United States
Environmental Protection
Agency

Predicting Chemical Exposure Pathways

John F. Wambaugh

National Center for Computational Toxicology
Office of Research and Development
United States Environmental Protection Agency
Research Triangle Park, North Carolina 27711

National Academics of Sciences, Engineering, & Medicine
Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions
June 6, 2019

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

https://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
  - 562 peer-reviewed journal articles in 2018

- Research is conducted by ORD's three national laboratories, four national centers, and two offices organized to address:
  - Hazard, exposure, risk assessment, and risk management

- 13 facilities across the United States

- Research conducted by a combination of Federal scientists (including uniformed members of the **Public Health Service)**; contract researchers; and postdoctoral, graduate student, and post-baccalaureate trainees

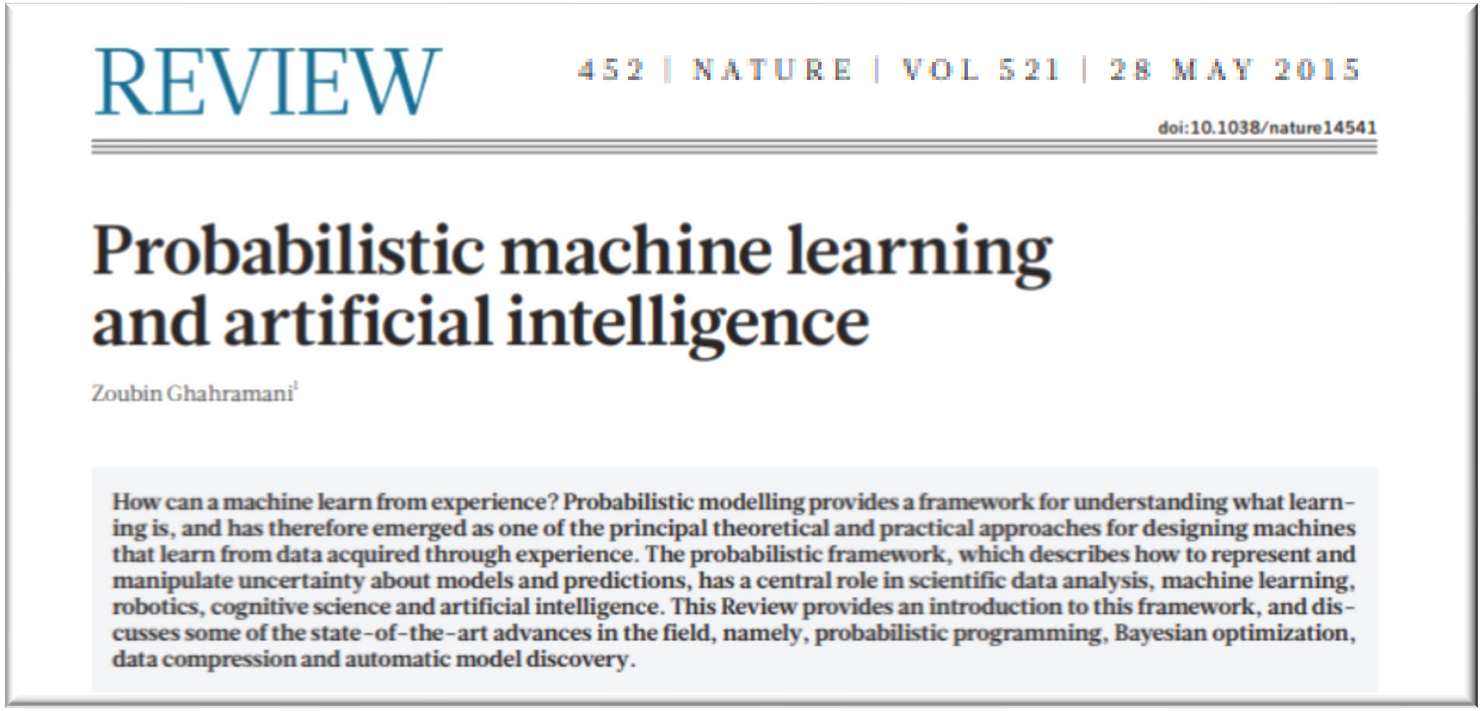
U.S. PUBLIC HEALTH SERVICE • 1798 •

Credit: the Research Triangle Foundati

ORD Facility in
Research Triangle Park, NC

Machine Learning: A Subset of Artificial Intelligence

"…machine learning can be thought of as inferring plausible models to explain observed data."



REVIEW

452 | NATURE | VOL 521 | 28 MAY 2015

doi:10.1038/nature14541

Probabilistic machine learning and artificial intelligence

Zoubin Ghahramani[1]

How can a machine learn from experience? Probabilistic modelling provides a framework for understanding what learning is, and has therefore emerged as one of the principal theoretical and practical approaches for designing machines that learn from data acquired through experience. The probabilistic framework, which describes how to represent and manipulate uncertainty about models and predictions, has a central role in scientific data analysis, machine learning, robotics, cognitive science and artificial intelligence. This Review provides an introduction to this framework, and discusses some of the state-of-the-art advances in the field, namely, probabilistic programming, Bayesian optimization, data compression and automatic model discovery.

At the EPA we are applying publicly available machine learning algorithms to bridge data gaps and draw inferences from complex data sets.

United States
Environmental Protection
Agency

Chemical Regulation in the United States

- Park *et al.* (2012): At least 3221 chemical signatures in pooled human blood samples, many appear to be exogenous

- A tapestry of laws covers the chemicals people are exposed to in the United States (Breyer, 2009)

- Different testing requirements exist for food additives, pharmaceuticals, and pesticide active ingredients (NRC, 2007)

- Most other chemicals, ranging from industrial waste to dyes to packing materials, are covered by the Toxic Substances Control Act (TSCA)

GIVE A DOG A PHONE
Technology for our furry friends

NewScientist

WEEKLY November 29 - December 5, 2014

We've made
150,000 new chemicals

We touch them,
we wear them, we eat them

But which ones should
we worry about?

SPECIAL REPORT, *page 34*

THE GOOD FIGHT
Most violence
is also virtuous

CHAMBER OF SECRETS
The greatest ever find
of early human bones

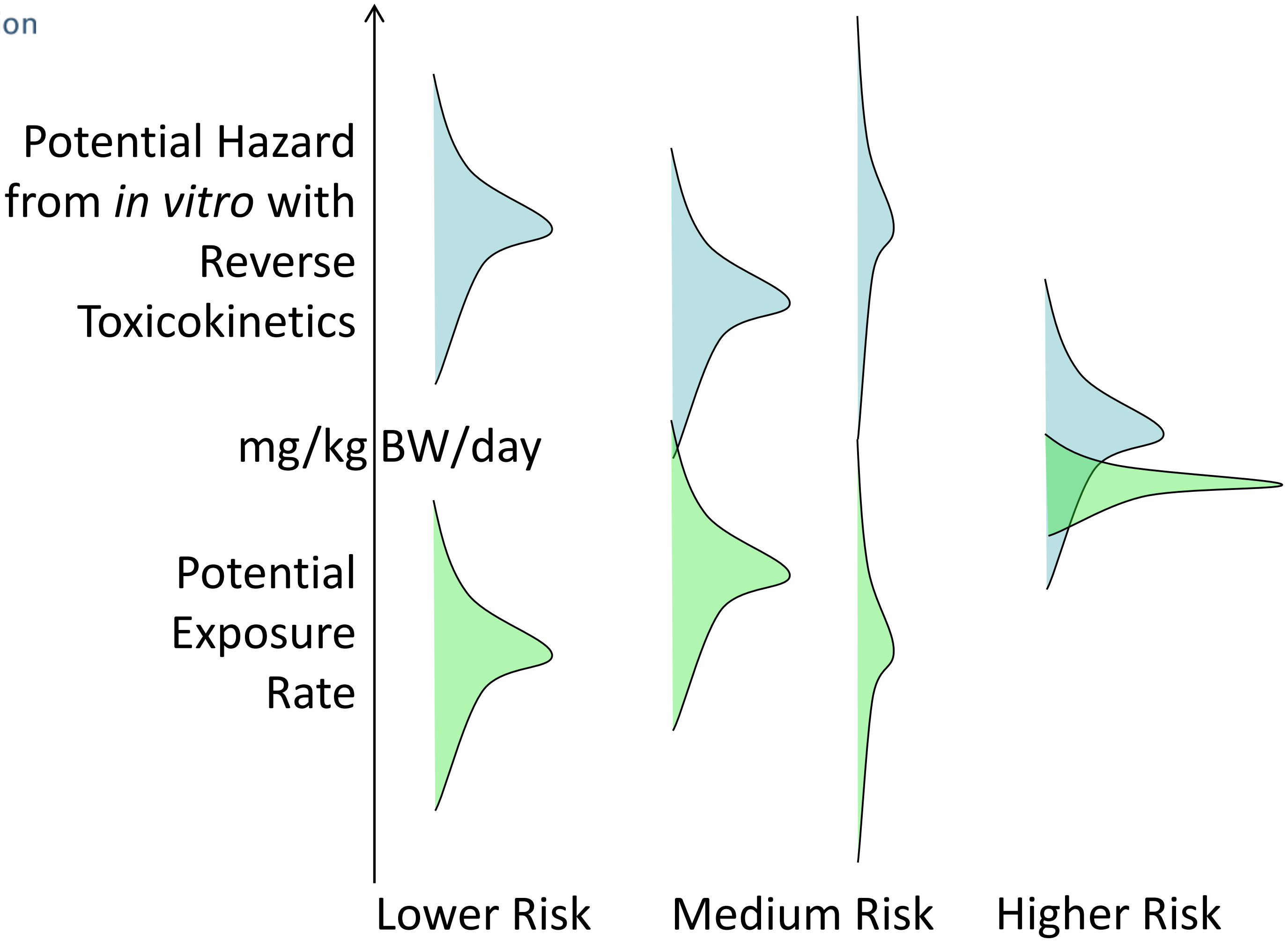
IS IT ALIVE?
Artificial worm could
be first digital animal

No2997 US$5.95 CAN$5.95

Science and technology news www.newscientist.com US jobs in science

November 29, 2014

Chemical Risk Assessment in the 21st Century



Potential Hazard from *in vitro* with Reverse Toxicokinetics

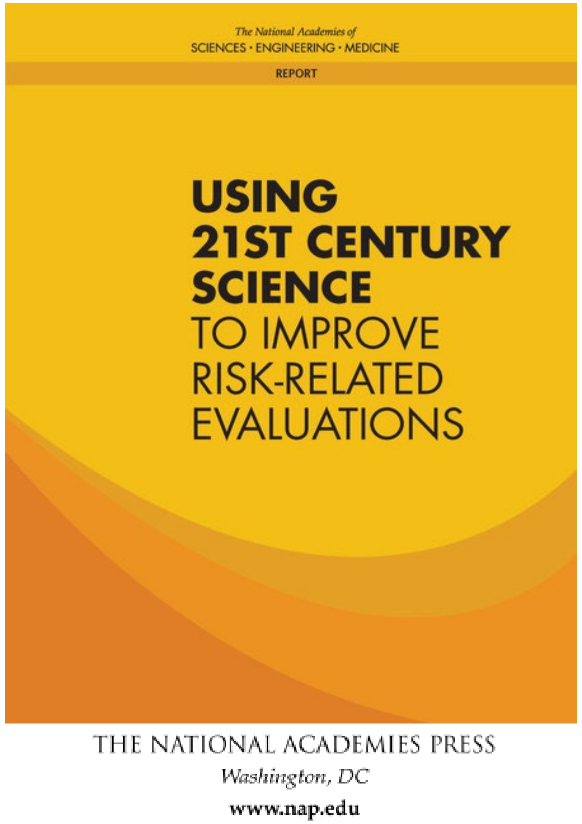
mg/kg BW/day

Potential Exposure Rate

Lower Risk

Medium Risk

Higher Risk



The National Academies of
SCIENCES · ENGINEERING · MEDICINE

REPORT

USING 21ST CENTURY SCIENCE TO IMPROVE RISK-RELATED EVALUATIONS

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

January, 2017

"…The committee sees the potential for the application of **computational exposure science** to be highly valuable and credible for comparison and priority-setting among chemicals in a **risk-based context**."

What Do We Know About Exposure? Biomonitoring Data

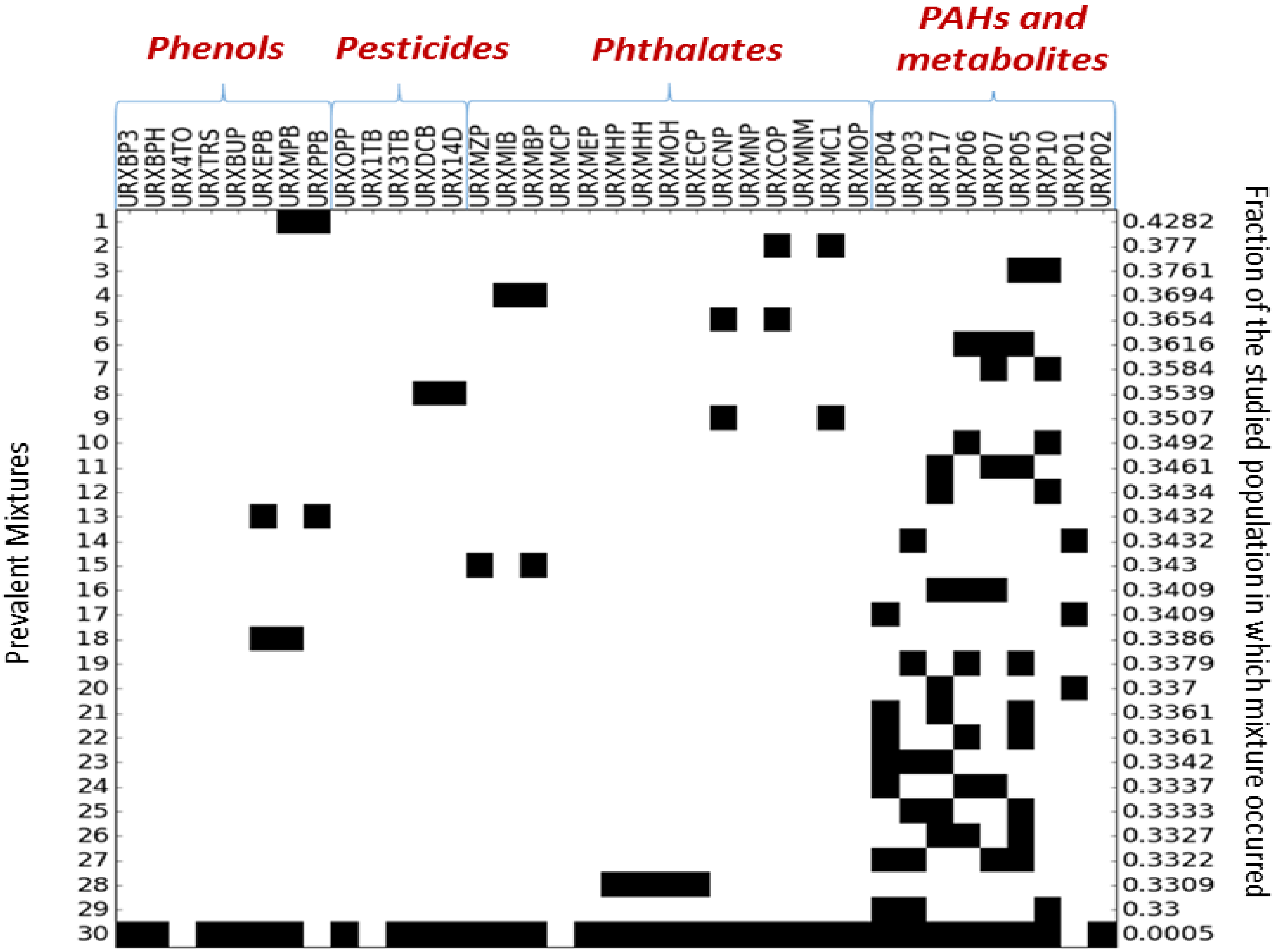
- Centers for Disease Control and Prevention (CDC) National Health and Nutrition Examination Survey (NHANES) provides an important tool for monitoring public health
- Large, ongoing CDC survey of US population: demographic, body measures, medical exam, biomonitoring (health and exposure), …
- Designed to be representative of US population according to census data
- Data sets publicly available (http://www.cdc.gov/nchs/nhanes.htm)
- Includes measurements of:
  - Body weight
  - Height
  - **Chemical analysis of blood and urine**

nhanes
National Health and Nutrition Examination Survey

Identifying Prevalent Mixtures in the NHANES Data

- We used data-mining methods (frequent itemset mining or FIM, Borgelt, 2012) to identify combinations of items (chemicals) that co-occur together within samples from same individual

- Identified a few dozen mixtures present in >30% of U.S. population

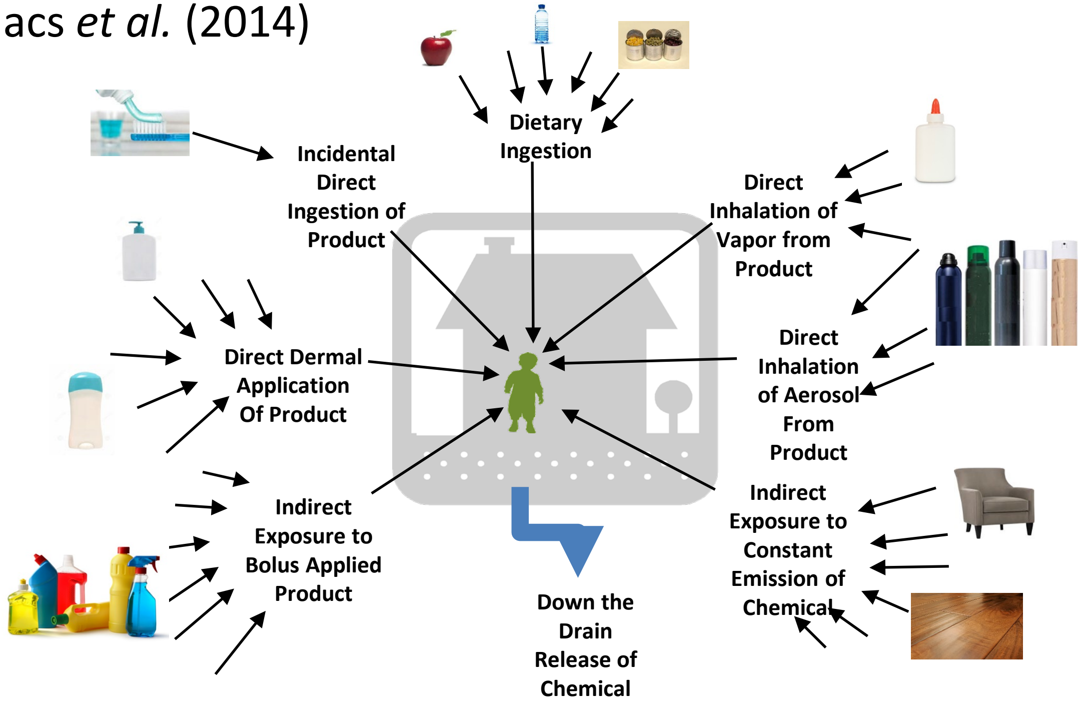


Kapraun *et al*. (2017)

What Else Do We Know About Exposure? Exposure Models

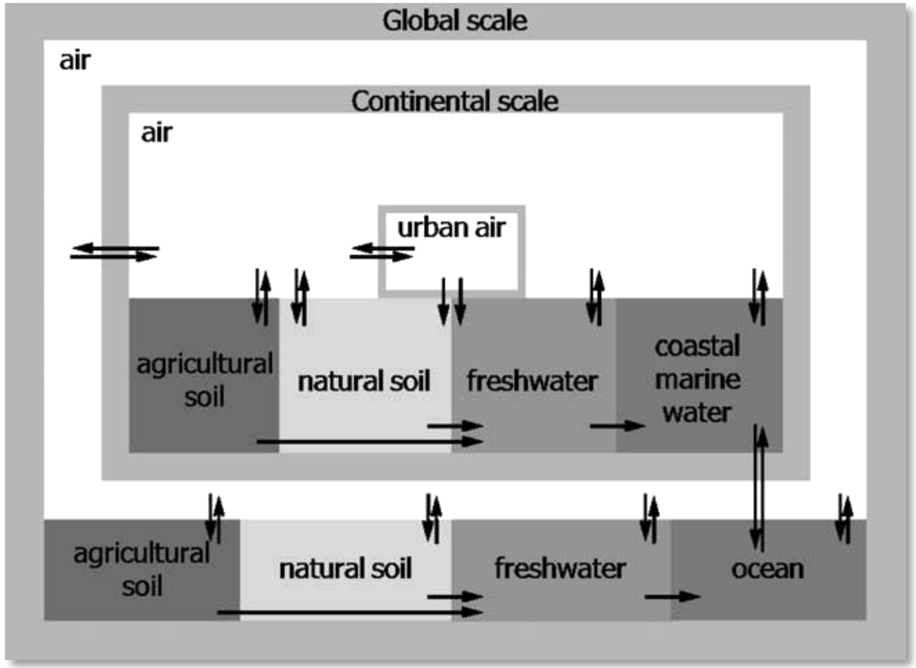
- A model captures knowledge and a hypothesis of how the world works (MacLeod *et al.*, 2010)

Isaacs *et al.* (2014)



Dietary Ingestion
Incidental Direct Ingestion of Product
Direct Inhalation of Vapor from Product
Direct Dermal Application Of Product
Direct Inhalation of Aerosol From Product
Indirect Exposure to Bolus Applied Product
Indirect Exposure to Constant Emission of Chemical
Down the Drain Release of Chemical

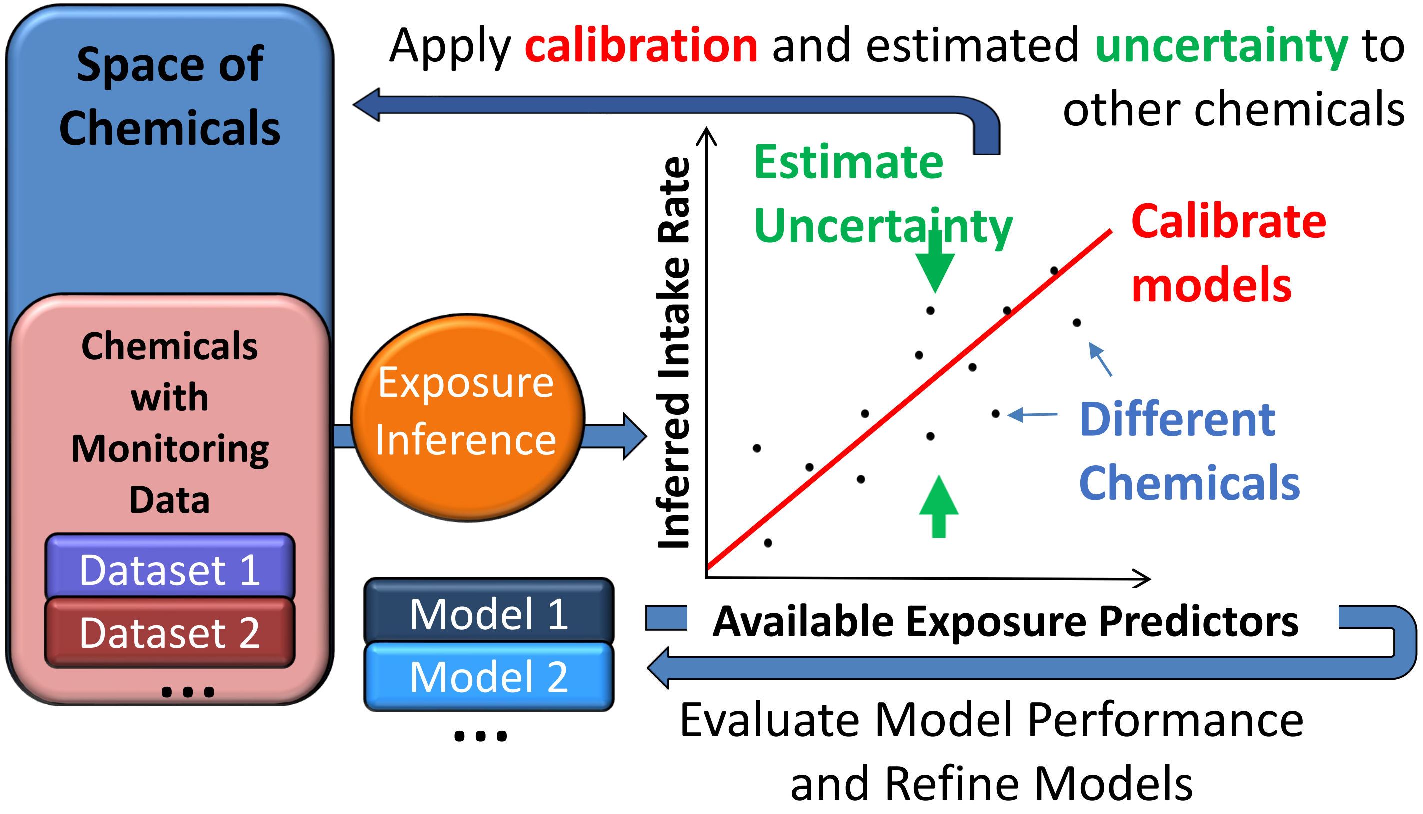
Rosenbaum *et al.* (2008)



“Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations… The only question of interest is ‘Is the model illuminating and useful?’” George Box

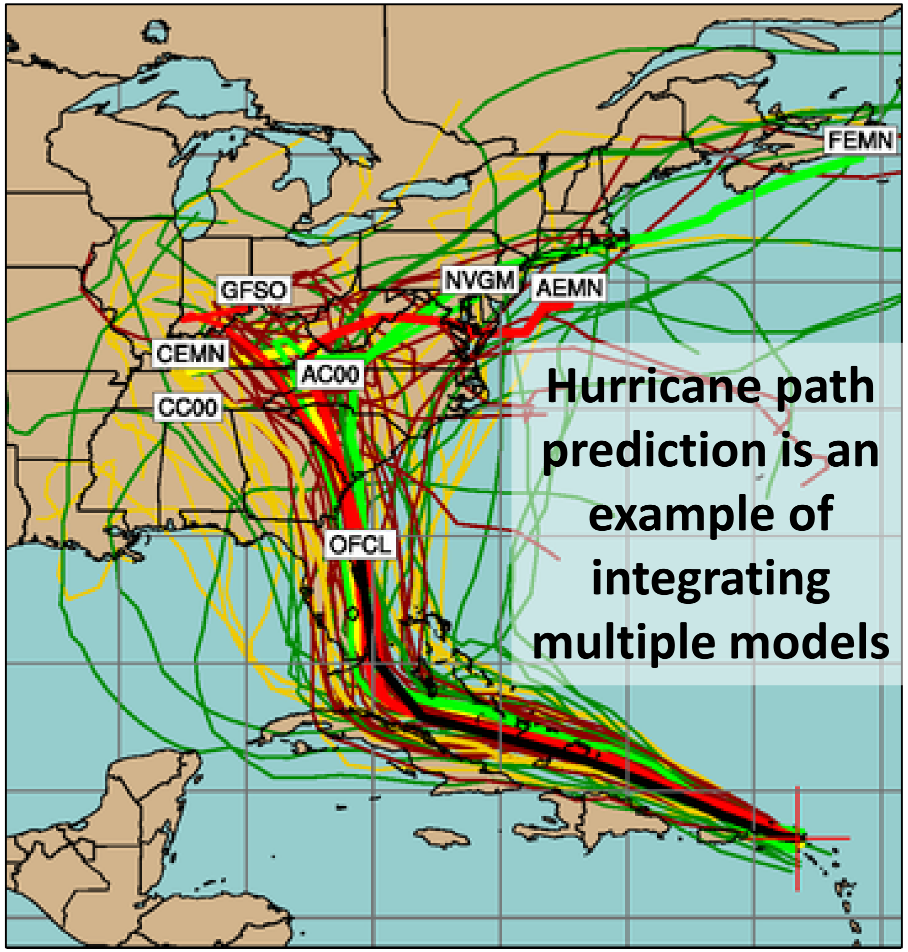
Consensus Exposure Predictions with the SEEM Framework

- We use Bayesian methods to incorporate multiple models into consensus predictions for 1000s of chemicals within the **Systematic Empirical Evaluation of Models (SEEM)** (Wambaugh et al., 2013, 2014; Ring et al., 2018)



Apply calibration and estimated uncertainty to other chemicals

Evaluate Model Performance and Refine Models

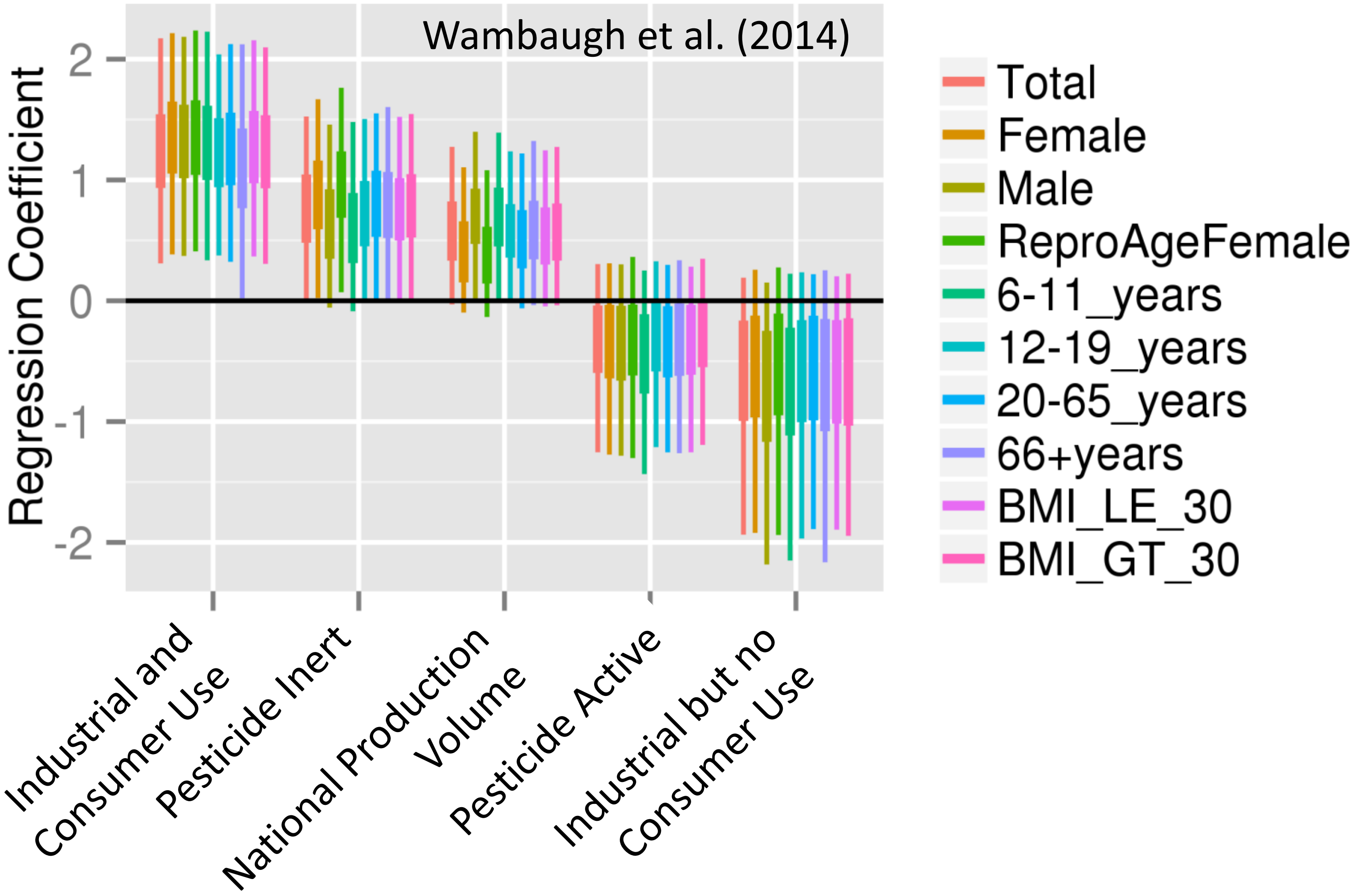


Hurricane path prediction is an example of integrating multiple models

United States Environmental Protection Agency

Heuristics of Exposure

This is just a fancy linear regression

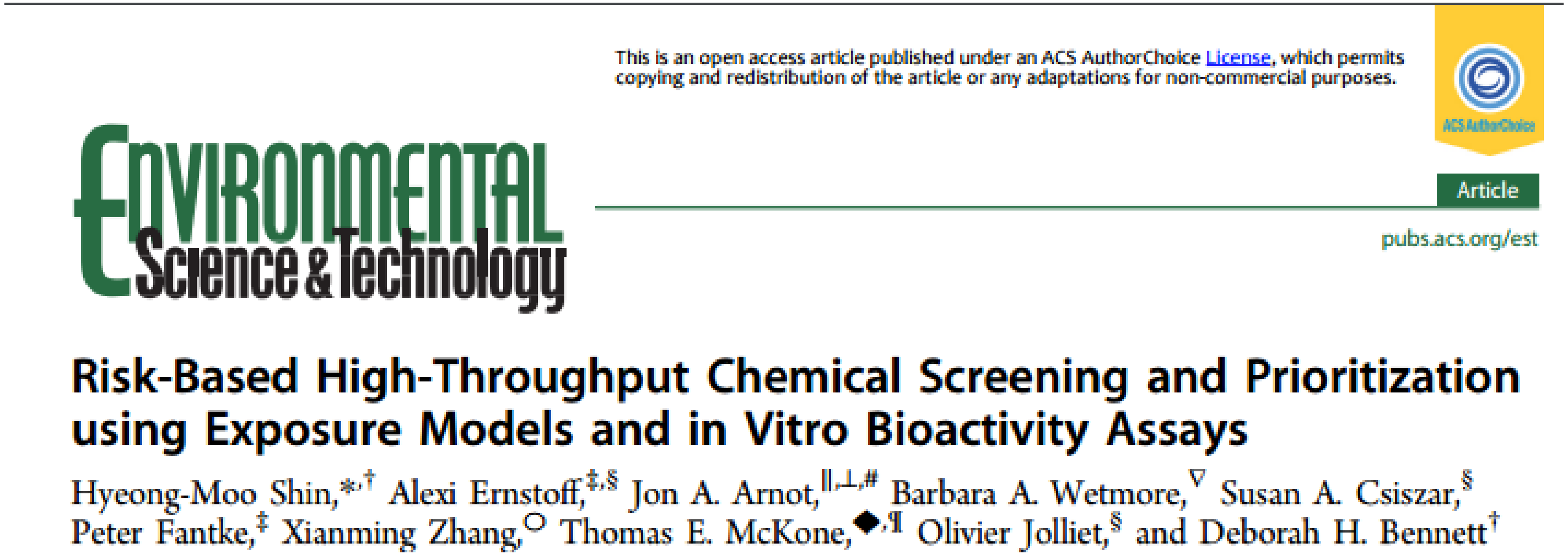


Same five predictors work for all NHANES demographic groups analyzed – stratified by age, sex, and body-mass index:

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

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."



This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS AuthorChoice

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[†]

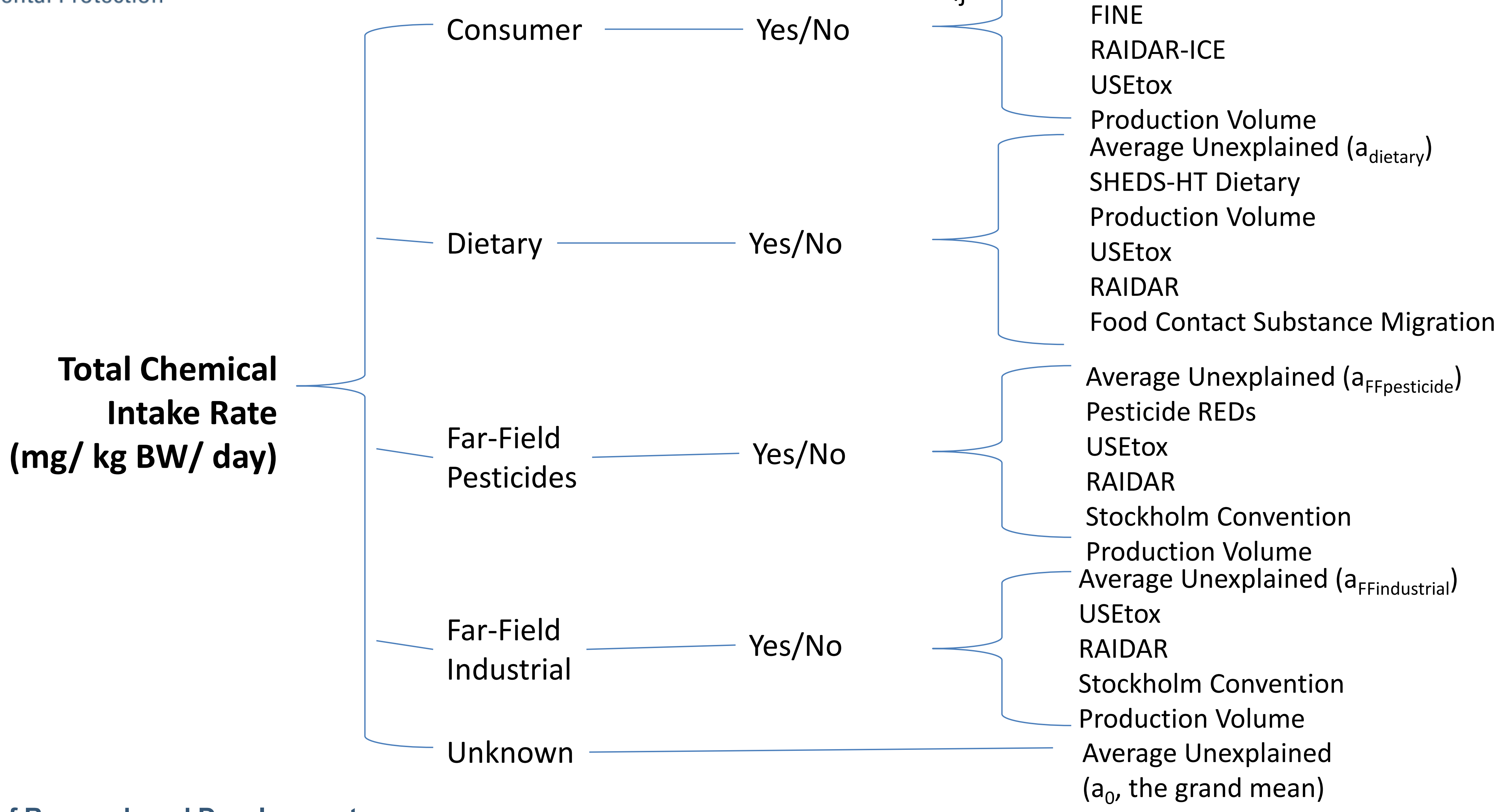
Collaboration on High Throughput Exposure Predictions

Jon Arnot, Deborah H. Bennett, Peter P. Egeghy, Peter Fantke, Lei Huang, Kristin K. Isaacs, Olivier Jolliet, Hyeong-Moo Shin, Katherine A. Phillips, Caroline Ring, R. Woodrow Setzer, John F. Wambaugh, Johnny Westgate

Predictor	Reference(s)	Chemicals Predicted	Pathways
EPA Inventory Update Reporting and Chemical Data Reporting (CDR) (2015)	US EPA (2018)	7856	All
Stockholm Convention of Banned Persistent Organic Pollutants (2017)	Lallas (2001)	248	Far-Field Industrial and Pesticide
EPA Pesticide Reregistration Eligibility Documents (REDs) Exposure Assessments (Through 2015)	Wetmore et al. (2012, 2015)	239	Far-Field Pesticide
United Nations Environment Program and Society for Environmental Toxicology and Chemistry toxicity model (USEtox) Industrial Scenario (2.0)	Rosenbaum et al. (2008)	8167	Far-Field Industrial
USEtox Pesticide Scenario (2.0)	Fantke et al. (2011, 2012, 2016)	940	Far-Field Pesticide
Risk Assessment IDentification And Ranking (RAIDAR) Far-Field (2.02)	Arnot et al. (2008)	8167	Far-Field Pesticide
EPA Stochastic Human Exposure Dose Simulator High Throughput (SHEDS-HT) Near-Field Direct (2017)	Isaacs (2017)	7511	Far-Field Industrial and Pesticide
SHEDS-HT Near-field Indirect (2017)	Isaacs (2017)	1119	Residential
Fugacity-based INdoor Exposure (FINE) (2017)	Bennett et al. (2004), Shin et al. (2012)	645	Residential
RAIDAR-ICE Near-Field (0.803)	Arnot et al., (2014), Zhang et al. (2014)	1221	Residential
USEtox Residential Scenario (2.0)	Jolliet et al. (2015), Huang et al. (2016,2017)	615	Residential
USEtox Dietary Scenario (2.0)	Jolliet et al. (2015), Huang et al. (2016), Ernstoff et al. (2017)	8167	Dietary

Ring *et al.* (2018)

Organizing Models by Pathway



Pathway

Chemical-Specific Pathway Relevancy (δ_{ij})

Exposure Predictors

Total Chemical Intake Rate (mg/ kg BW/ day)

Consumer — Yes/No
- Average Unexplained ($a_{consumer}$)
- SHEDS-HT
- FINE
- RAIDAR-ICE
- USEtox
- Production Volume

Dietary — Yes/No
- Average Unexplained ($a_{dietary}$)
- SHEDS-HT Dietary
- Production Volume
- USEtox
- RAIDAR
- Food Contact Substance Migration

Far-Field Pesticides — Yes/No
- Average Unexplained ($a_{FFpesticide}$)
- Pesticide REDs
- USEtox
- RAIDAR
- Stockholm Convention
- Production Volume

Far-Field Industrial — Yes/No
- Average Unexplained ($a_{FFindustrial}$)
- USEtox
- RAIDAR
- Stockholm Convention
- Production Volume

Unknown — Average Unexplained (a_0, the grand mean)

Ring *et al.* (2018)

Machine Learning to Predicting Exposure Pathways

We use the method of Random Forests to relate chemical structure and properties to exposure pathway

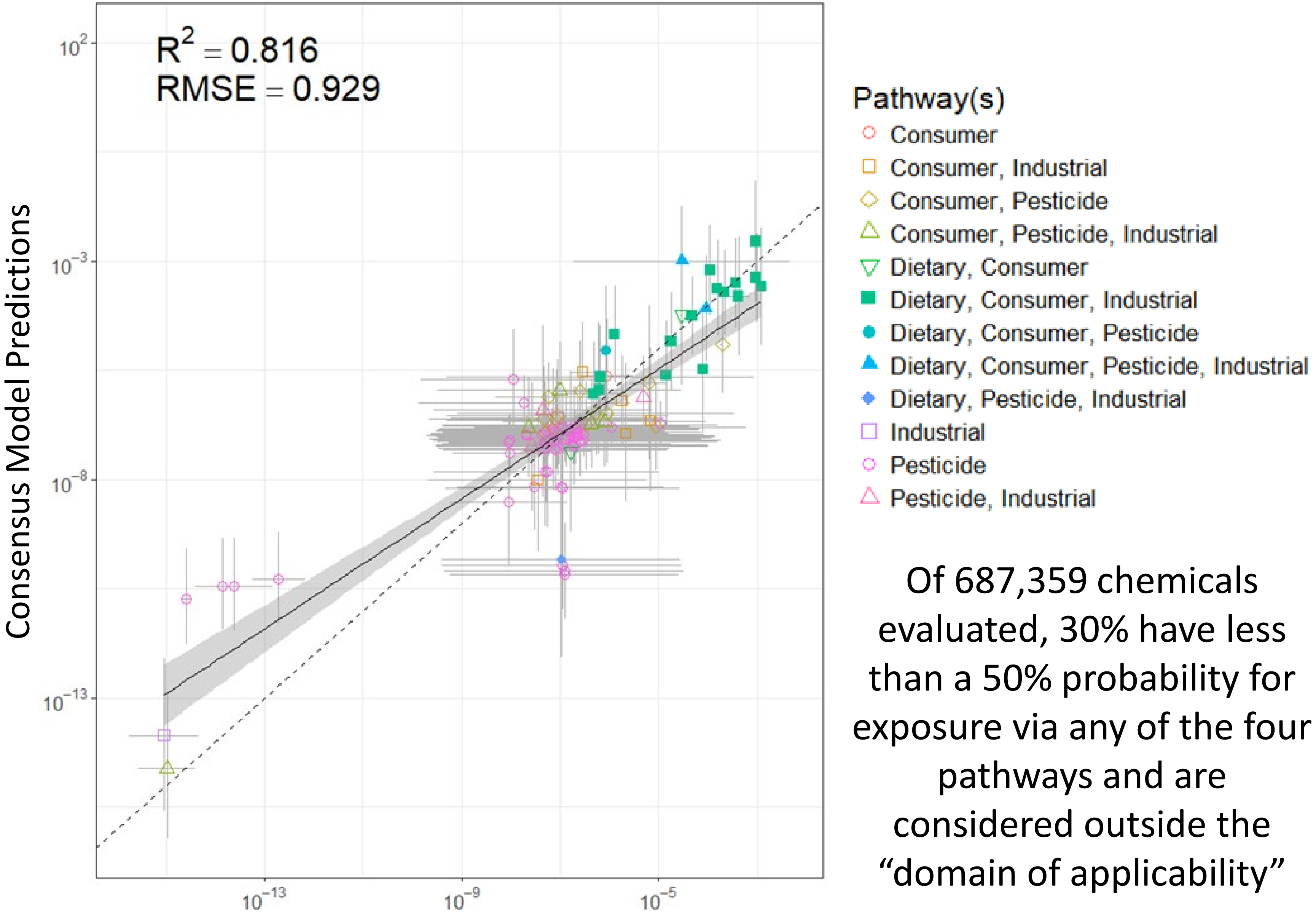
	NHANES Chemicals	Positives	Negatives	OOB Error Rate	Positives Error Rate	Balanced Accuracy	Sources of Positives	Sources of Negatives
Dietary	24	2523	8865	27	32	73	FDA CEDI, ExpoCast, CPDat (Food, Food Additive, Food Contact), NHANES Curation	Pharmapendium, CPDat (non-food), NHANES Curation
Near-Field	**49**	1622	567	26	24	74	CPDat (consumer_use, building_material), ExpoCast, NHANES Curation	CPDat (Agricultural, Industrial), FDA CEDI, NHANES Curation
Far-Field Pesticide	**94**	1480	6522	21	36	80	REDs, Swiss Pesticides, Stockholm Convention, CPDat (Pesticide), NHANES Curation	Pharmapendium, Industrial Positives, NHANES Curation
Far Field Industrial	**42**	5089	2913	19	16	81	CDR HPV, USGS Water Occurrence, NORMAN PFAS, Stockholm Convention, CPDat (Industrial, Industrial_Fluid), NHANES Curation	Pharmapendium, Pesticide Positives, NHANES Curation

Ring *et al.* (2018)

United States Environmental Protection Agency

Pathway-Based Consensus Modeling of NHANES

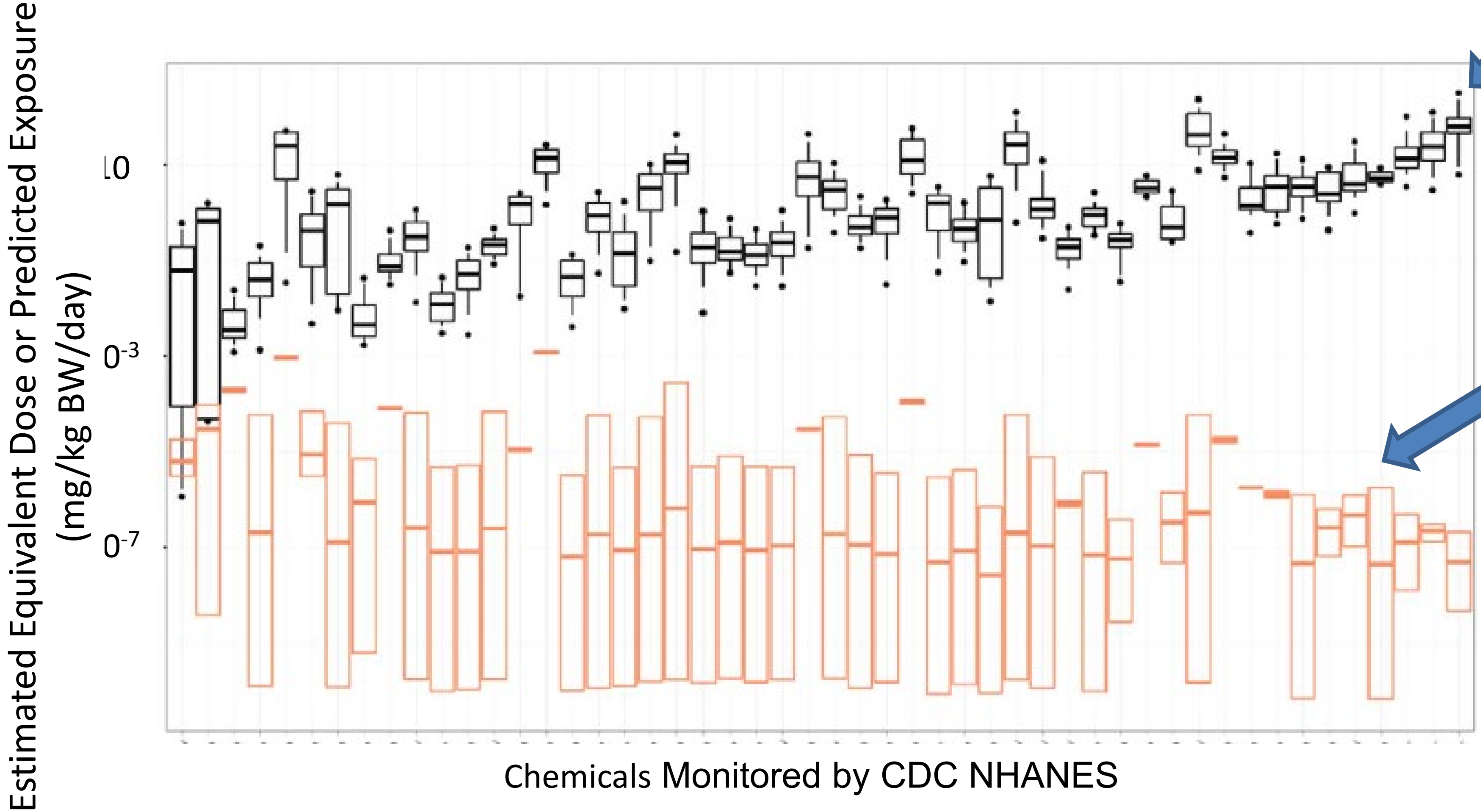
- Machine learning models were built for each of four exposure pathways
- Pathway predictions can be used for large chemical libraries
- Use prediction (and accuracy of prediction) as a prior for Bayesian analysis
- Each chemical may have exposure by multiple pathways



Of 687,359 chemicals evaluated, 30% have less than a 50% probability for exposure via any of the four pathways and are considered outside the "domain of applicability"

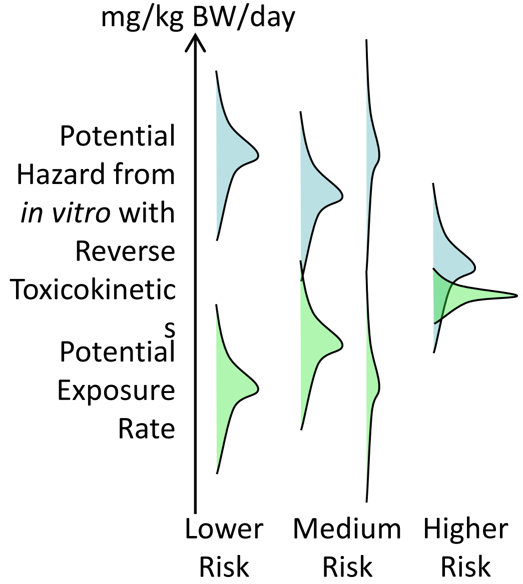
Ring *et al.* (2018)

Exposure Estimates Allow Chemical Prioritization



High throughput in vitro screening can estimate doses needed to cause bioactivity (e.g., Wetmore et al., 2015)

Exposure intake rates can be inferred from biomarkers (e.g., Ring et al., 2018)


mg/kg BW/day

Potential Hazard from *in vitro* with Reverse Toxicokinetics

Potential Exposure Rate

Lower Risk

Medium Risk

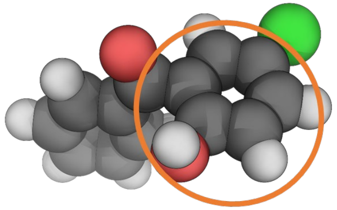
Higher Risk

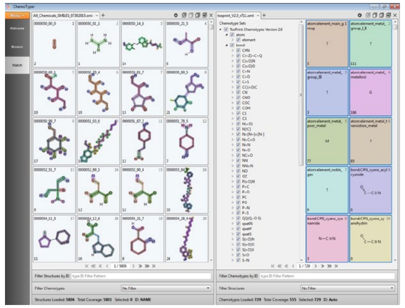
Ring *et al*. (2017)

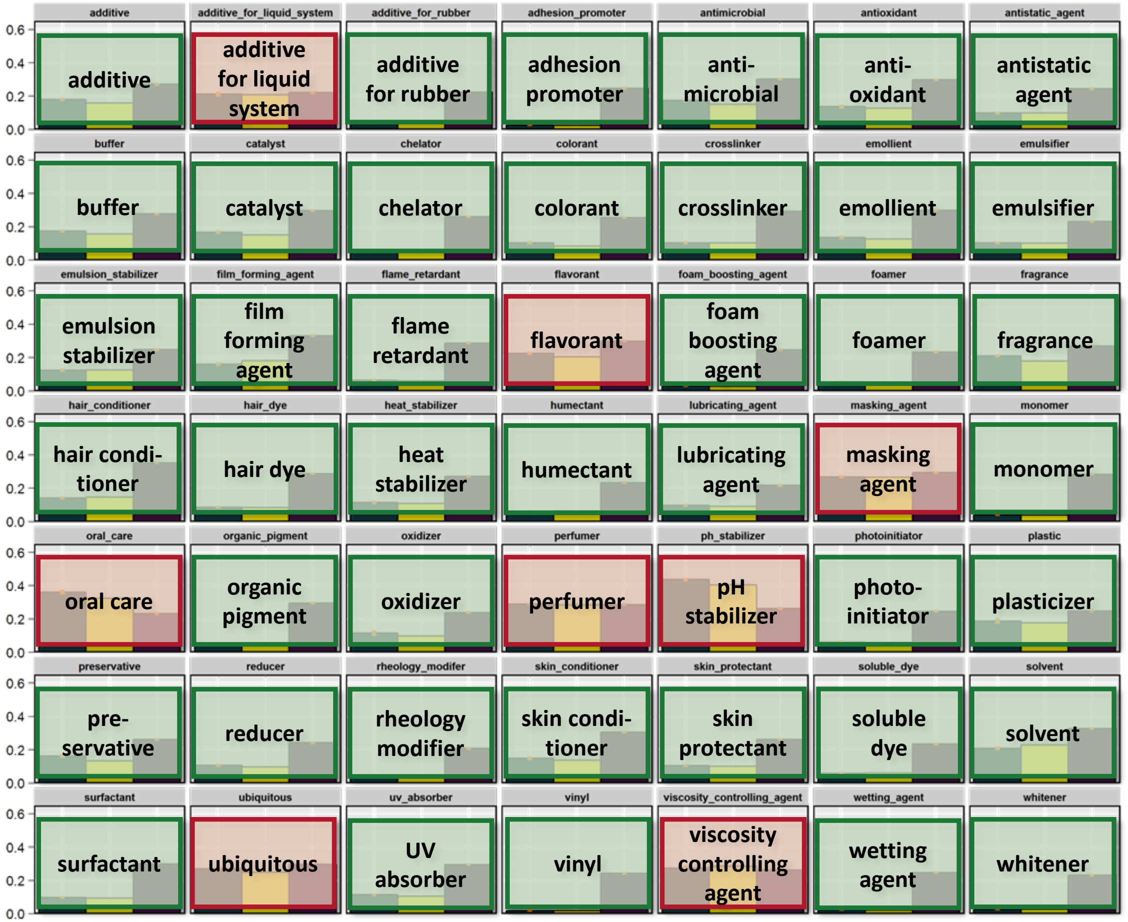
Predicting Chemical Function From Structure

Use Database (FUSE)

Chemical Structure and Property Descriptors







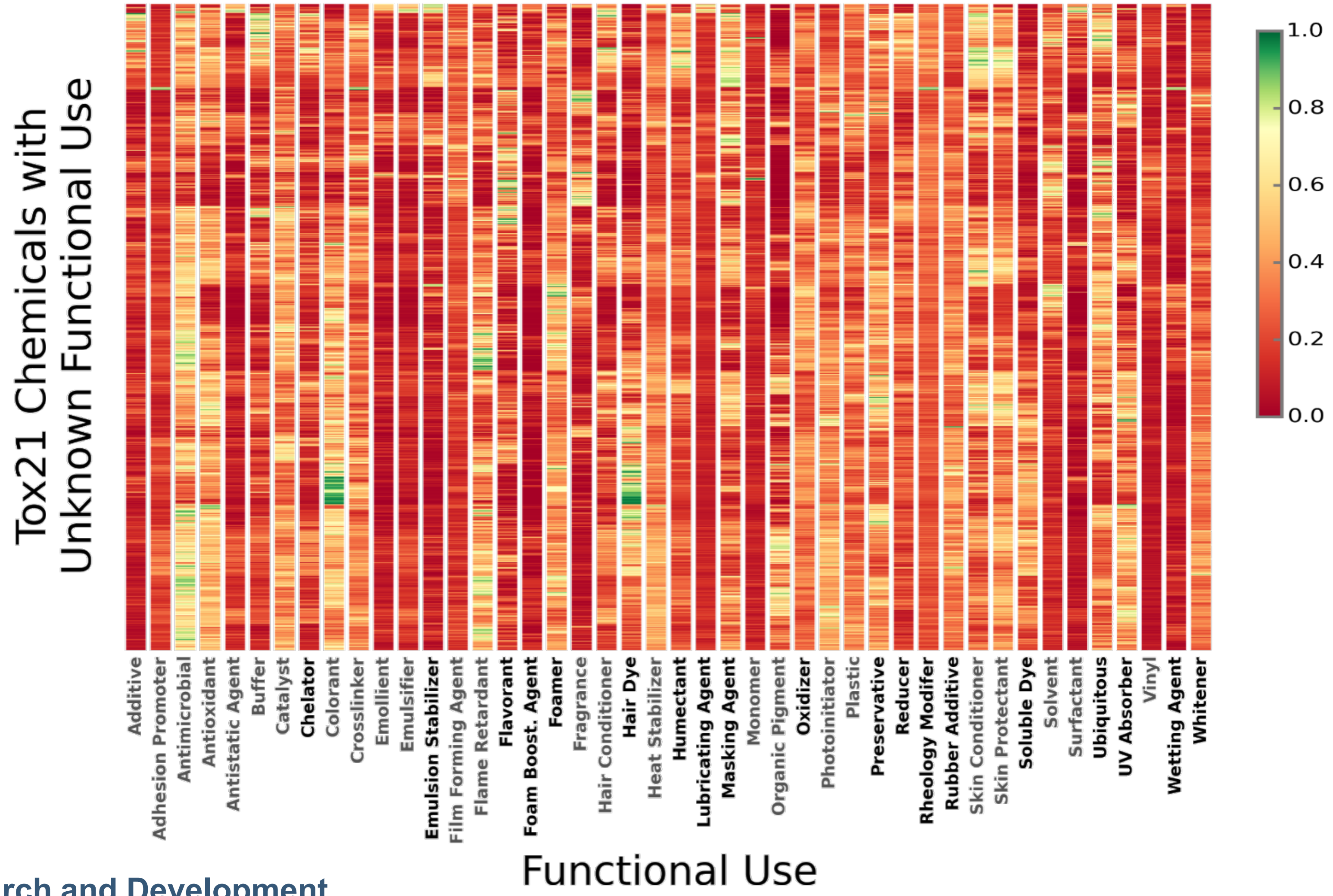
Machine Learning Based Classification Models
(Random Forest, Breiman, 2001)

Prediction of Of Potential Alternatives from Chemical Libraries

Phillips *et al.* (2017)

Screening for Alternatives By Function and Bioactivity

Combine high throughput screening data and chemical use prediction:



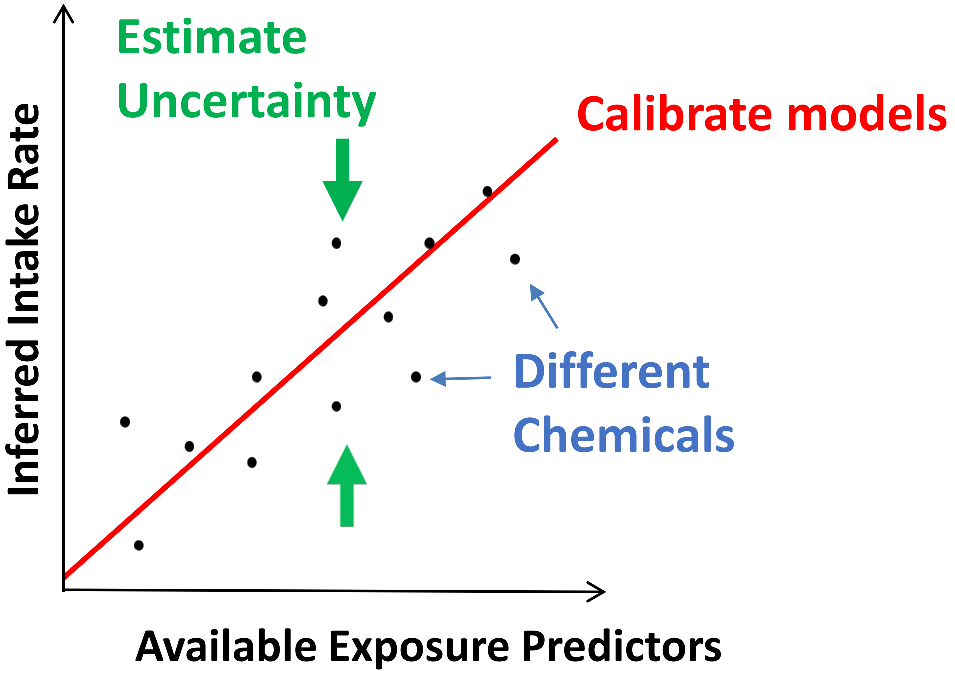
Probability of Chemical Performing Same Function

Phillips *et al.* (2017)

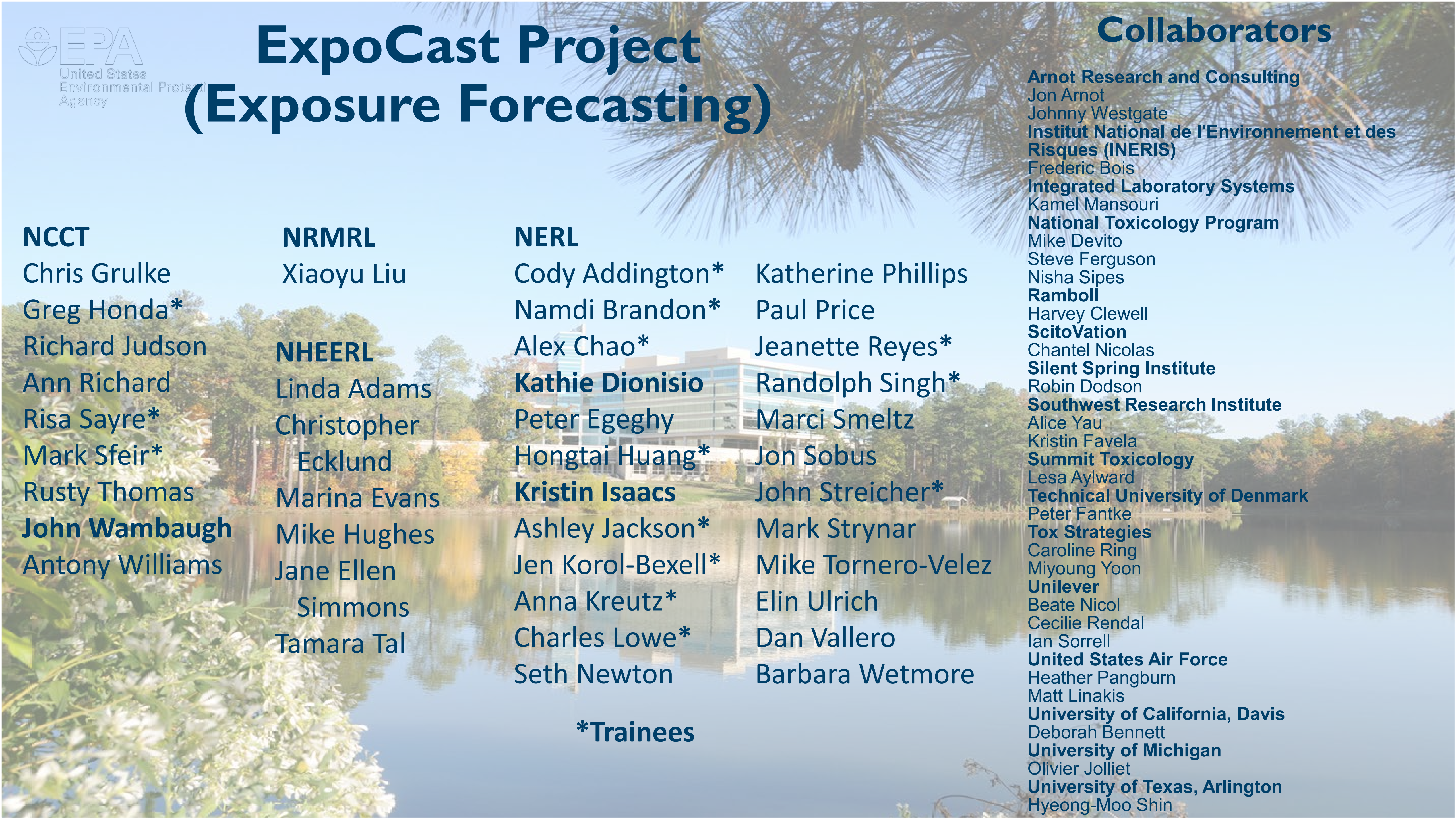
United States Environmental Protection Agency

Conclusions

- At the EPA we are applying publicly available machine learning algorithms to bridge data gaps and draw inferences from complex data sets.
- We can make chemical-specific estimates of intake rate for hundreds of thousands of chemical
  - Synthesizing as many models and other data as we can find
- Different models incorporate Knowledge, Assumptions and Data (Macleod, et al., 2010)
  - The trick is to know which model to use and when
  - Machine learning models allow educated guesses
- We are using existing chemical data to predict pathways
  - Not all chemicals fit within the domain of applicability
  - Need better training data for machine learning



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



United States
Environmental Protecti
Agency

ExpoCast Project (Exposure Forecasting)

NCCT
Chris Grulke
Greg Honda*
Richard Judson
Ann Richard
Risa Sayre*
Mark Sfeir*
Rusty Thomas
John Wambaugh
Antony Williams

NRMRL
Xiaoyu Liu

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

NERL
Cody Addington*
Namdi Brandon*
Alex Chao*
Kathie Dionisio
Peter Egeghy
Hongtai Huang*
Kristin Isaacs
Ashley Jackson*
Jen Korol-Bexell*
Anna Kreutz*
Charles Lowe*
Seth Newton

Katherine Phillips
Paul Price
Jeanette Reyes*
Randolph Singh*
Marci Smeltz
Jon Sobus
John Streicher*
Mark Strynar
Mike Tornero-Velez
Elin Ulrich
Dan Vallero
Barbara Wetmore

***Trainees**

Collaborators

Arnot Research and Consulting
Jon Arnot
Johnny Westgate
Institut National de l'Environnement et des Risques (INERIS)
Frederic Bois
Integrated Laboratory Systems
Kamel Mansouri
National Toxicology Program
Mike Devito
Steve Ferguson
Nisha Sipes
Ramboll
Harvey Clewell
ScitoVation
Chantel Nicolas
Silent Spring Institute
Robin Dodson
Southwest Research Institute
Alice Yau
Kristin Favela
Summit Toxicology
Lesa Aylward
Technical University of Denmark
Peter Fantke
Tox Strategies
Caroline Ring
Miyoung Yoon
Unilever
Beate Nicol
Cecilie Rendal
Ian Sorrell
United States Air Force
Heather Pangburn
Matt Linakis
University of California, Davis
Deborah Bennett
University of Michigan
Olivier Jolliet
University of Texas, Arlington
Hyeong-Moo Shin

References

- Arnot, J. A.; et al., Develop Sub-Module for Direct Human Exposures to Consumer Products. Technical Report for the U.S. Environmental Protection Agency; ARC Arnot Research & Consulting, Inc.: Toronto, ON, Canada, 2014.
- Bennett, D. H.; Furtaw, E. J., Fugacity-based indoor residential pesticide fate model. Environmental Science & Technology 2004, 38, (7), 2142-2152.
- Breyer, Stephen. Breaking the vicious circle: Toward effective risk regulation. Harvard University Press, 2009
- Burwell, Sylvia M., et al. "Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy--Managing Information as an Asset." (2013).
- Collins, Francis S., George M. Gray, and John R. Bucher. "Transforming environmental health protection." *Science (New York, NY)* 319.5865 (2008): 906.
- Dix, David J., et al. "The ToxCast program for prioritizing toxicity testing of environmental chemicals." *Toxicological Sciences* 95.1 (2006): 5-12.
- Egeghy, P. P., et al. (2012). The exposure data landscape for manufactured chemicals. Science of the Total Environment, 414, 159-166.
- Ernstoff, A. S., et al.., High-throughput migration modelling for estimating exposure to chemicals in food packaging in screening and prioritization tools. Food and Chemical Toxicology 2017, 109, 428-438.
- Huang, Lt al.., A review of models for near-field exposure pathways of chemicals in consumer products. Science of The Total Environment 2017, 574, 1182-1208.
- Huang, L.; Jolliet, O., A parsimonious model for the release of volatile organic compounds (VOCs) encapsulated in products. Atmospheric Environment 2016, 127, 223-235.
- Jolliet, O. et al. Defining Product Intake Fraction to Quantify and Compare Exposure to Consumer Products. Environmental Science & Technology 2015, 49, (15), 8924-8931.
- Kavlock, Robert J., et al. "Accelerating the pace of chemical risk assessment." Chemical research in toxicology 31.5 (2018): 287-290.
- MacLeod, Matthew, et al. "The state of multimedia mass-balance modeling in environmental science and decision-making." (2010): 8360-8364
- McEachran, Andrew D., Jon R. Sobus, and Antony J. Williams. "Identifying known unknowns using the US EPA's CompTox Chemistry Dashboard." *Analytical and bioanalytical chemistry* 409.7 (2017): 1729-1735.
- National Research Council. (1983). Risk Assessment in the Federal Government: Managing the Process Working Papers. National Academies Press.
- Obama, B. H. "Executive Order 13642: Making Open and Machine Readable the New Default for Government Information. Washington, DC: Office of the Executive." (2013).
- Park, Youngja, H., et al. "High-performance metabolic profiling of plasma from seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring." Toxicology 295:47-55 (2012)
- Pearce, Robert G., et al. "Httk: R package for high-throughput toxicokinetics." Journal of statistical software 79.4 (2017): 1.
- Rager, Julia E., et al. "Linking high resolution mass spectrometry data with exposure and toxicity forecasts to advance high-throughput environmental monitoring." *Environment international* 88 (2016): 269-280.
- Rappaport, Stephen M., et al. "The blood exposome and its role in discovering causes of disease." Environmental health perspectives 122.8 (2014): 769-774.
- Ring, Caroline L., et al. "Identifying populations sensitive to environmental chemicals by simulating toxicokinetic variability." Environment International 106 (2017): 105-118.
- Ring, Caroline L., et al. "Consensus Modeling of Median Chemical Intake for the US Population Based on Predictions of Exposure Pathways." Environmental science & technology 53.2 (2018): 719-732
- Shin, H.-M.; McKone, T. E.; Bennett, D. H., Intake Fraction for the Indoor Environment: A Tool for Prioritizing Indoor Chemical Sources. Environmental Science & Technology 2012, 46, (18), 10063-10072.
- 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.
- Sobus, Jon R., et al. "Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA." *Journal of exposure science & environmental epidemiology* (2017): 1.
- Tan, Yu-Mei, Kai H. Liao, and Harvey J. Clewell III. "Reverse dosimetry: interpreting trihalomethanes biomonitoring data using physiologically based pharmacokinetic modeling." Journal of Exposure Science and Environmental Epidemiology 17.7 (2007): 591.
- 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. "Integration of dosimetry, exposure and high-throughput screening data in chemical toxicity assessment." Toxicological Sciences (2012): kfr254.
- 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.
- Zhang, X.; Arnot, J. A.; Wania, F., Model for screening-level assessment of near-field human exposure to neutral organic chemicals released indoors. Environmental science & technology 2014, 48, (20), 12312-12319.