United States Environmental Protection Agency

Predicting Exposure Pathways with Machine Learning

John F. Wambaugh[1], Caroline L. Ring[1,3,], Kristin K. Isaacs[2], Katherine A. Phillips[2], Peter P. Egeghy[2], R. Woodrow Setzer[1]*

1. National Center for Computational Toxicology, Office of Research and Development
2. National Exposure Research Laboratory, Office of Research and Development
3. Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee 3783

*Currently ToxStrategies, Austin, TX

International Society of Exposure Science
"Chemical Prioritization via Computational Exposure and Hazard Screening"
Research Triangle Park, NC
October 18, 2017

orcid.org/0000-0002-4024-534X

EPA Office of Research and Development

- The Office of Research and Development (ORD) is the scientific research arm of EPA
- Research is conducted by ORD's three national laboratories, four national centers, and two offices
  - Includes **National Center for Computational Toxicology** and **National Exposure Research Laboratory**
- 14 facilities across the country and in Washington, D.C.
- Six research programs
  - Includes **Chemical Safety for Sustainability**
- Research conducted by a combination of Federal scientists; contract researchers; and postdoctoral, graduate student, and post-baccalaureate trainees

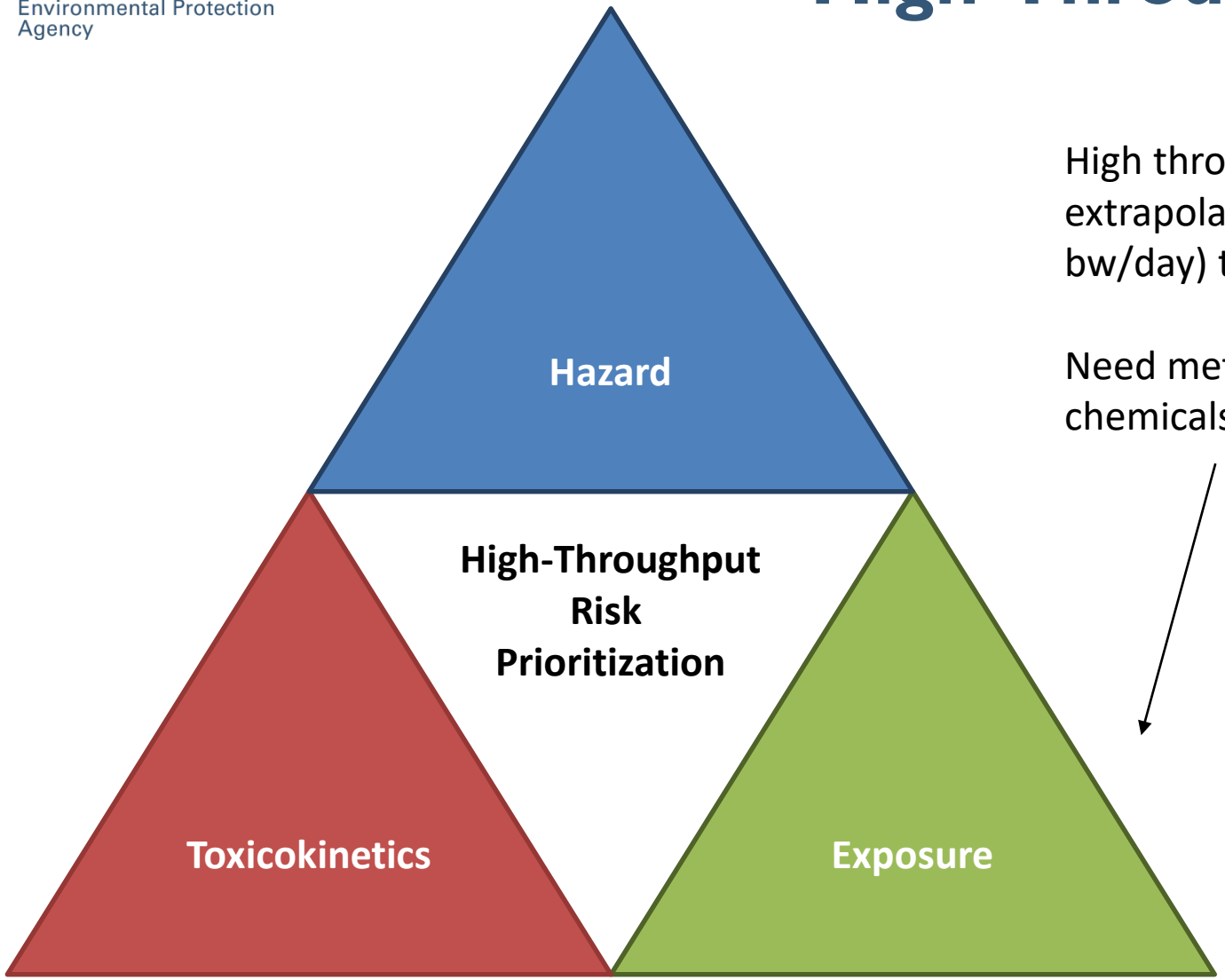
ORD Facility in Research Triangle Park, NC

Introduction

- Park *et al*. (2012): At least 3221 chemicals in pooled human blood samples, many appear to be exogenous
- Prioritizing the risk posed to human health from the thousands of chemicals in the environment requires tools that can estimate exposure rates from limited information
- High throughput models exist to make predictions of exposure via specific, important pathways such as residential product use, diet, and environmental fate and transport (Arnot et al., 2006, Rosenbaum et al., 2008, Wambaugh et al., 2014, Isaacs et al., 2014)
- These models can be parameterized in terms of physico-chemical properties that can be predicted with reasonable accuracy from chemical structure

November 29, 2014

High-Throughput Risk Prioritization



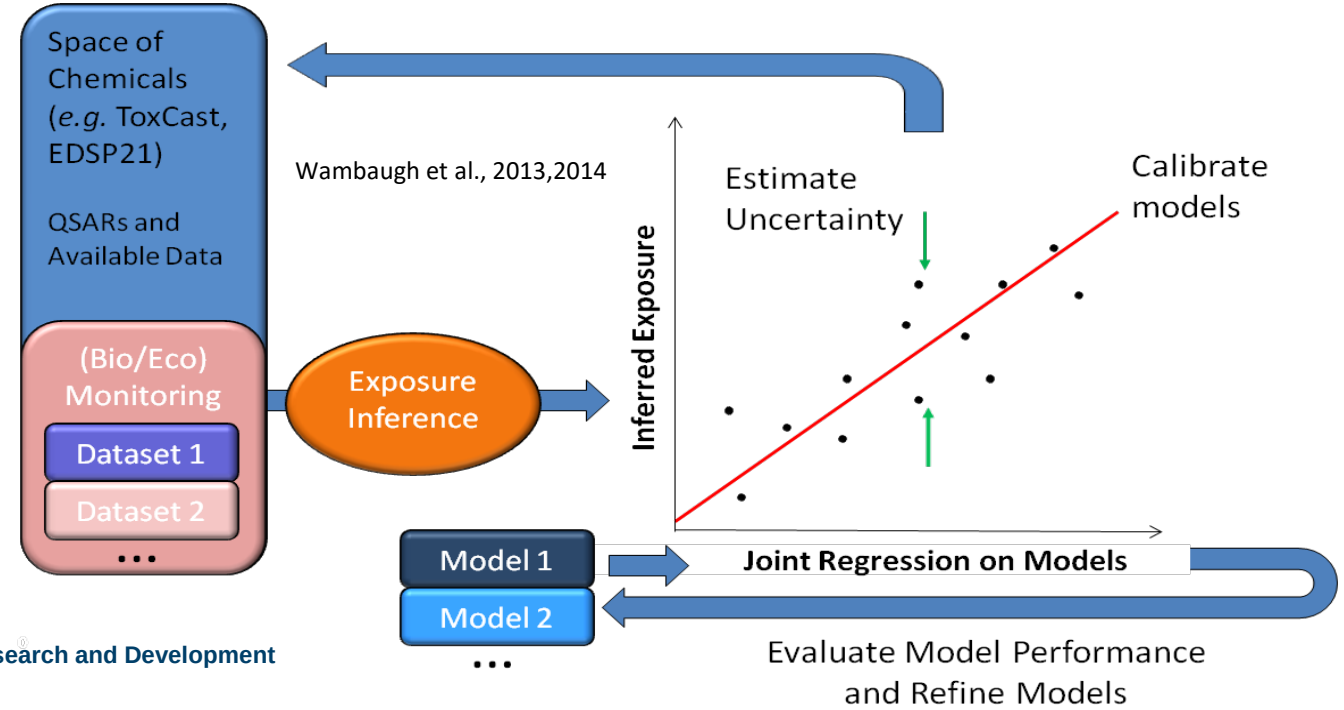
High throughput screening + *in vitro-in vivo* extrapolation (IVIVE can predict a dose (mg/kg bw/day) that might be adverse

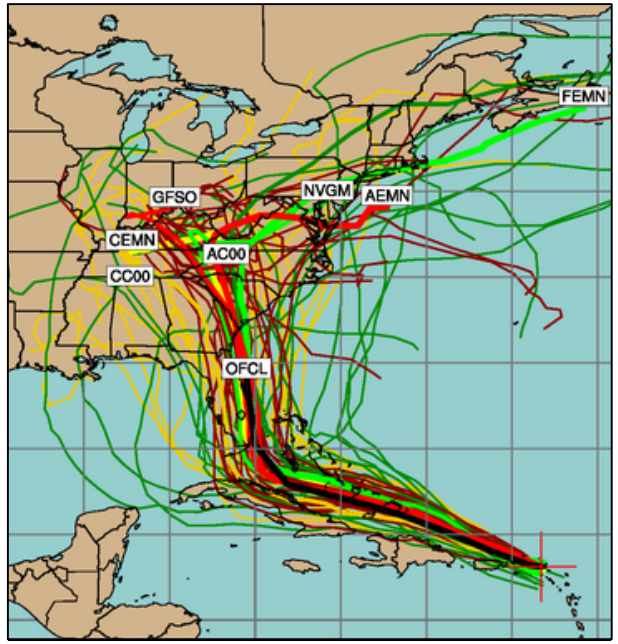
Need methods to forecast exposure for thousands of chemicals (Wetmore *et al.*, 2015)

Egeghy *et al.* (2012) – Most chemicals lack exposure data

Consensus Exposure Predictions with the SEEM Framework

- We incorporate multiple models (including SHEDS-HT, ExpoDat) into consensus predictions for 1000s of chemicals within the **Systematic Empirical Evaluation of Models (SEEM) framework**
- We evaluate/calibrate predictions with available monitoring data
- This provides information similar to a sensitivity analysis: What models are working? What data are most needed? This is an iterative process





Integrating Multiple Models

Exposures Inferred from NHANES

National Health and Nutrition Examination Survey

- Annual survey, data released on 2-year cycle
- Separate evaluations can be done for various demographics
- ~2000 individuals per chemical, with statistical weights allowing inference for larger U.S. populations
- To date, we have used this to draw inference about median exposure rates

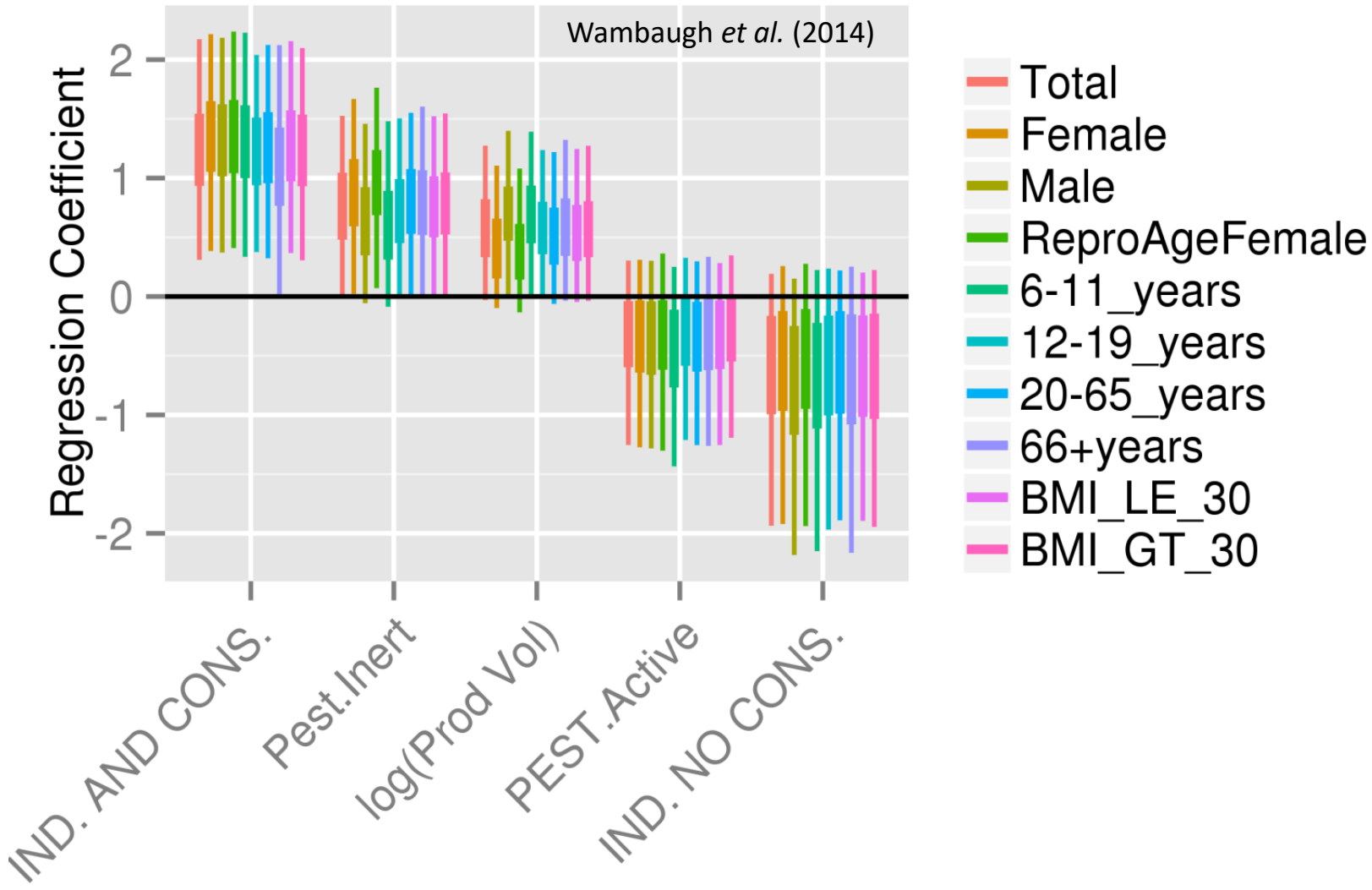
Urinary Bisphenol A (2,2-*bis*[4-Hydroxyphenyl] propane)

Geometric mean and selected percentiles of urine concentrations (in µg/L) for the U.S. populatio
and Nutrition Examination Survey.

	Survey years	Geometric mean (95% conf. interval)	Selected percentiles (95% confidence interval) 50th	75th	90th
Total	03-04	2.64 (2.38-2.94)	2.80 (2.50-3.10)	5.50 (5.00-6.20)	10.6 (9.40-
	05-06	1.90 (1.79-2.02)	2.00 (1.90-2.00)	3.70 (3.50-3.90)	7.00 (6.40-
	07-08	2.08 (1.92-2.26)	2.10 (1.90-2.30)	4.10 (3.60-4.60)	7.70 (6.80-
Age group					
6-11 years	03-04	3.55 (2.95-4.29)	3.80 (2.70-5.00)	6.90 (6.00-8.30)	12.6 (9.50-
	05-06	2.86 (2.52-3.24)	2.70 (2.30-2.90)	5.00 (4.40-5.80)	13.5 (9.30-
	07-08	2.46 (2.20-2.75)	2.40 (1.90-3.00)	4.50 (3.70-5.50)	7.00 (6.30-
12-19 years	03-04	3.74 (3.31-4.22)	4.30 (3.60-4.60)	7.80 (6.50-9.00)	13.5 (11.8-
	05-06	2.42 (2.18-2.68)	2.40 (2.10-2.70)	4.30 (3.90-5.20)	8.40 (6.50-
	07-08	2.44 (2.14-2.78)	2.30 (2.10-2.60)	4.40 (3.70-5.50)	9.70 (7.30-
20 years and older	03-04	2.41 (2.15-2.72)	2.60 (2.30-2.80)	5.10 (4.50-5.70)	9.50 (8.10-
	05-06	1.75 (1.62-1.89)	1.80 (1.70-2.00)	3.40 (3.10-3.70)	6.40 (5.80-
	07-08	1.99 (1.82-2.18)	2.00 (1.80-2.30)	3.90 (3.40-4.60)	7.40 (6.60-

CDC, Fourth National Exposure Report (2011)

Heuristics of Exposure



Five descriptors explain roughly 50% of the chemical to chemical variability in median NHANES exposure rates

Same five predictors work for all NHANES demographic groups analyzed – stratified by age, sex, and body-mass index:

- Industrial and Consumer use
- Pesticide Inert
- Pesticide Active
- Industrial but no Consumer use
- Production Volume

What we are really doing is identifying chemical exposure pathway

Chemical Use Identifies Relevant Pathways

- **Exposure event unobservable**
  - Can try to predict exposure by characterizing pathway
- Some pathways have much higher average exposures!
  - In home "Near field" sources significant (Wallace, *et al.,* 1987)
- Chemical-Product Database (https://actor.epa.gov/cpcat/) provides chemical use information (Dionisio et al., 2015)

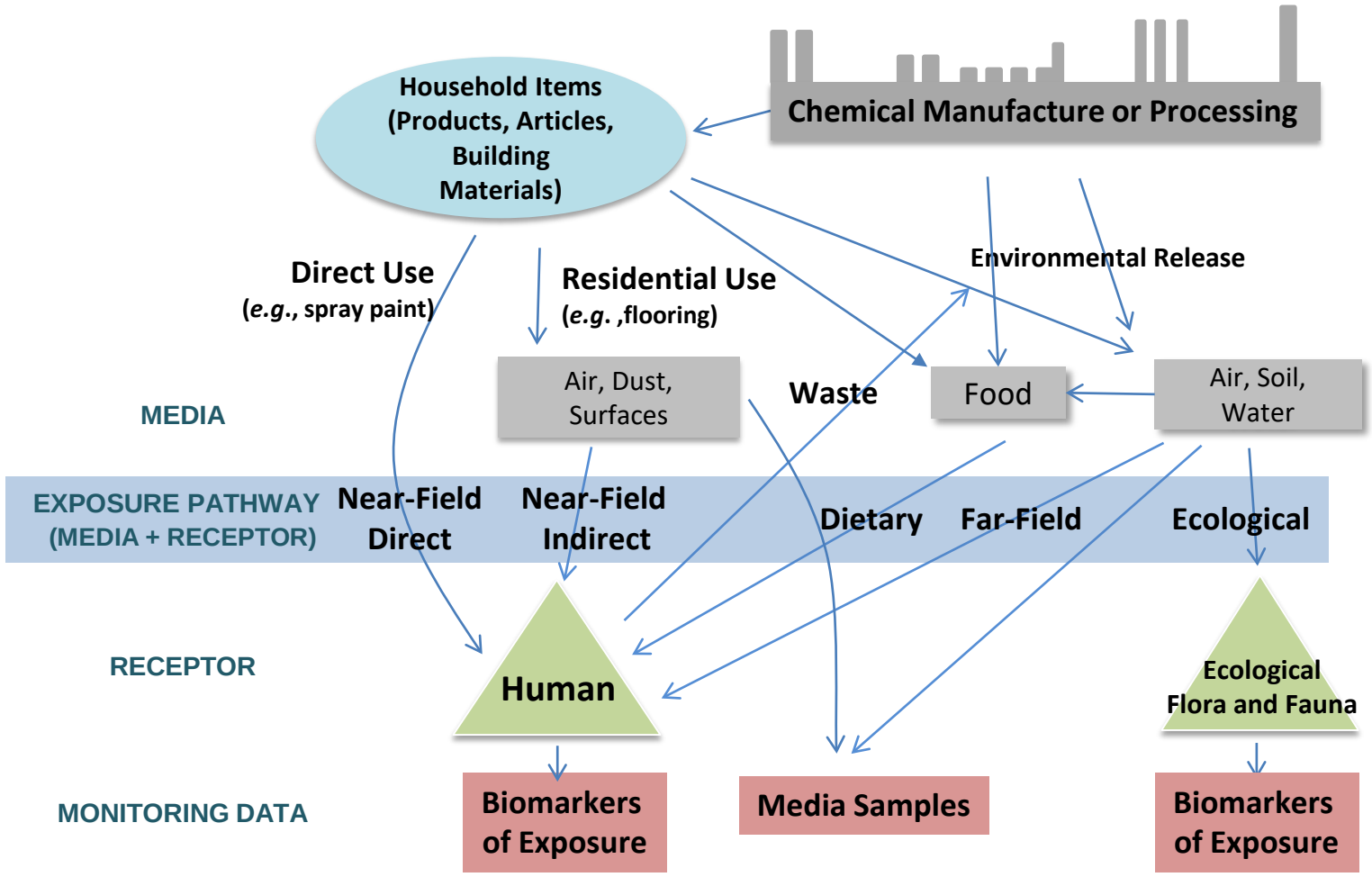


Figure from Kristin Isaacs

Knowledge of Exposure Pathways Limits High Throughput Exposure Models

"In particular, the assumption that 100% of [quantity emitted, applied, or ingested] is being applied to each individual use scenario is a very conservative assumption for many compound / use scenario pairs."

Environmental Science & Technology

Article

pubs.acs.org/est

Risk-Based High-Throughput Chemical Screening and Prioritization using Exposure Models and in Vitro Bioactivity Assays

Hyeong-Moo Shin,*[,†] Alexi Ernstoff,[‡,§] Jon A. Arnot,[∥,⊥,#] Barbara A. Wetmore,[∇] Susan A. Csiszar,[§] Peter Fantke,[‡] Xianming Zhang,[○] Thomas E. McKone,[◆,¶] Olivier Jolliet,[§] and Deborah H. Bennett[†]

[†]Department of Public Health Sciences, University of California, Davis, California 95616, United States

[‡]Quantitative Sustainability Assessment Division, Department of Management Engineering, Technical University of Denmark, Kgs. Lyngby 2800, Denmark

[§]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, Michigan 48109, United States

[∥]ARC Arnot Research and Consulting, Toronto, Ontario M4M 1W4 , Canada

[⊥]Department of Physical and Environmental Sciences, University of Toronto, Scarborough, Toronto, Ontario M1C 1A4, Canada

[#]Department of Pharmacology and Toxicology, University of Toronto, Toronto, Ontario M5S 1A8, Canada

[∇]The Hamner Institutes for Health Sciences, Research Triangle Park, North Carolina 27709, United States

[○]Harvard School of Public Health and School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, United States

[◆]Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720 , United States

[¶]School of Public Health, University of California, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: We present a risk-based high-throughput screening

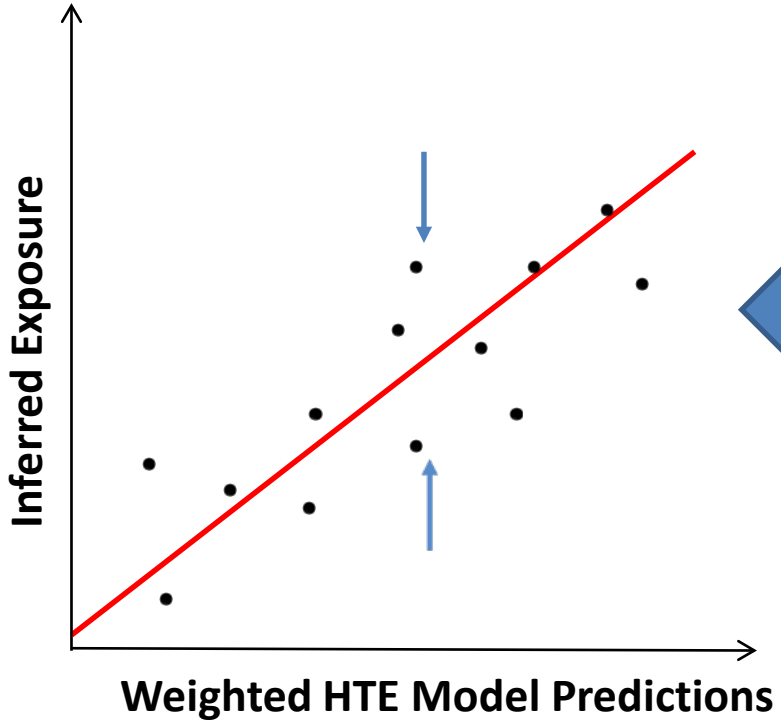
Potential exposure from exposure

Potential hazard from in vitro

SEEM is a Linear Regression

Multiple regression models:

Log(Parent Exposure) = $a + m * \log(\text{Model Prediction}) + b * \text{Near Field} + \varepsilon$

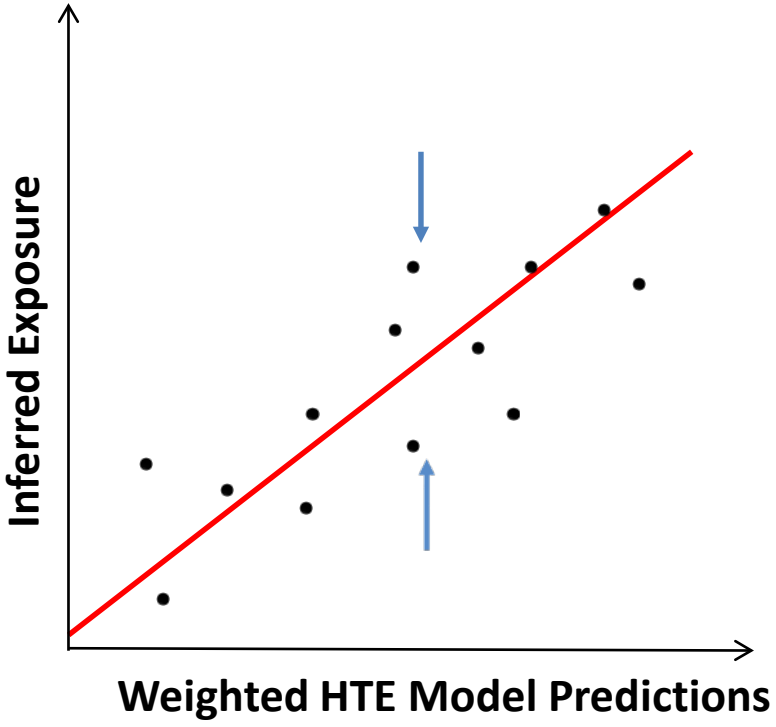


$\varepsilon \sim N(0, \sigma^2)$
Residual error, unexplained by the regression model

SEEM is a Linear Regression

Multiple regression models:

Log(Parent Exposure) = $a + m * \log(\text{Model Prediction}) + b * \text{Near Field} + \varepsilon$



Not all models have predictions for all chemicals

- We can run SHEDS-HT (Isaacs et al., 2014) for ~2500 chemicals

What do we do for the rest?

- Assign the average value?
- Zero?

Pathway Predictors:

Chemical Use Identifies Relevant Pathways

When averaging over many exposure models, the trick is to know which one to use…

Machine learning models were built for each four exposure pathways:

1. Far-field pesticide use
2. Non-pesticide dietary exposure
3. Far-field industrial exposure (e.g. drinking water)
4. Near-field exposure (e.g., consumer products).

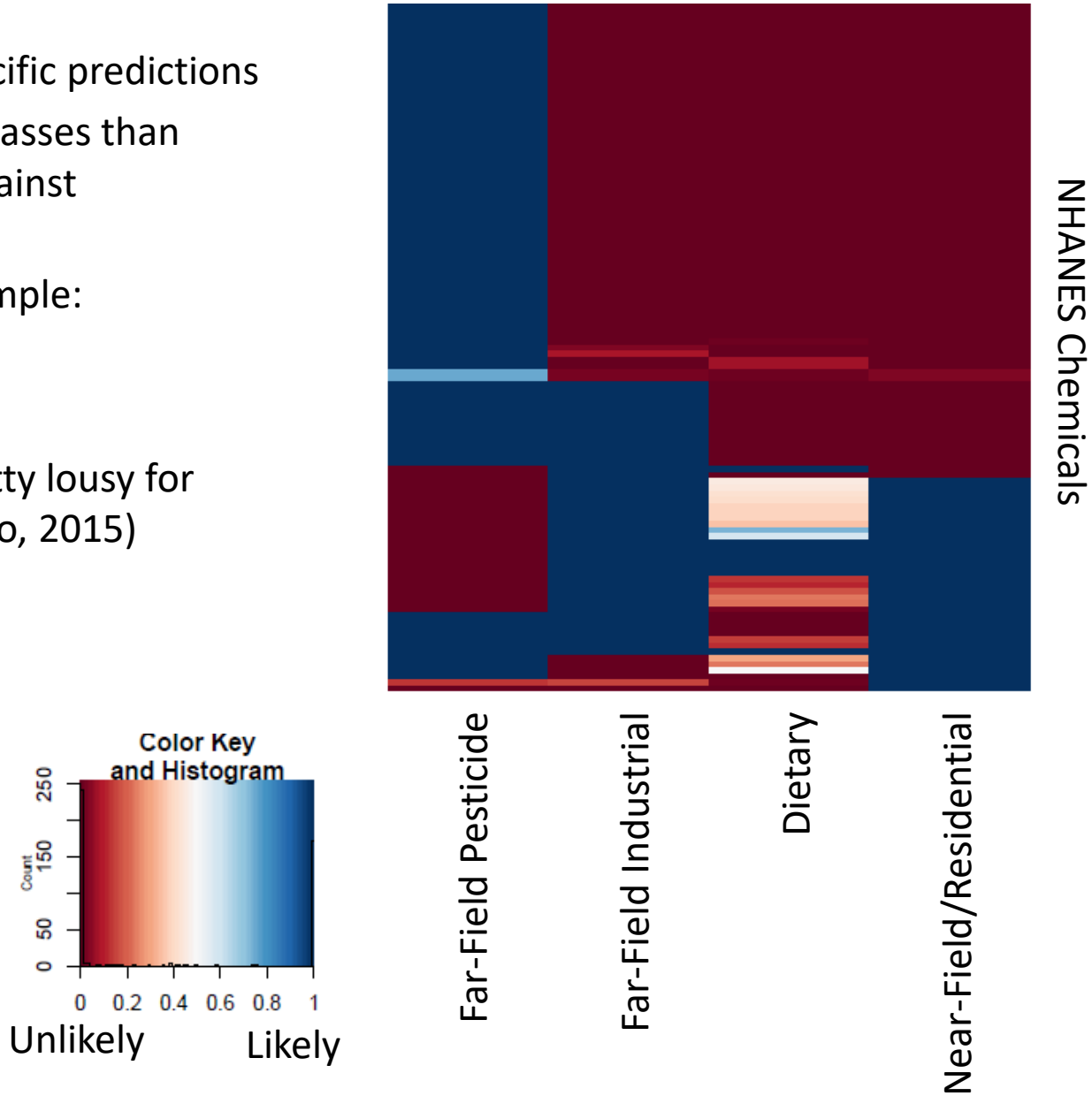
Pathway	Positives	Negatives	OOB Error Rate	Positives Error Rate	Balanced Accuracy	Sources of Positives	Sources of Negatives
Dietary	2429	13331	7.8	34	92	FDA CEDI, ACToR USEdb, NHANES Curation	ACToR USEdb, NHANES Curation
Near-Field	1382	3498	20	51	80	CPCPdb, Household Products Non-Targeted Analysis*, NHANES Curation	ACToR USEdb, NHANES Curation
Far-Field Pesticide	1726	9204	9.2	48	91	REDs, ACToR USEdb, NHANES Curation	NHANES curation, Diet Positives, ACToR USEdb, NHANES Curation
Far Field Industrial	3183	3792	18	21	82	USGS Water Occurence, ACToR USEdb, NHANES Curation	ACToR USEdb, Dietary and Pesticde Positives

Use Random Forest (Breiman, 2001) to predict based upon production volume, OPERA phys-chem (Mansouri et al., submitted), and ToxPrint structure descriptors (Yang, 2015)

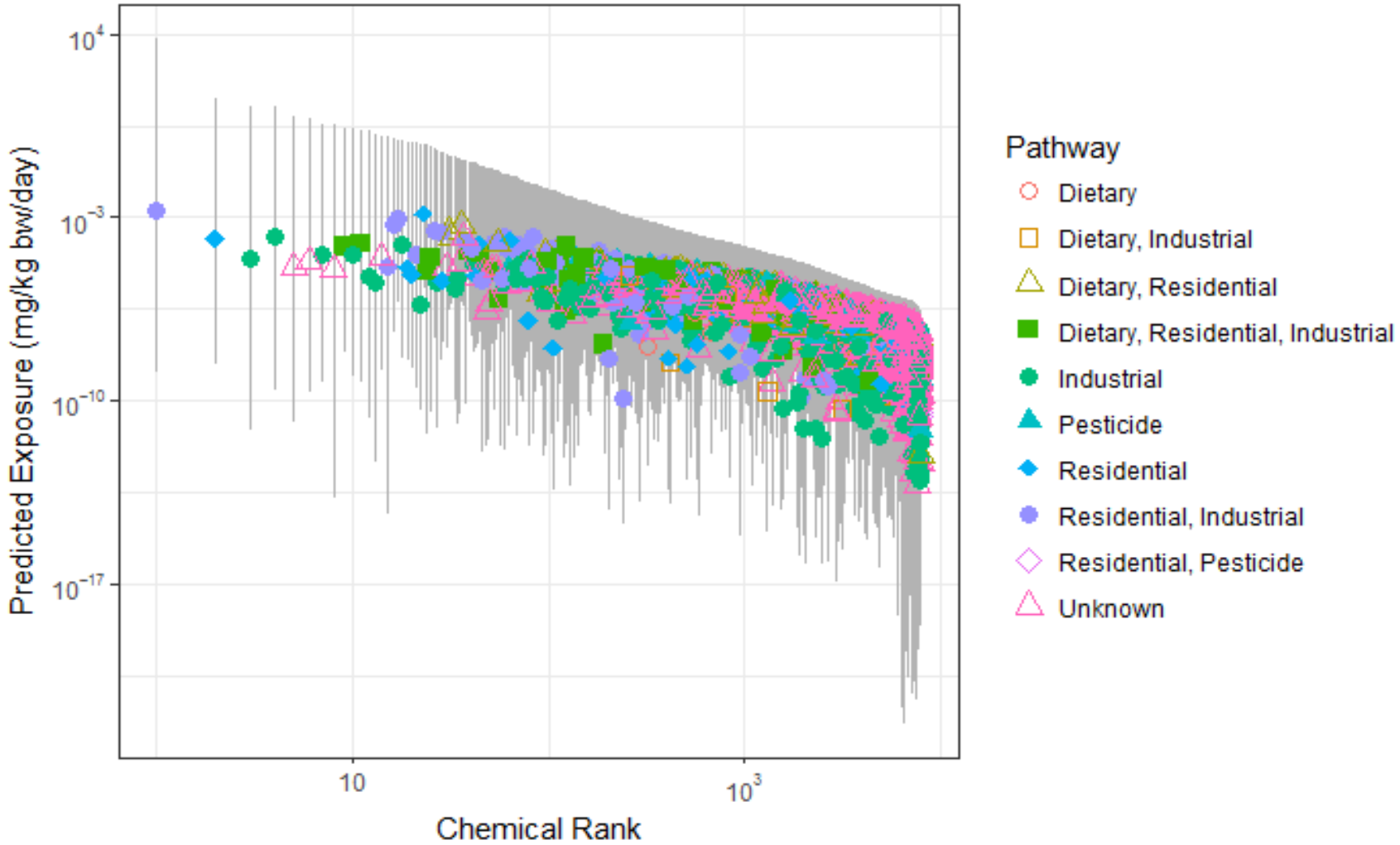
*Phillips et al., submitted

Pathway Probabilities

- Different predictive models provide different chemical-specific predictions
  - Some models may do a better job for some chemical classes than others overall, so we want to evaluate performance against monitoring data
- Hard to identify positives and especially negatives. For example:
  - What is a non-industrial chemical?
  - How do I know something isn't in consumer products?
- Manual inspection determined that tools we had were pretty lousy for NHANES, so did a manual curation guided by CPcat (Dionisio, 2015)



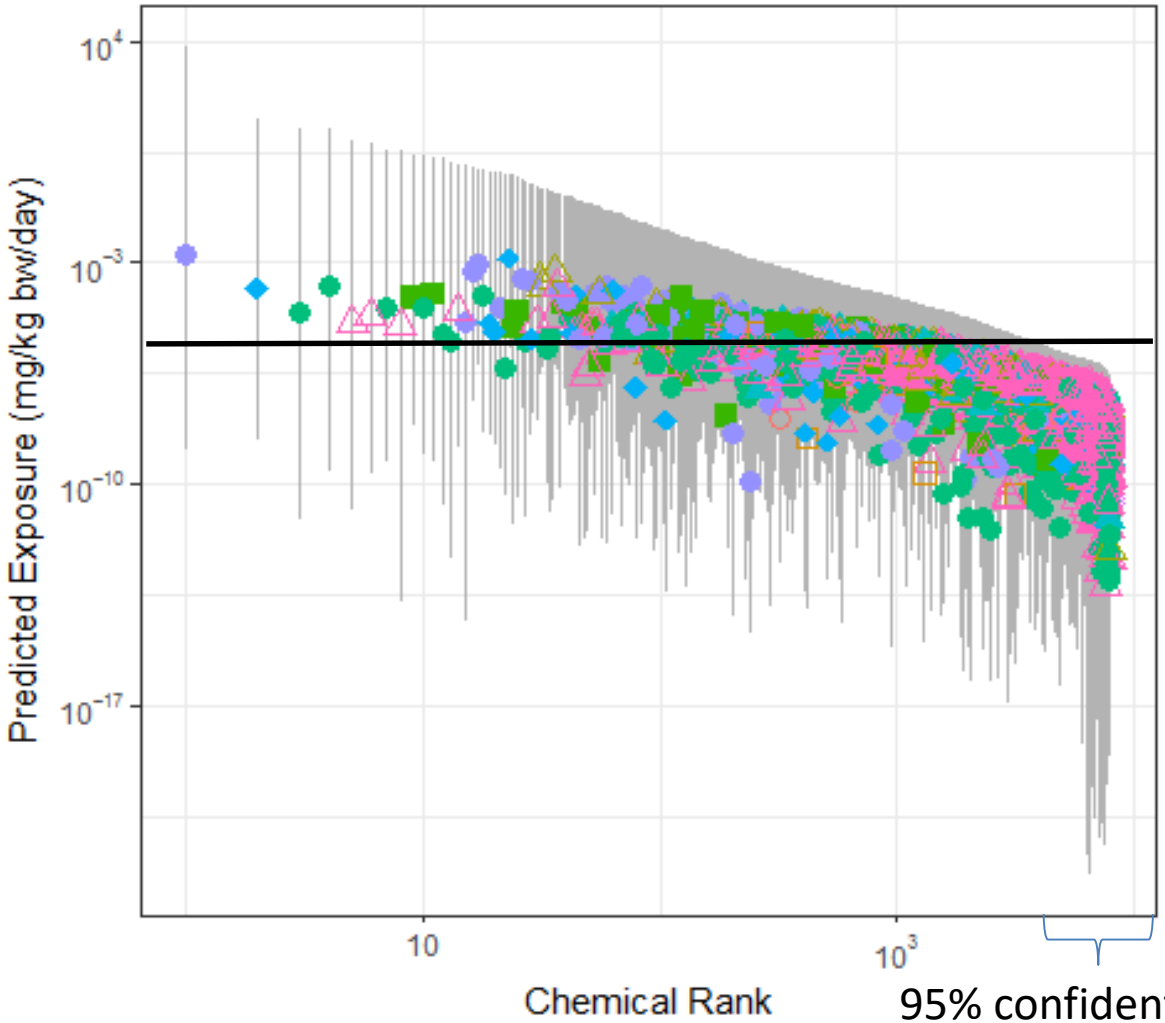
Human Exposure Predictions for 134,521 Chemicals



- Pathway predictions can be used for large chemical libraries
- Use prediction (and accuracy of prediction) as a prior for Bayesian analysis
- Each chemical may have exposure by multiple pathways

Ring et al. (*in prep.*)

Human Exposure Predictions for 134,521 Chemicals



Lowest NHANES limit of detection (LOD) roughly corresponds to ~10^{-6} mg/kg BW/day

95% confident that median population would be <LOD for thousands of chemicals

Ring et al. (*in prep.*)

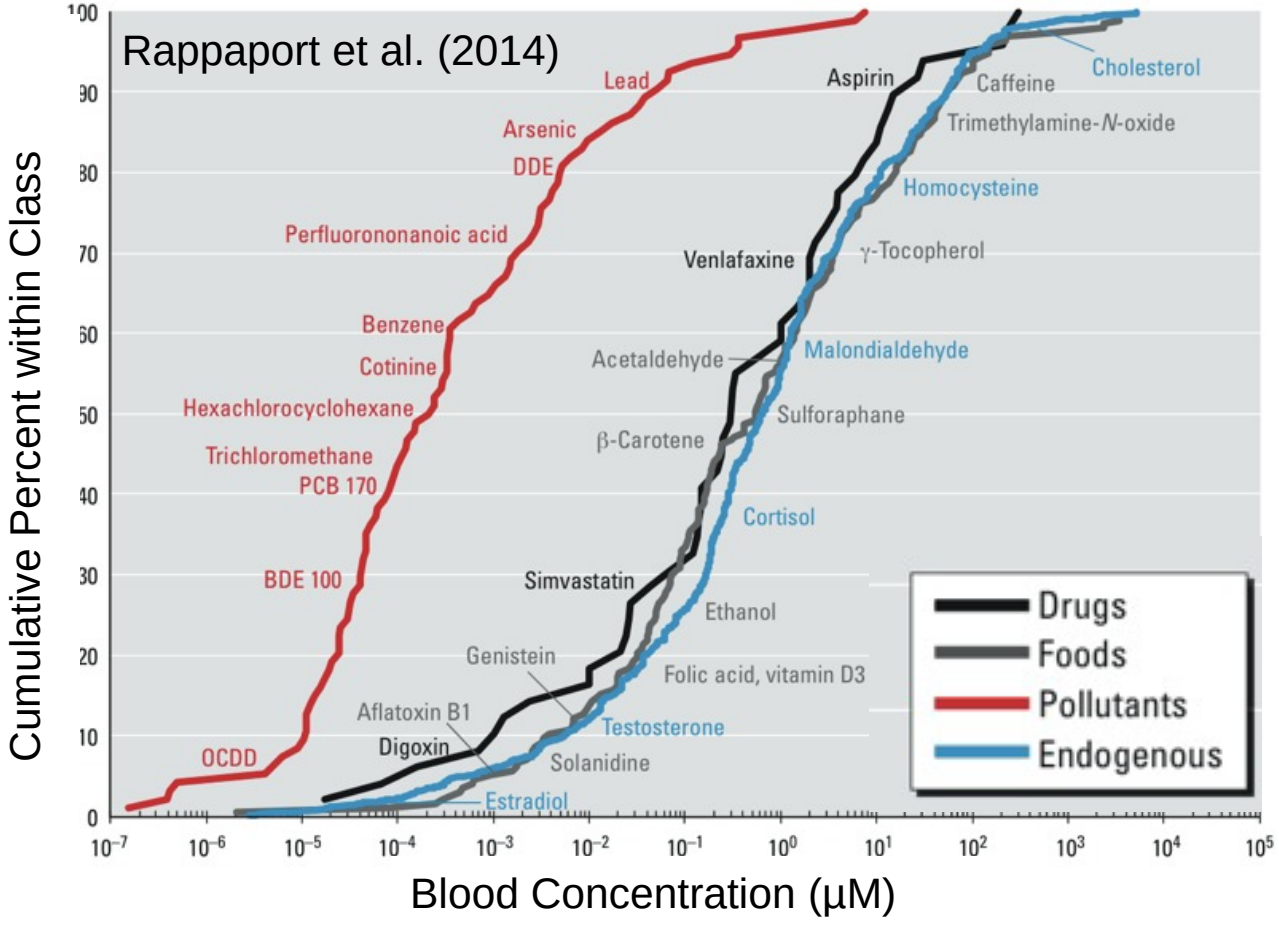
Conclusions

- Rough exposure assessments may be potentially useful if the uncertainty can be quantified and is acceptable (i.e., “fit for purpose”)

- Models incorporate Knowledge, Assumptions and Data (Macleod, et al., 2010)

- The trick is to know which model to use and when

- Using existing chemical data to predict pathways
  - Need better training data for random forest
  - (How do you know something isn’t an industrial chemical?)
- Eventually we have got to go beyond NHANES (~100 chemicals)
  - Non-targeted analysis of blood may eventually be possible



EPA United States Environmental Protection Agency

Chemical Safety for Sustainability (CSS) Research Program

Rapid Exposure and Dosimetry (RED) Project

NCCT
Chris Grulke
Greg Honda*
Richard Judson
Andrew McEachran*
Robert Pearce*
Ann Richard
Risa Sayre*
Woody Setzer
Rusty Thomas
John Wambaugh
Antony Williams

NRMRL
Yirui Liang*
Xiaoyu Liu

NHEERL
Linda Adams
Christopher Ecklund
Marina Evans
Mike Hughes
Jane Ellen Simmons

NERL
Craig Barber
Namdi Brandon*
Peter Egeghy
Jarod Grossman*
Hongtai Huang*
Brandall Ingle*
Kristin Isaacs
Sarah Laughlin-Toth*
Seth Newton
Katherine Phillips
Paul Price
Jeanette Reyes*
Jon Sobus
John Streicher*
Mark Strynar
Mike Tornero-Velez
Elin Ulrich
Dan Vallero
Barbara Wetmore

*Trainees

Lead CSS Matrix Interfaces:
John Kenneke (NERL)
John Cowden (NCCT)

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

Collaborators

Arnot Research and Consulting
Jon Arnot
Johnny Westgate
Battelle Memorial Institute
Anne Louise Sumner
Anne Gregg
Chemical Computing Group
Rocky Goldsmith
National Institute for Environmental Health Sciences (NIEHS) National Toxicology Program
Mike Devito
Steve Ferguson
Nisha Sipes
Netherlands Organisation for Applied Scientific Research (TNO)
Sieto Bosgra
Research Triangle Institute
Timothy Fennell
ScitoVation
Harvey Clewell
Kamel Mansouri
Chantel Nicolas
Silent Spring Institute
Robin Dodson
Southwest Research Institute
Alice Yau
Kristin Favela
Summit Toxicology
Lesa Aylward
Tox Strategies
Caroline Ring
University of California, Davis
Deborah Bennett
Hyeong-Moo Shin
University of Michigan
Olivier Jolliet
University of North Carolina, Chapel Hill
Alex Tropsha

United States Environmental Protection Agency

References

- Arnot, Jon A., et al. "Screening level risk assessment model for chemical fate and effects in the environment." Environmental science & technology 40.7 (2006): 2316-2323.
- Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
- Dionisio, Kathie L., et al. “Exploring Consumer Exposure Pathways and Patterns of Use for Chemicals in the Environment.” Toxicology Reports (2015)
- Egeghy, Peter P., et al. "The exposure data landscape for manufactured chemicals." Science of the Total Environment 414: 159-166 (2012)
- Isaacs, Kristin K., et al. "SHEDS-HT: an integrated probabilistic exposure model for prioritizing exposures to chemicals with near-field and dietary sources." Environmental science & technology 48.21 (2014): 12750-12759.
- MacLeod, Matthew, et al. "The state of multimedia mass-balance modeling in environmental science and decision-making." (2010): 8360-8364
- Mansouri, Kamel, et al. “OPERA (OPEn saR App) ” in preparation
- National Academies of Sciences, Engineering, and Medicine. Using 21st century science to improve risk-related evaluations. National Academies Press, 2017.
- Park, Youngja H., et al. "High-performance metabolic profiling of plasma from seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring." Toxicology 295.1 (2012): 47-55.
- Phillips, Katherine A., et al. “Suspect Screening Analysis of Chemicals in Consumer Products”, submitted.
- Rappaport, Stephen M., et al. "The blood exposome and its role in discovering causes of disease." Environmental Health Perspectives (Online) 122.8 (2014): 769.,
- Ring, Caroline, et al.. “Chemical Exposure Pathway Prediction for Screening and Prioritization,” in preparation
- Rosenbaum, Ralph K., et al. "USEtox—the UNEP-SETAC toxicity model: recommended characterisation factors for human toxicity and freshwater ecotoxicity in life cycle impact assessment." The International Journal of Life Cycle Assessment 13.7 (2008): 532.
- Shin, Hyeong-Moo, et al. "Risk-based high-throughput chemical screening and prioritization using exposure models and in vitro bioactivity assays." Environmental science & technology 49.11 (2015): 6760-6771.
- Wallace et al., “The TEAM Study: Personal exposures to toxic substances in air, drinking water, and breath of 400 residents of New Jersey, North Carolina, and North Dakota .” Environmental Research 43: 209-307 (1987)
- Wambaugh, John F., et al. "High-throughput models for exposure-based chemical prioritization in the ExpoCast project." Environmental science & technology 47.15 (2013): 8479-848.
- Wambaugh, John F., et al. "High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals." Environmental science & technology (2014).
- Wetmore, Barbara A., et al. "Incorporating high-throughput exposure predictions with dosimetry-adjusted in vitro bioactivity to inform chemical toxicity testing." Toxicological Sciences 148.1 (2015): 121-136.
- Yang, Chihae, et al. "New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling." Journal of chemical information and modeling 55.3 (2015): 510-528.