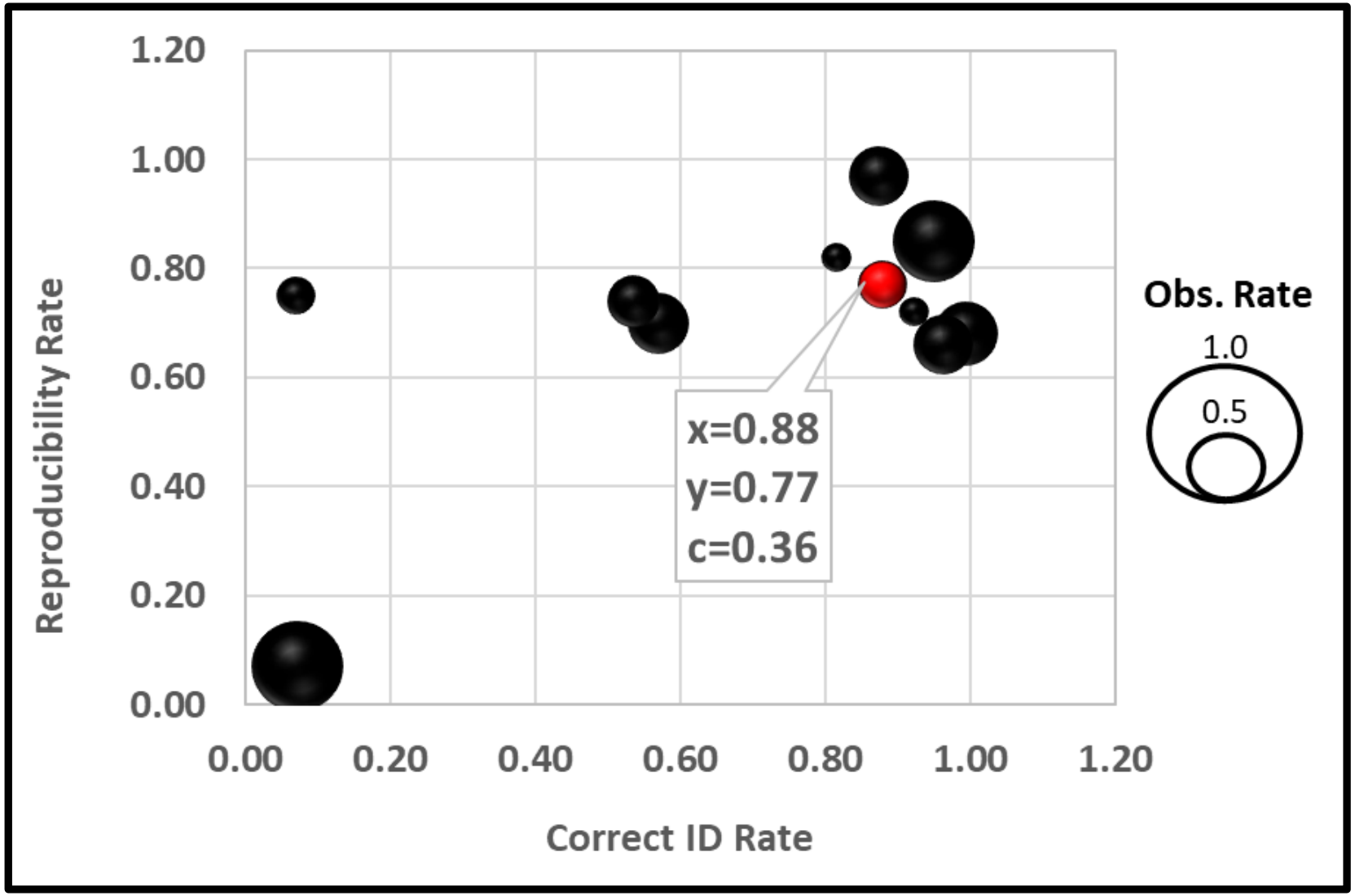
Metrics (all %):

Bubble Size →
How often observed?

X-Axis →
How often correctly ID'd if observed?

Y-Axis →
How consistently ID'd?

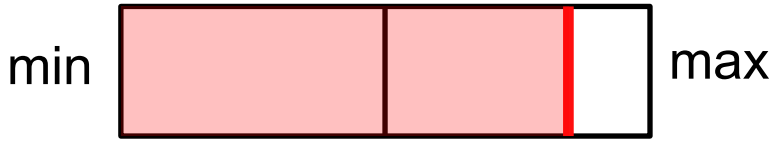
Example Performance Report



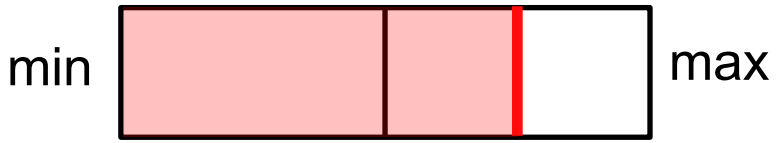
Performance Scores:

(% of max score)

Correct ID Rate: **88%**

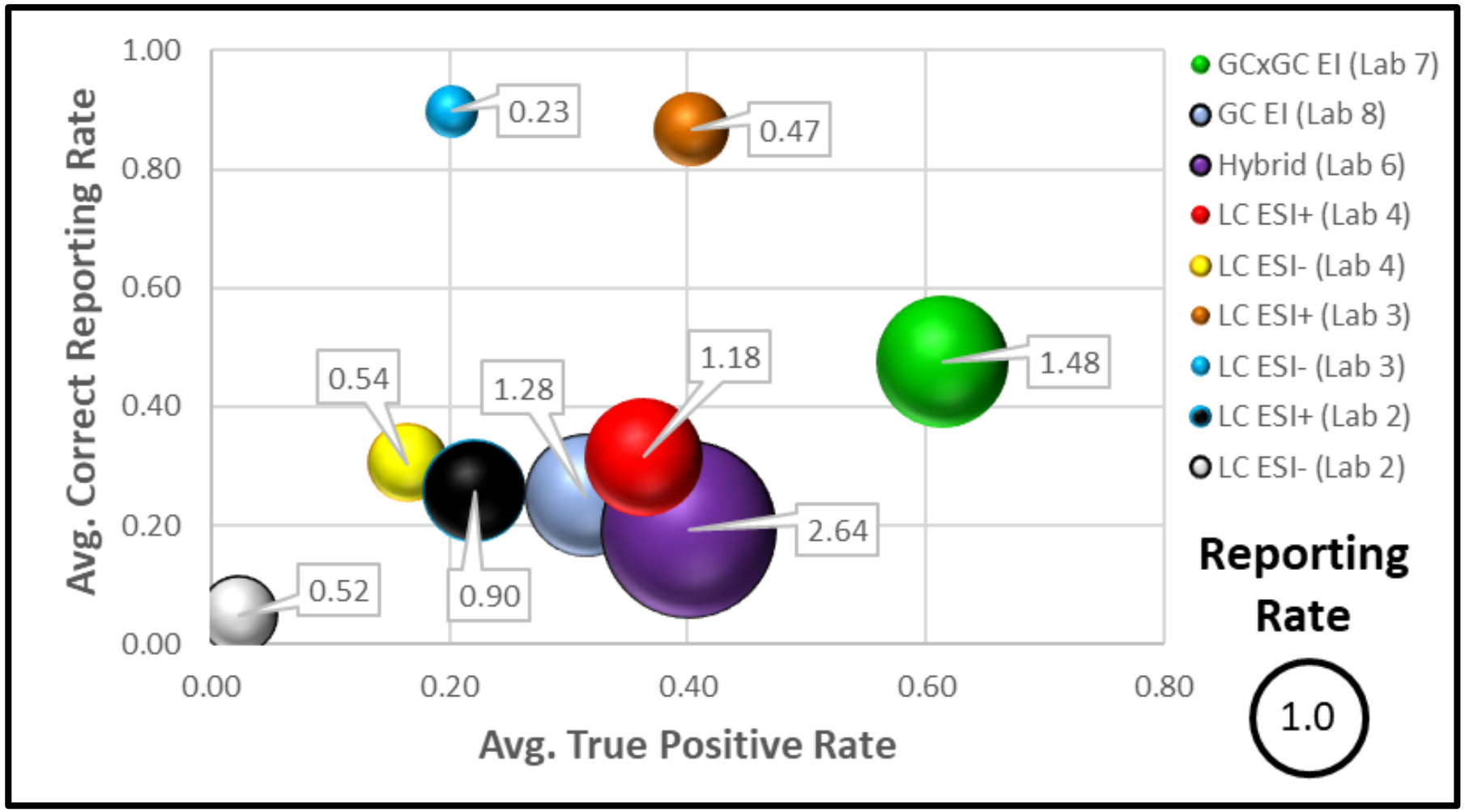


Repro. Rate: **78%**



Obs. Rate: **30%**

Method Comparison: TPR & Reverse Metrics



Metrics (all %):

Bubble Size → Amount reported

X-Axis → Correct IDs across all spikes

Y-Axis → Correct IDs across all reported

Additional Results for Collaborators

- Simple performance summary file (n=1 per method):
  - # and % correct identifications per sample
- Individual results files (n=10 per method):
  - Mass match (yes/no), formula match (yes/no), compound match (yes/no)
  - Highest confidence level (as reported or after consensus revision)
- Composite results file (n=1 per method):
  - For each spiked substance (n=1,269)
    - # of spikes (1-10), # of isomer spikes (1-5)
    - # mass hits, # formula hits, # compound hits
    - Observed (yes/no/undetermined), Correct ID (yes/no), Reproducible (yes/no)

EPA United States Environmental Protection Agency

Some Challenges (to date)

- Multiple chemical candidate submissions per feature
- Inconsistent & inaccurate use of scoring metrics
- Inconsistent & inaccurate chemical reporting procedures
- Inconsistent and unclear feature filtering protocols
- Limited engagement regarding collaborator follow-up
- Determining FPs vs. uTPs
- Determining TNs and dependent metrics
- Slow evaluation process vs. rapid method development processes

Summary of ENTACT Findings

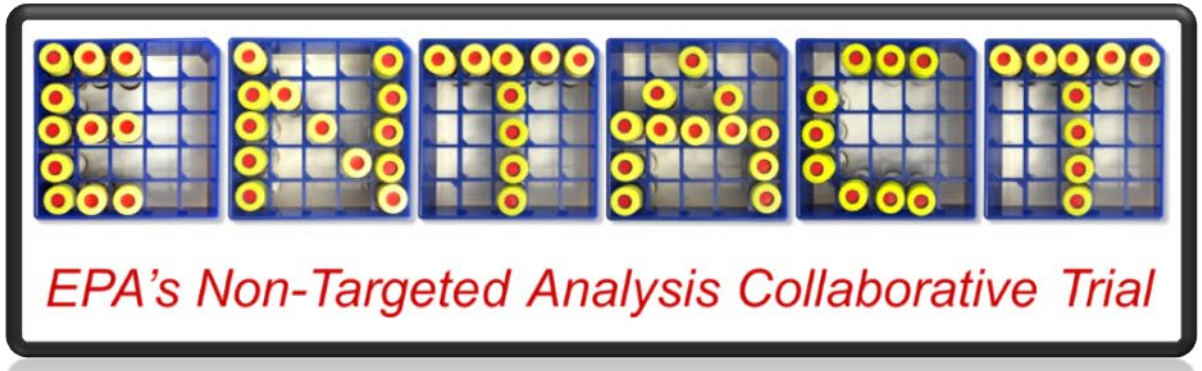
- **NTA methods are suitable for many ToxCast chemicals**
  - ~5% of ENTACT compounds not observed by any method
- **Multiple methods required for broad characterization**
  - No “one size fits all” method
  - <1% of ENTACT compounds observed using all methods
- **Performance determined across multiple metrics:**
  - **Observability Rate** = Ability to observe those spiked → (22% to 69%)
  - **True Positive Rate** = Ability to identify those spiked → (2% to 61%)
  - **Correct ID Rate** = Ability to identify those observed → (7% to 99%)
  - **Reproducibility Rate** = Ability to consistently identify → (7% to 97%)
  - **Reporting Rate** = Amount reported vs. spiked → (23% to 264%)
  - **Correct Reporting Rate** = Amount correctly ID’d vs. reported → (5% to 90%)

Take-Away Messages from ENTACT (to date…)

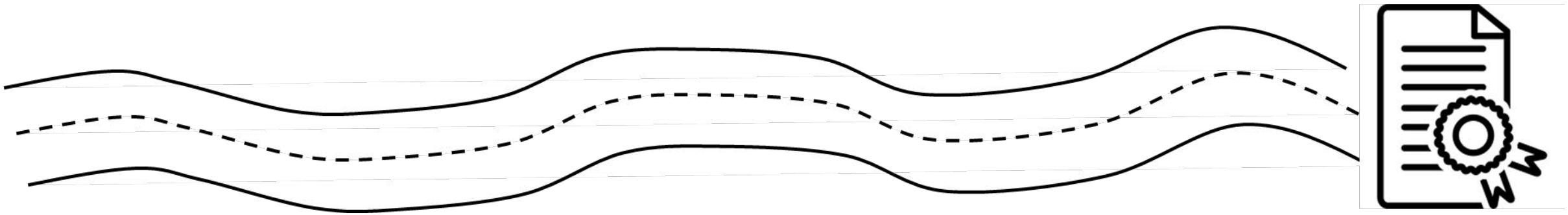
- Lack of transparency in methods/results reporting
- Method procedures change over short time increments
- Biased self-reporting → highlight strengths, mask weaknesses
- Blinded ToxCast mixtures allow for NTA performance assessment
- Performance measures highly variable across labs/methods
- Standard performance assessment methods/benchmarks should be adopted
- Benchmarks require input/consensus from NTA community
- Community focus should be on QA/QC

The Path to NTA Lab Credentialing

BP4NTA
BENCHMARKING AND PUBLICATIONS FOR NON-TARGETED ANALYSIS



Definitions & Guidance + **Performance Metrics** + **Reporting Criteria** + **Reference Materials** + **Evaluation Protocols** + **Performance Benchmarks**



Questions?

sobus.jon@epa.gov

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