https://orcid.org/0000-0002-6357-0508

United States Environmental Protection Agency

Estimating Uncertainty of Predicted Chemical Concentrations Via Quantitative Non-Targeted Analysis

Louis Groff, Jarod Grossman, Anneli Kruve, Jeffrey Minucci, Charles Lowe, James McCord, Dustin Kapraun, Katherine Phillips, S. Thomas Purucker, Alex Chao, Caroline Ring, Antony Williams, Jon Sobus

Office of Research and Development
Center for Computational Toxicology and Exposure

May 23, 2022

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

Toxic Substances Control Act (TSCA) Compliance Monitoring

To protect
federal, sta
with statut
import), pr
chemical su
substances

Safe Drinking Water Act (SDWA) Compliance Monitoring

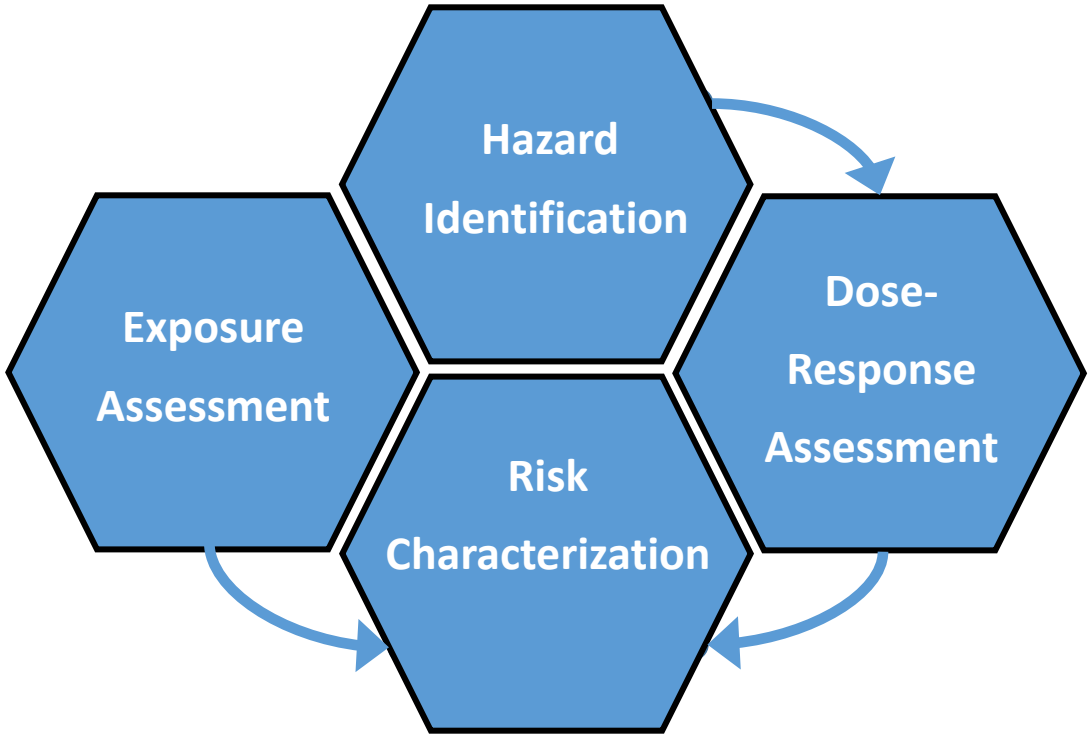
Providing safe drin
states, tribes, publi
certified laboratori
water samples coll
the tribes monitor
Water Act regulato

Federal Insecticide, Fungicide and Rodenticide Act Compliance Monitoring

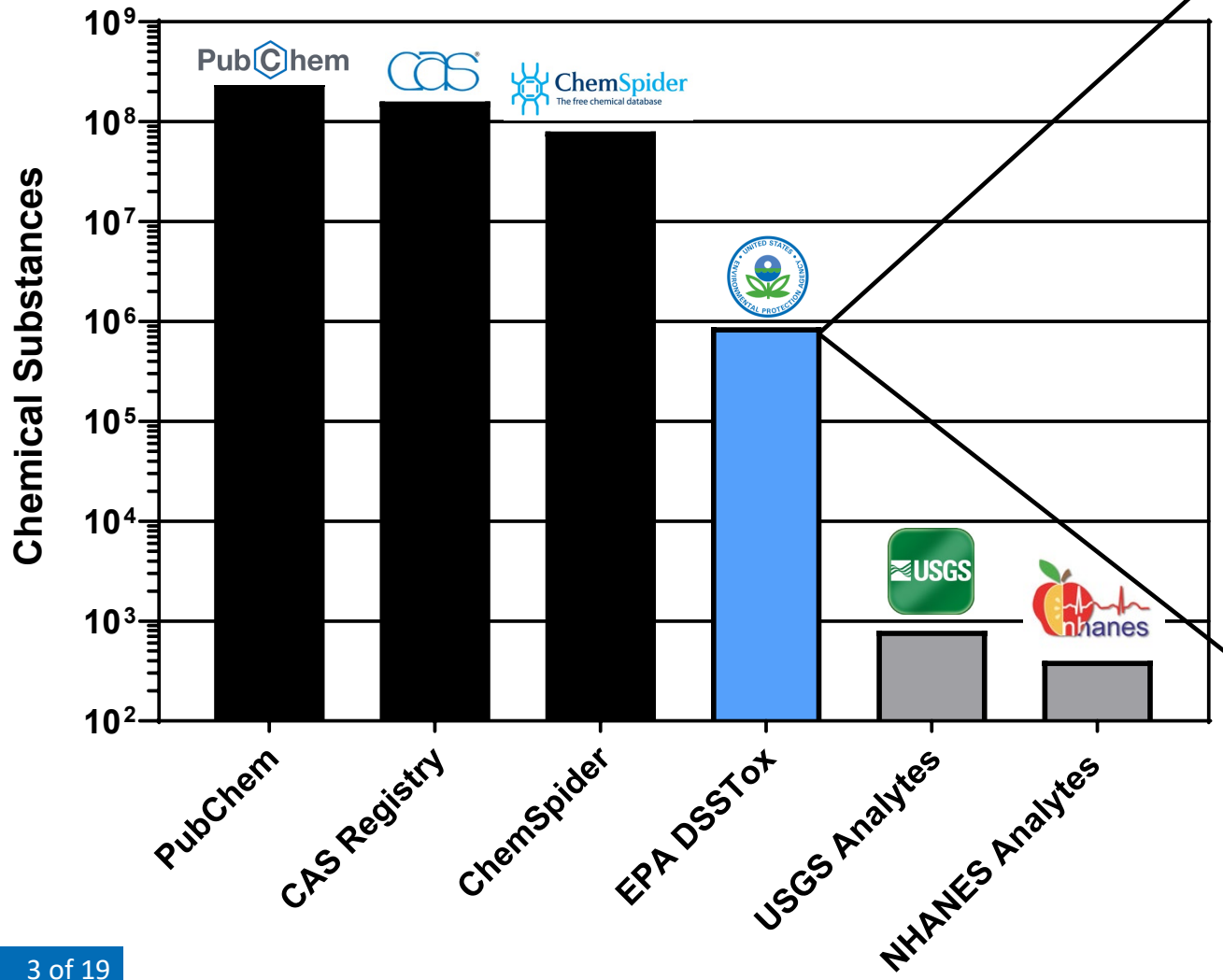
The Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) gives EPA the authority to regulate the registration, distribution, sale and use of pesticides. FIFRA applies to all types of pesticides, including:

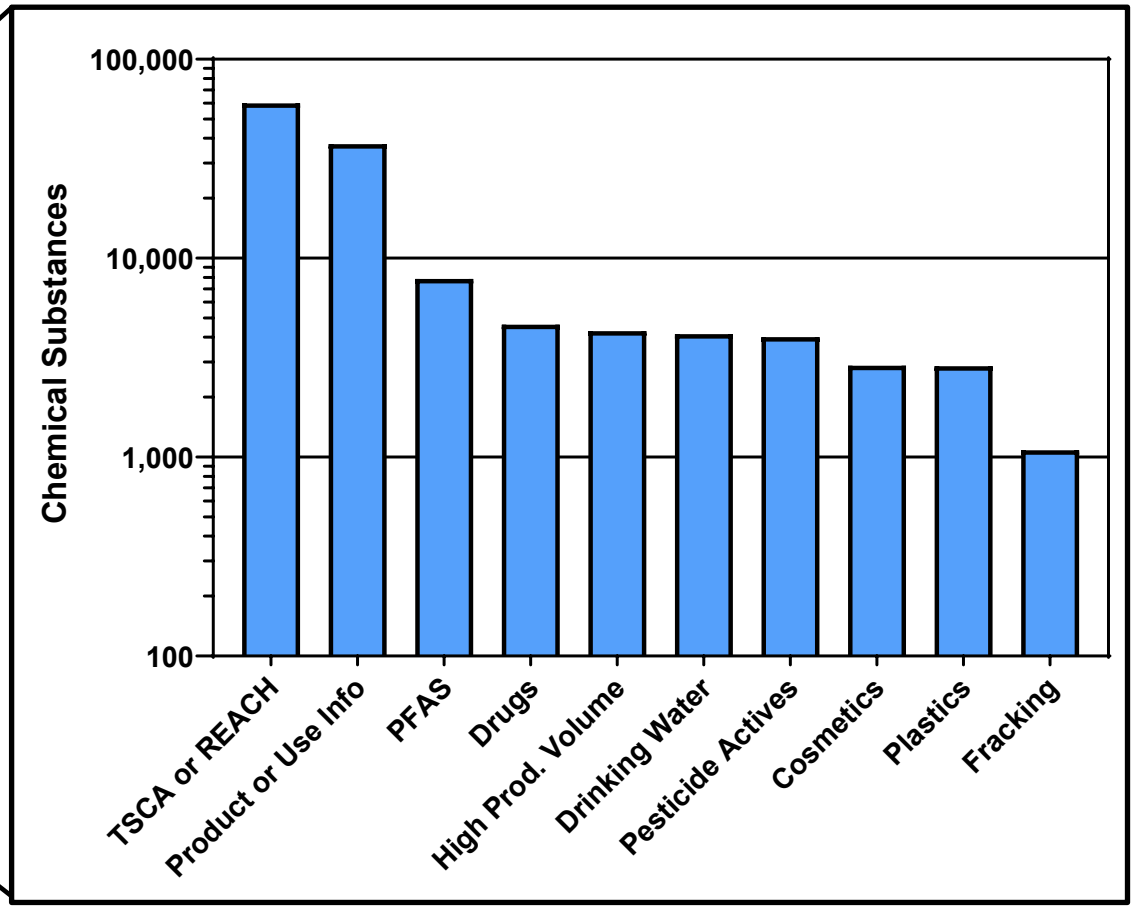
Resources and Guidance Documents

Chemical Monitoring Needs



Data Disparity: Have vs. Need





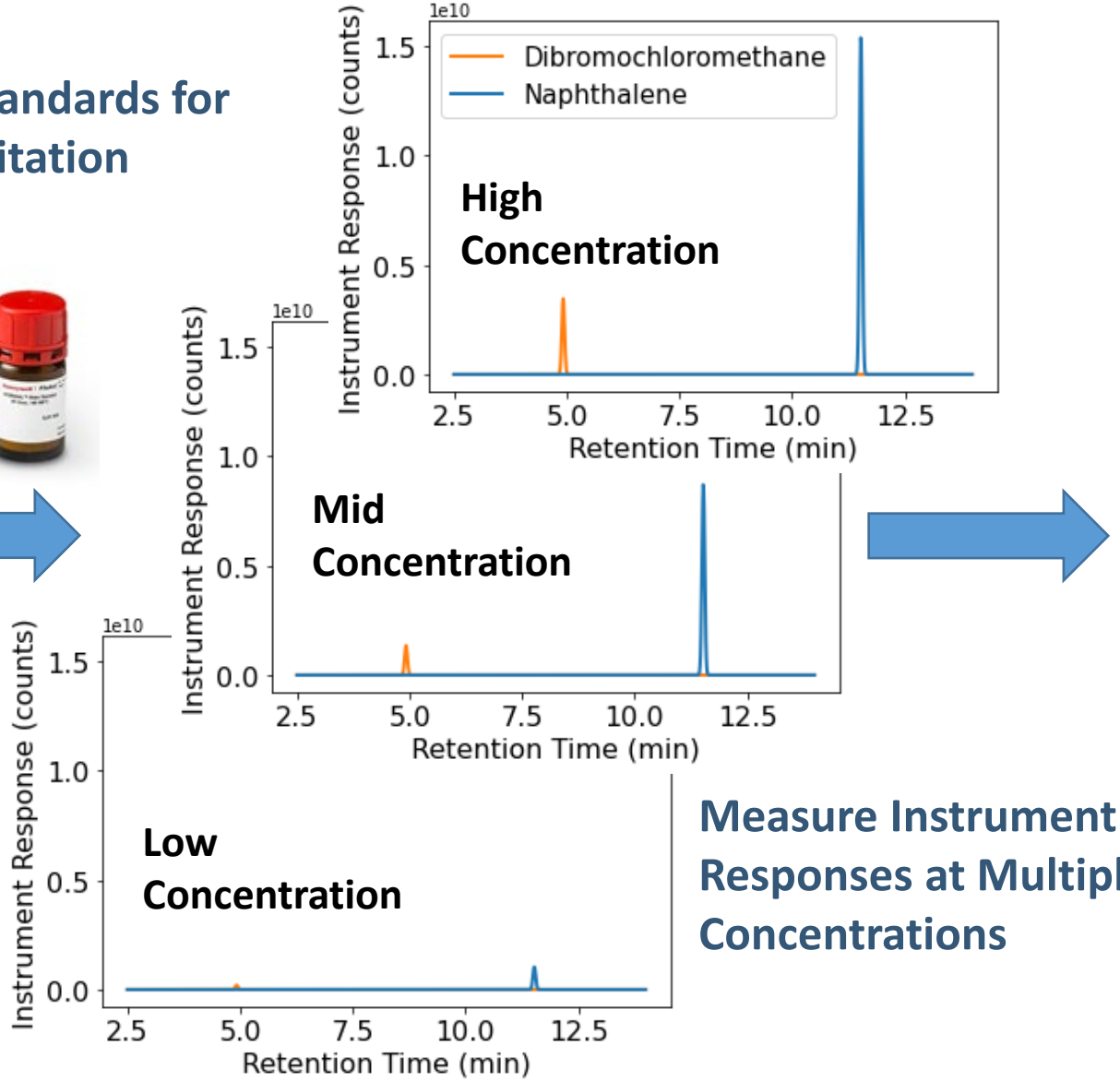
Challenges

- 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

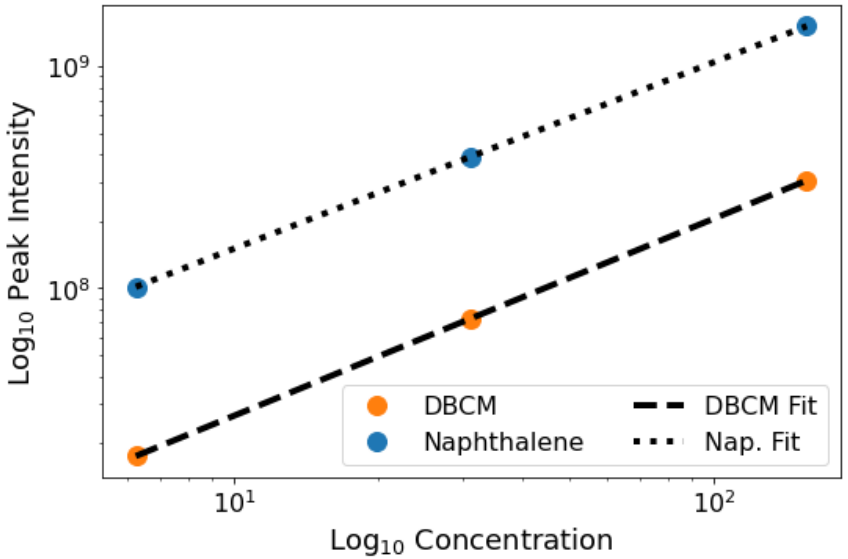
Traditional Targeted Analysis

Purchase standards for quantitation

Collect Sample of Interest



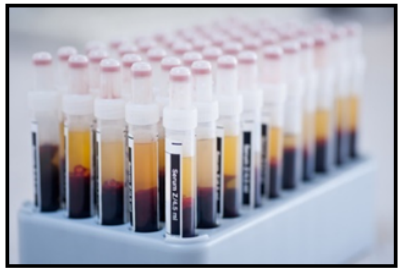
Measure Instrument Responses at Multiple Concentrations



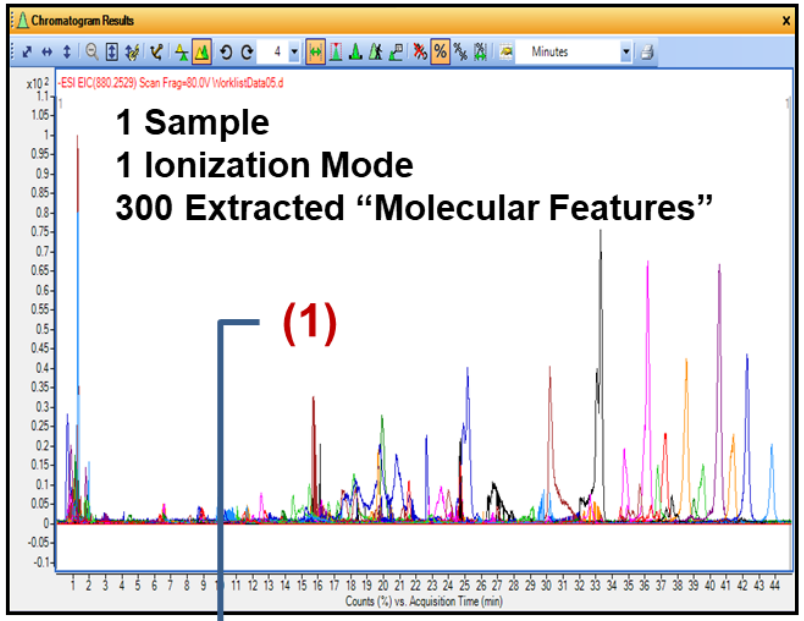
Generate Calibration Curves for Accurate Quantification of Individual Analytes

General NTA Workflow

Samples



High-Resolution MS

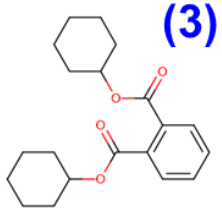


1) Prioritize "molecular features"
2) Correctly assign formulas
3) Correctly assign structures
4) Predict chemical concentrations
5) Determine chemical sources

(2)

$C_{20}H_{26}O_4$

(3)

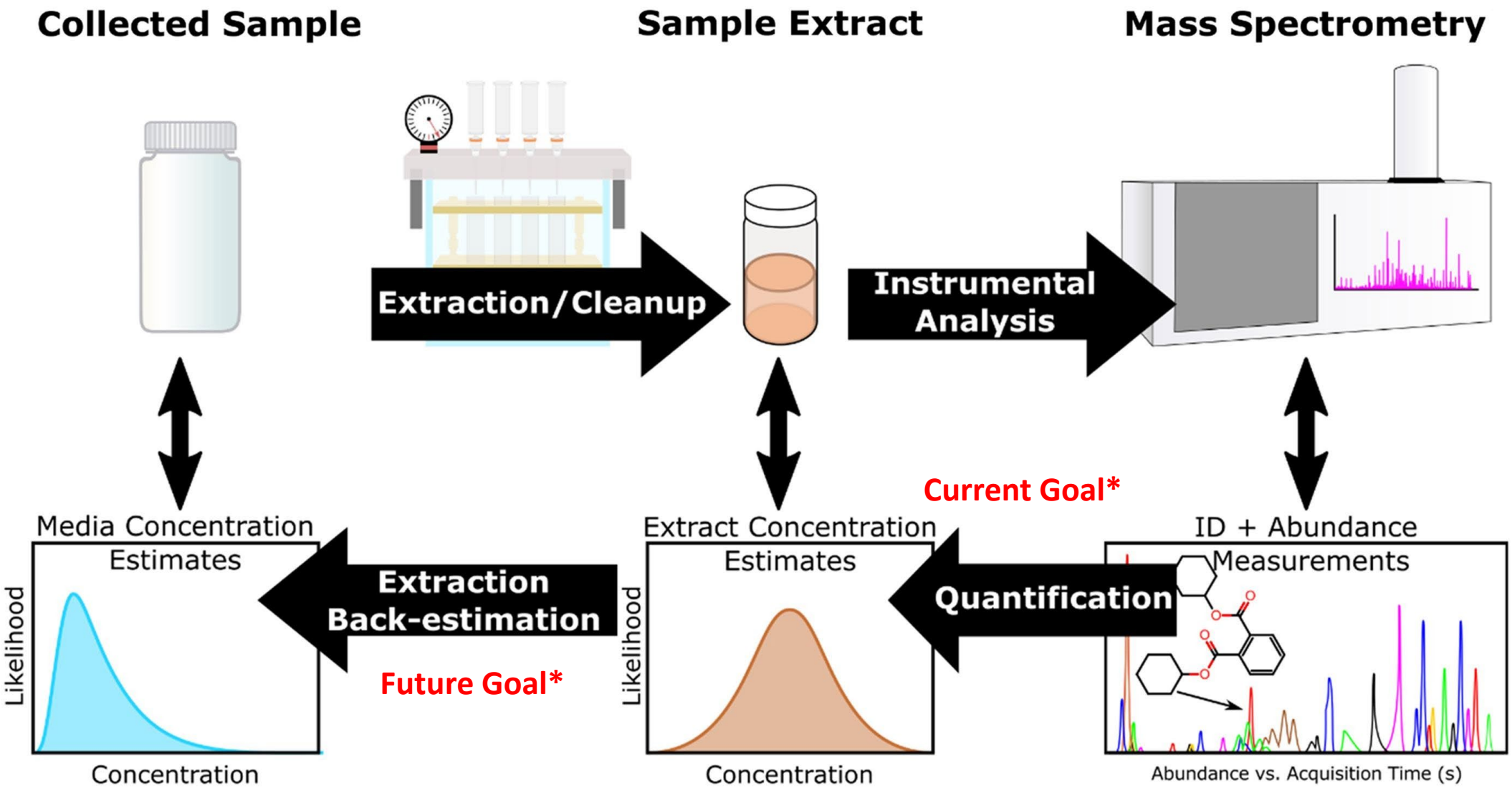


(4)

12 µg/mL

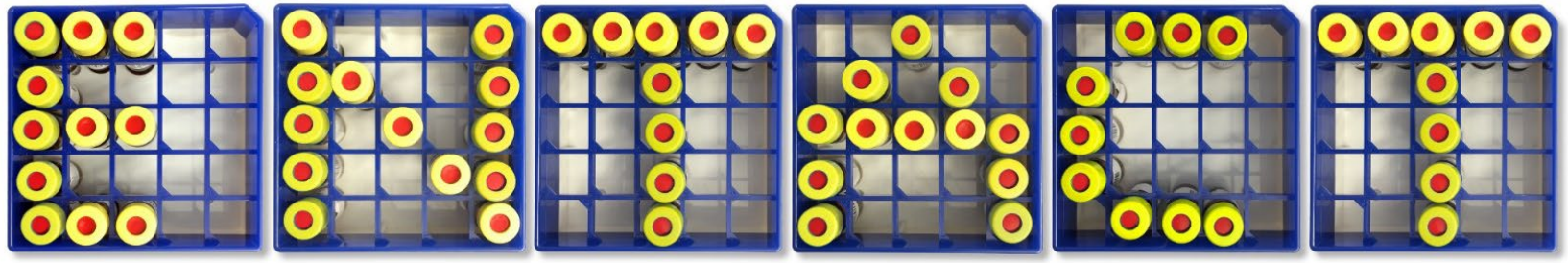
(5)

Quantitative NTA (qNTA) is a Multi-Step Process



McCord, J. P., Groff, L. C., and Sobus, J. R. *Environ. Int.* (2022) 158:107011.

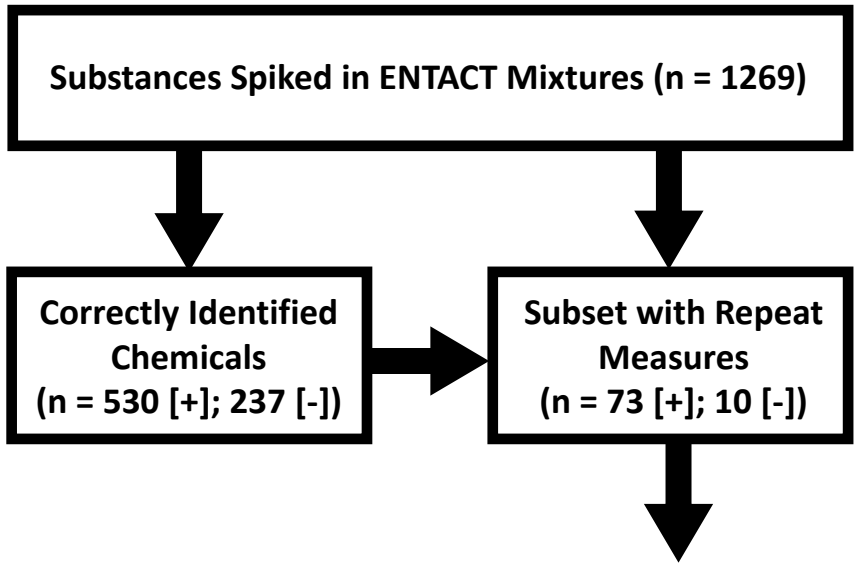
EPA's Non-Targeted Analysis Collaborative Trial as an NTA Dataset



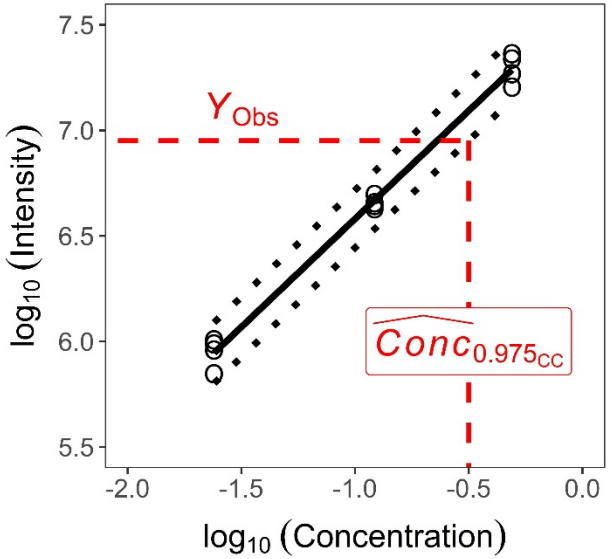
- Ten synthetic mixtures with 1269 chemical substances
- Each contains between 95 and 365 unique substances in DMSO
- Analyzed with LC-QToF high-resolution mass spectrometry (HRMS)
- 3 dilutions per mixture; chemical subset with replicate measures
- 530 compounds identified in ESI+; 237 in ESI-
- **Aim: develop and evaluate qNTA methods using ENTACT NTA data**

Sobus, J. R., et Al. *Anal. Bioanal. Chem.* (2019) 411:835–851.
Ulrich, E. M., et Al. *Anal. Bioanal. Chem.* (2019) 411:853–866.

Benchmark Method: Inverse Prediction Using Targeted Calibration Curves



Example Chemical



Prediction Error for Automated Analysis = ???

- Transform intensity & conc. data into log-log space
- Generate calibration curves for each chemical
- Fit → targeted (true) concentration
- 95% Prediction Interval → prediction error bound via inverse prediction
- **Use to compare to qNTA estimated concentrations**

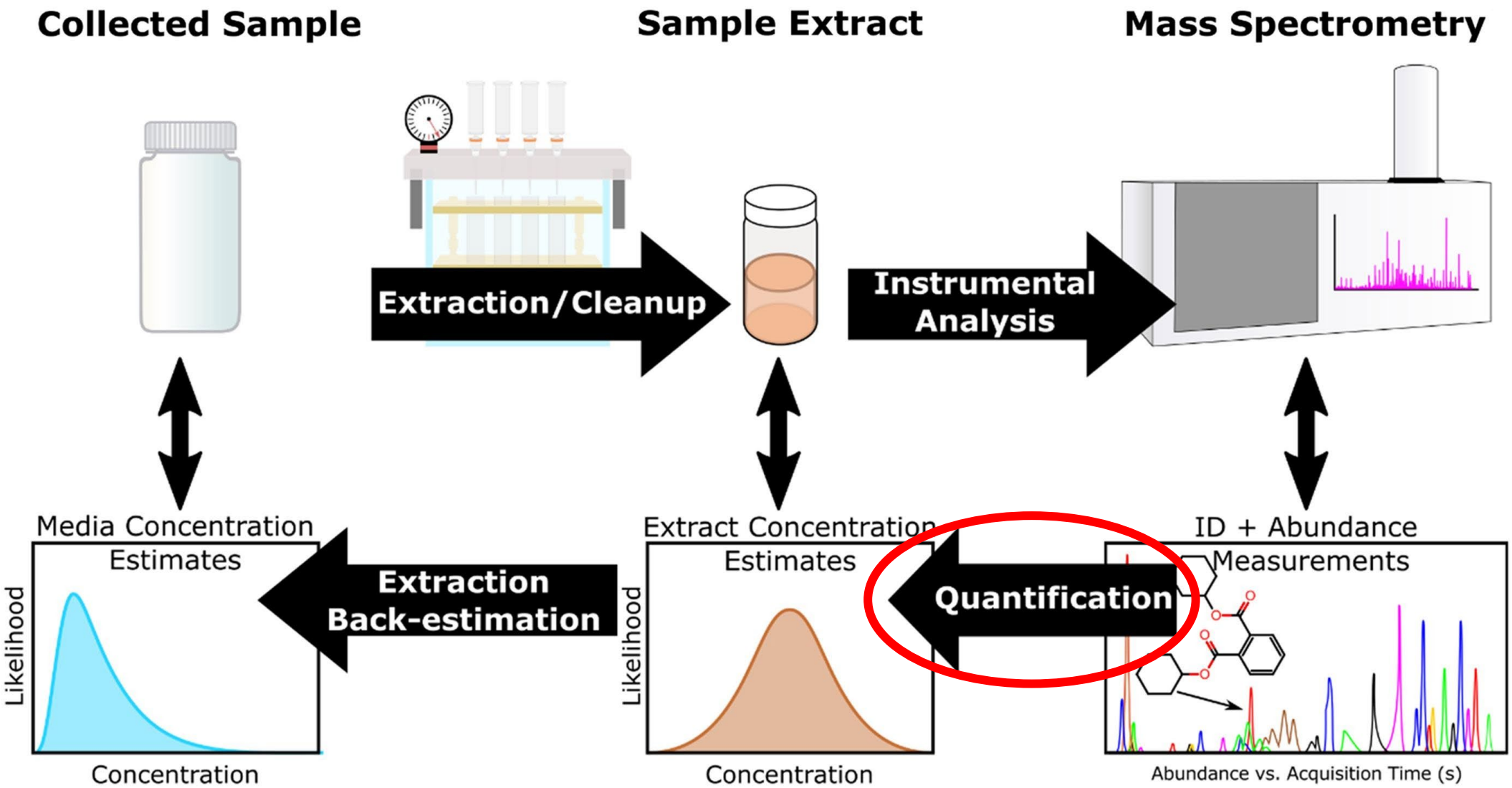
Given Y_{obs} →

$\widehat{Conc}_{CC}$
From Linear Fit

Given 95% Pred. Interval →

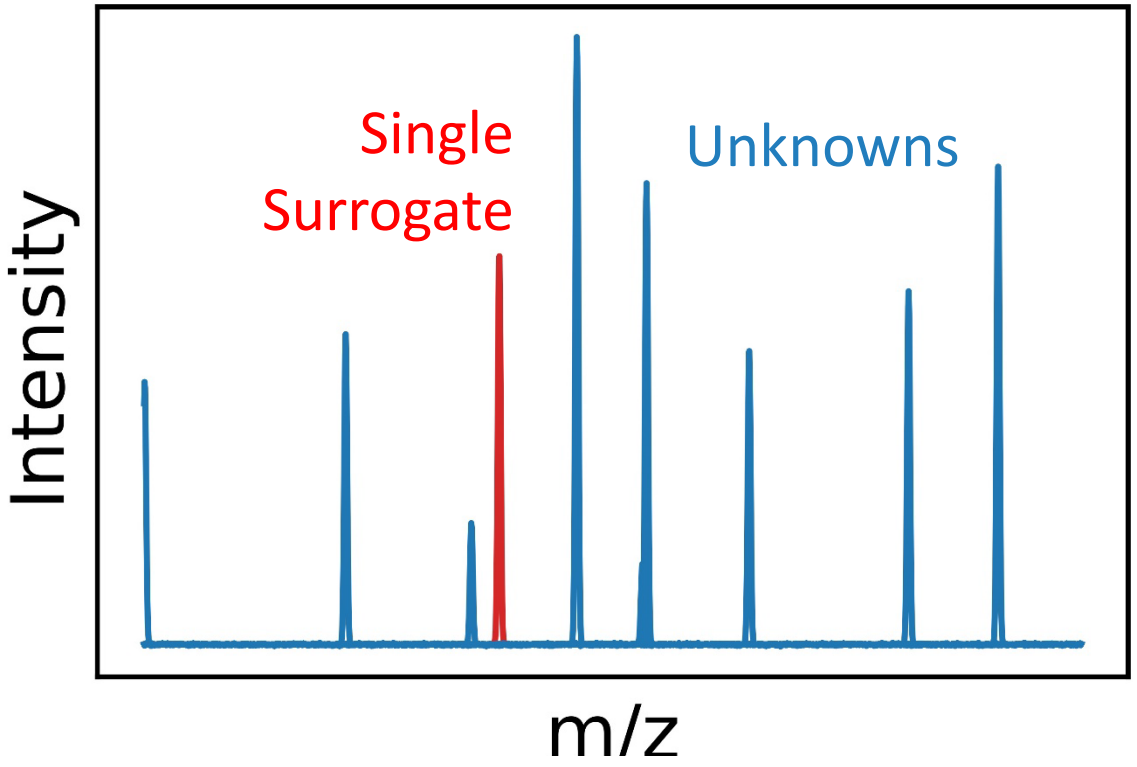
$\widehat{Conc}_{0.975_{CC}}$
From Lower PI Bound

Quantitative NTA (qNTA) is a Multi-Step Process



McCord, J. P., Groff, L. C., and Sobus, J. R. *Environ. Int.* (2022) 158:107011.

Simplest qNTA Model Uses Surrogate Response Factors



"Single Surrogate" → known chemical spiked at known conc. with observed intensity

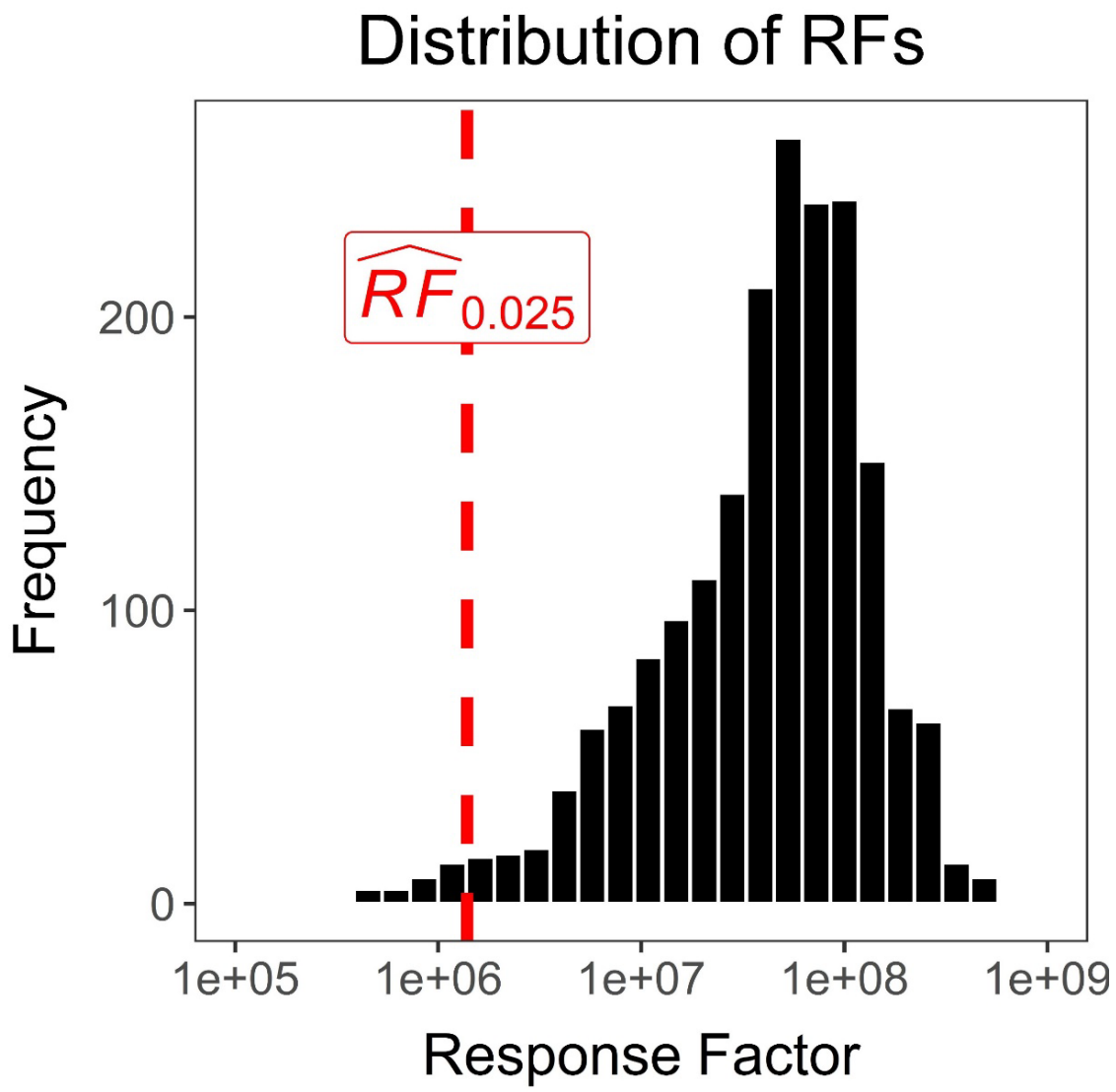
"Unknowns" → tentatively identified chemicals with unknown conc. and observed intensities

$$Response\ Factor\ (RF) = \frac{Obs.Intensity_{\ Surrogate}}{Known\ Conc._{\ Surrogate}}$$

$$Predicted\ Conc._{\ Unknown} = \frac{Obs.Intensity_{Unknown}}{RF}$$

Confidence Limit Strategy for qNTA Predictions Using Bootstrapped RF Distributions

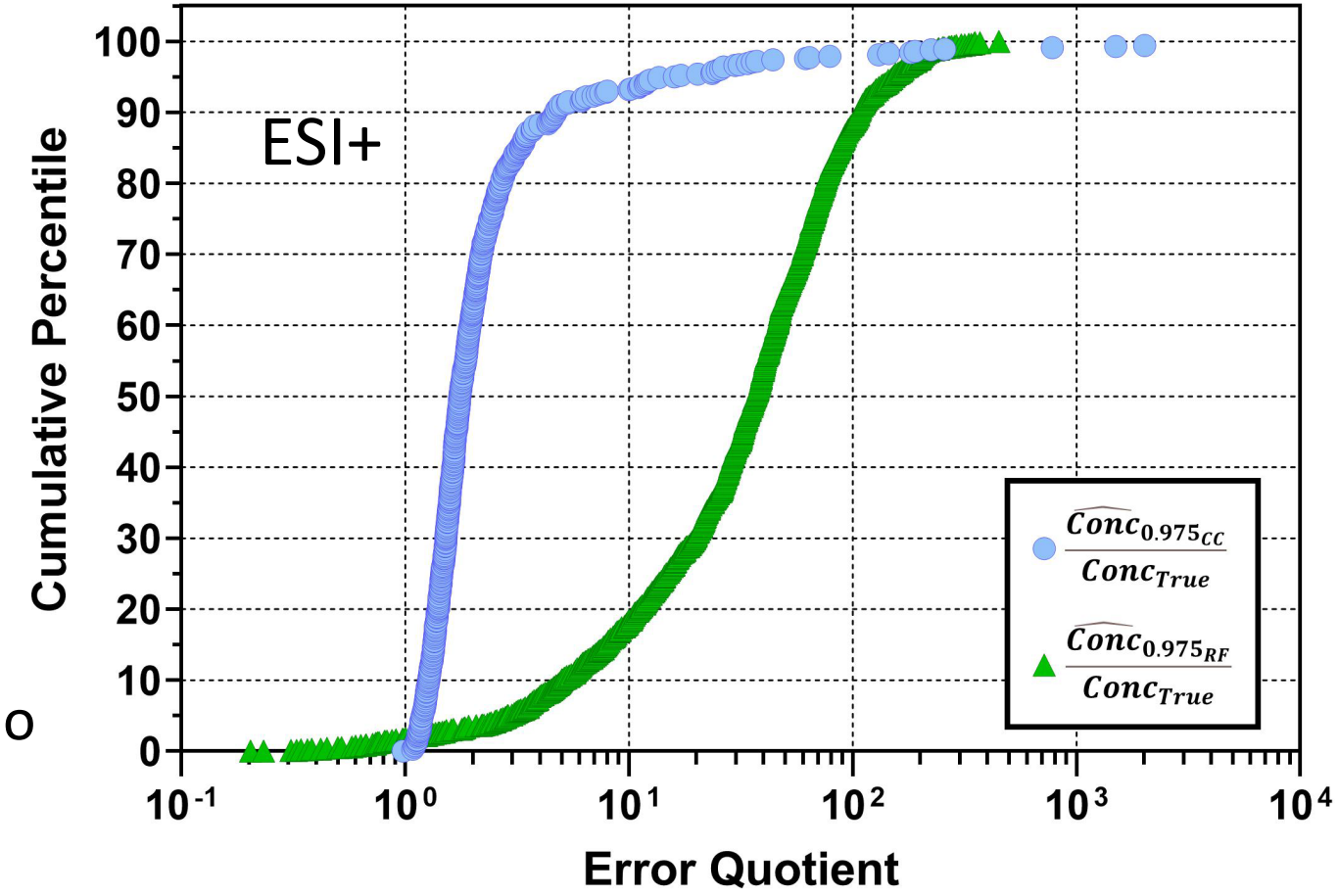
- Perform five-fold cross-validation to split ENTACT chemicals into training/test sets
- Bootstrap resample training set RF distribution many times (10k)
- Calculate 2.5th percentile RF for each resampled distribution
- Take average over 10k resamples and five CV folds to get $\widehat{RF}_{0.025}$
- Given $\widehat{Conc}_{RF} = Obs.\ Intensity/RF$
- Using $RF = \widehat{RF}_{0.025} \rightarrow \widehat{Conc}_{0.975_{RF}}$



Prediction Error for RF-Estimated Concentrations vs. Calibration Curve Estimates

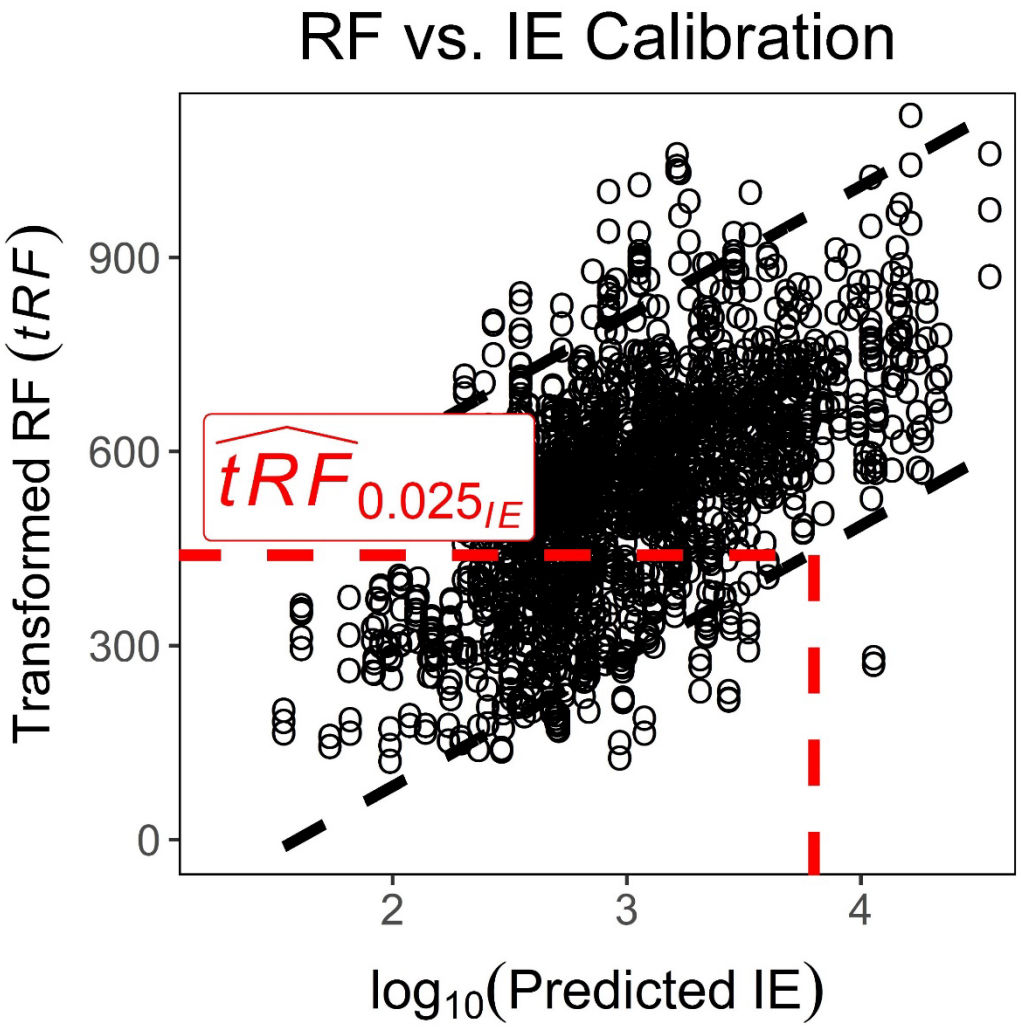
$$Error\ Quotient = \frac{\widehat{Conc}_{0.975}}{\widehat{Conc}_{True}}$$

- **Use cal. curve error quotient as benchmark:**
  - 50th percentile: 1.7× over-est.
  - 95th percentile: 16× over-est.
- **EQ $\widehat{Conc}_{0.975_{RF}}$ percentiles:**
  - 50th percentile: 37× over-est.
  - 95th percentile: 152× over-est.
  - ≤ 1.7th percentile: <5× under-est.
- RF method is naïve qNTA strategy, given no need for structural information



Groff, L. C. et al. *Anal. Bioanal. Chem.* (2022) Accepted.

Improving Concentration Estimates Using Ionization Efficiency Model Predictions



- Use physicochemical descriptors to predict ionization efficiency (IE) for each ENTACT chemical
- Beneficial statistical relationship between RF and predicted IE
- Predicted IE and RF were transformed to meet the assumptions of linear regression

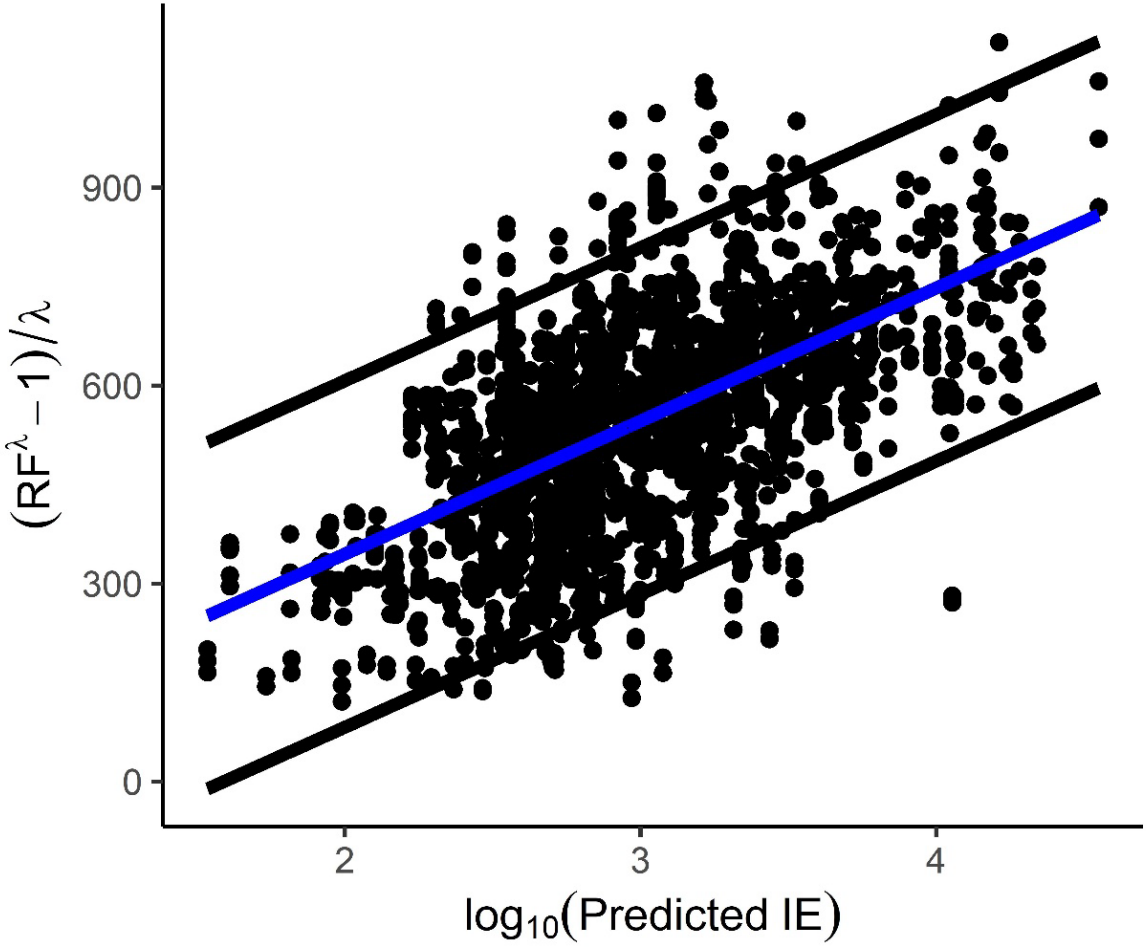
$$tRF = (RF^{\lambda} - 1)/\lambda$$

Box-Cox Transform Equation

$\lambda_{ESI+} = 0.285,\ \lambda_{ESI-} = -0.106$

Liigand, J., Wang, T., Kellogg, J., Smedsgaard, J. Cech, N., and Kruve, A. Sci Rep 10, 5808 (2020).

IE-Predicted Response Factors Using Linear Mixed-Effects Modeling

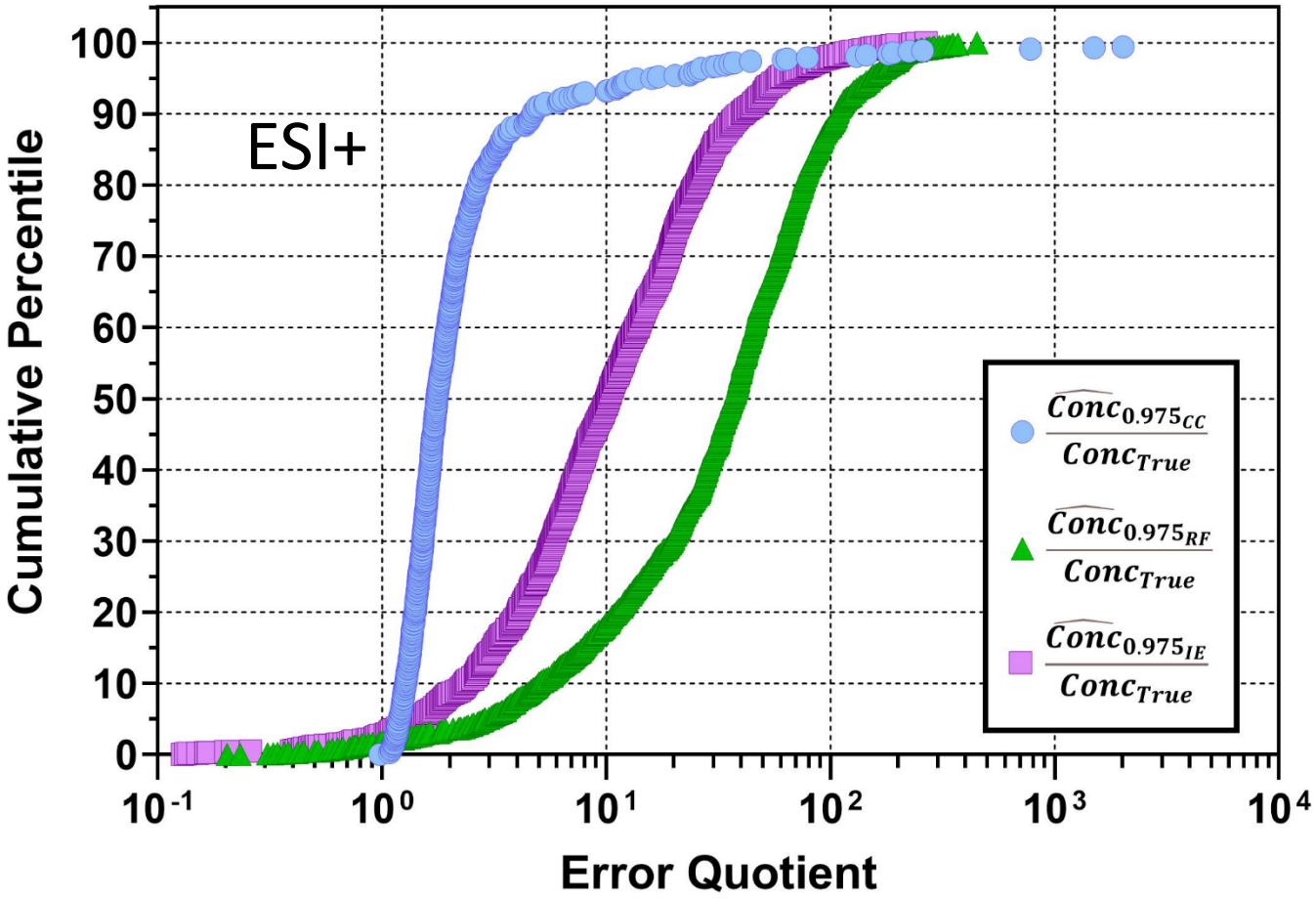


- Repeat five-fold cross-validation procedures
- Bootstrap resample training set tRF vs. log(IE) distribution many times (10k)
- Calculate linear mixed model regression coefficients on the resampled distributions
- Determine prediction interval for each CV fold
- Given predicted log(IE), we can calculate $\widehat{tRF}_{0.025_{IE}}$ and back-transform to $\widehat{RF}_{0.025_{IE}}$
- $\widehat{Conc}_{0.975_{IE}} = Obs.Intensity/\widehat{RF}_{0.025_{IE}}$

Prediction Error Across qNTA Methods

$$Error\ Quotient = \frac{\widehat{Conc}_{0.975}}{\widehat{Conc}_{True}}$$

- Use cal. curve error quotient as benchmark:
  - 50th percentile: 1.7× over-est.
  - 95th percentile: 16× over-est.
- EQ $\widehat{Conc}_{0.975_{RF}}$ percentiles:
  - 50th percentile: 37× over-est.
  - 95th percentile: 152× over-est.
- **EQ $\widehat{Conc}_{0.975_{IE}}$ percentiles:**
  - **50th percentile: 10× over-est.**
  - **95th percentile: 59× over-est.**



Groff, L. C. et al. *Anal. Bioanal. Chem.* (2022) Accepted.

Conclusions

- NTA is an integral tool for keeping pace with the discovery of chemicals of emerging concern
- qNTA provides a means to estimate confidence limits about concentration estimates, with high statistical confidence, for chemicals lacking authentic standards
- Interpretation: ***"There is a 95% probability that the true concentration lies between X_1 lower bound and X_2 upper bound."***
- **Upper-bound** concentration estimates can be used for provisional chemical safety screenings
- Using chemical specific calibration curves with automated NTA data processing, upper-bound concentration estimates are within ~15× of the true concentration (ESI+)
- Using a naïve response factor estimation method, upper-bound concentration estimates are within ~150× of the true concentration (ESI+)
- Using mixed model regressions of response factor vs. predicted ionization efficiency, upper-bound concentration estimates are within ~60× of the true concentration (ESI+)
- Using any of these methods, the upper bound concentration estimate will be LOWER than the true value ~2.5% of the time –inherent to the chosen 95% confidence level, but within ~5x of the true concentration (ESI+)

Future Activities

- Apply qNTA models to existing NTA sample datasets generated via GC & LC platforms (consumer products, environmental media, biological samples)
- Apply sample extraction data to extend bounded concentrations in prepared solution upward toward media concentrations
- Develop risk-prioritization strategies that combine qNTA media predictions with estimated thresholds of human and ecological toxicity
- Examine platform transferability for qNTA models
- Incorporate into EPA NTA Informatics Toolkit

Contributing Researchers

EPA ORD

Jon Sobus
Hussein Al-Ghoul*
Alex Chao
Jarod Grossman*
Kristin Isaacs
Dustin Kapraun
Charles Lowe
James McCord
Jeff Minucci
Katherine Phillips
Tom Purucker
Caroline Ring
Randolph Singh*
Elin Ulrich
Dimitri Panagopoulos Abrahamsson*

EPA ORD (cont.)

Chris Grulke
Kamel Mansouri*
Andrew McEachran*
Ann Richard
Antony Williams

* = ORISE/ORAU

Stockholm University

Anneli Kruve

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

Questions?

Groff.Louis@epa.gov
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.