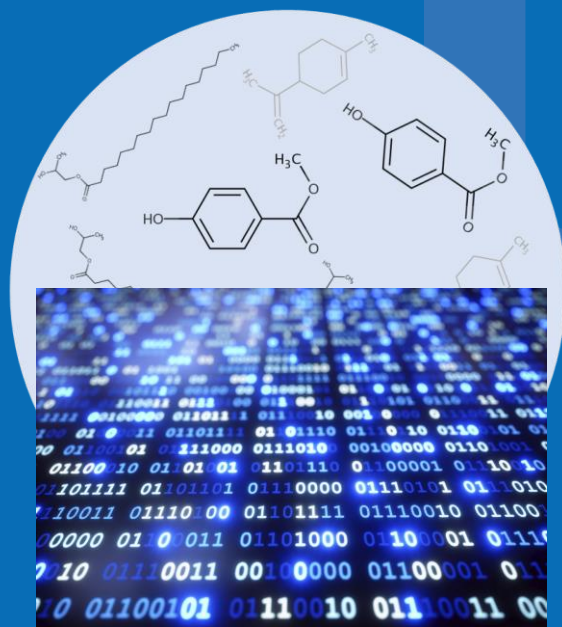


Filling Gaps in Exposure Data from Chemical Descriptors with Machine Learning

Kristin Isaacs

Center for Computational Toxicology and Exposure
United States Environmental Protection Agency

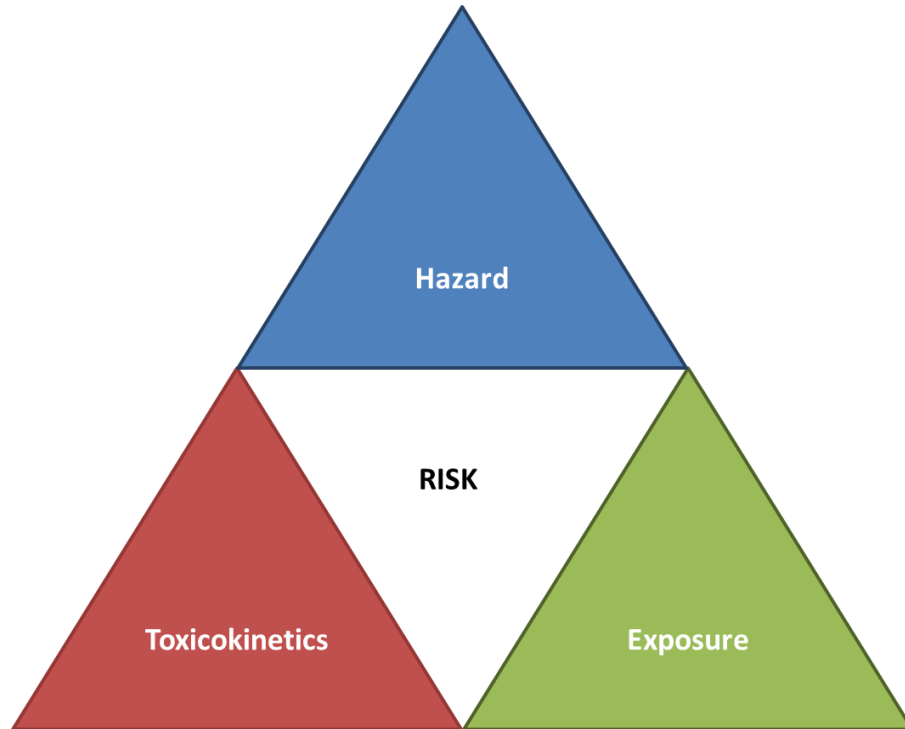


**New Approach Methodologies for Exposure:
Advancing Chemical Risk Assessment**
Society of Toxicology Annual Meeting
March 16, 2021



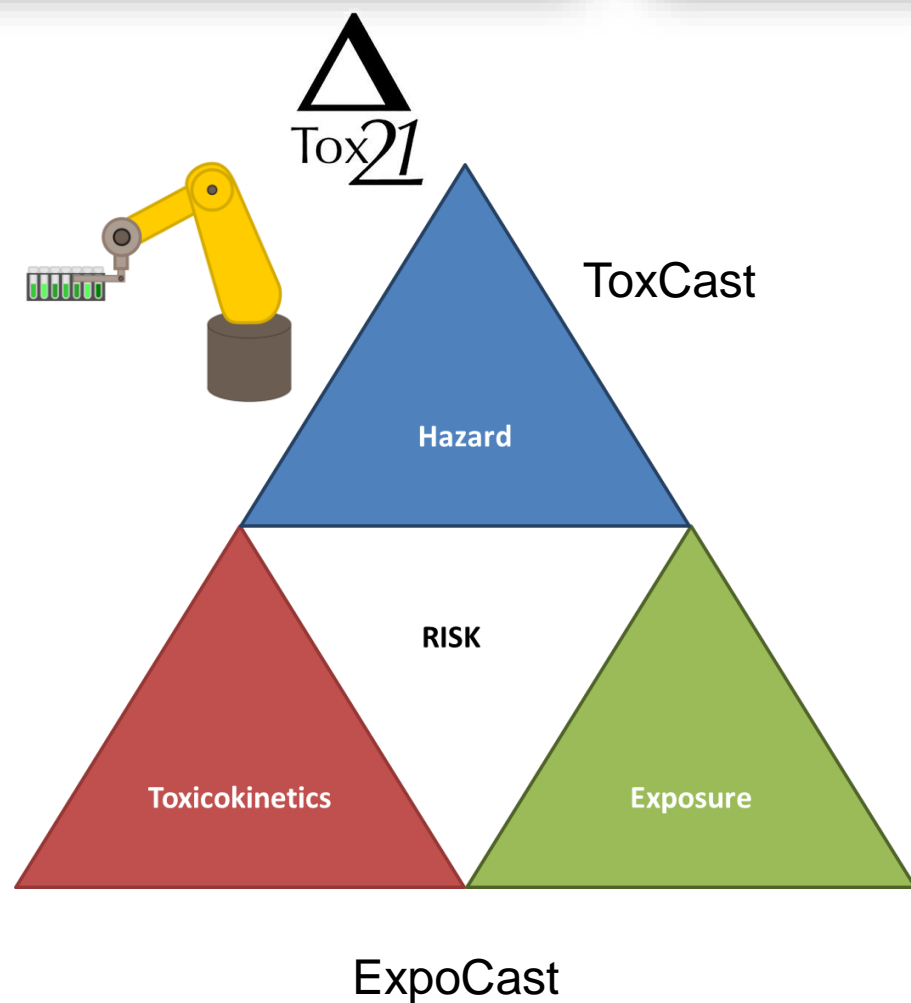
The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA





- Evaluating chemicals for risk to humans or the environment requires information on hazard and exposure potential
- Exposure potential quantifies the degree of contact between a chemical and a receptor
- Toxicokinetic information is required to bridge hazard and exposure (what real-world exposure is required to produce an internal concentration consistent with a potential hazard?)



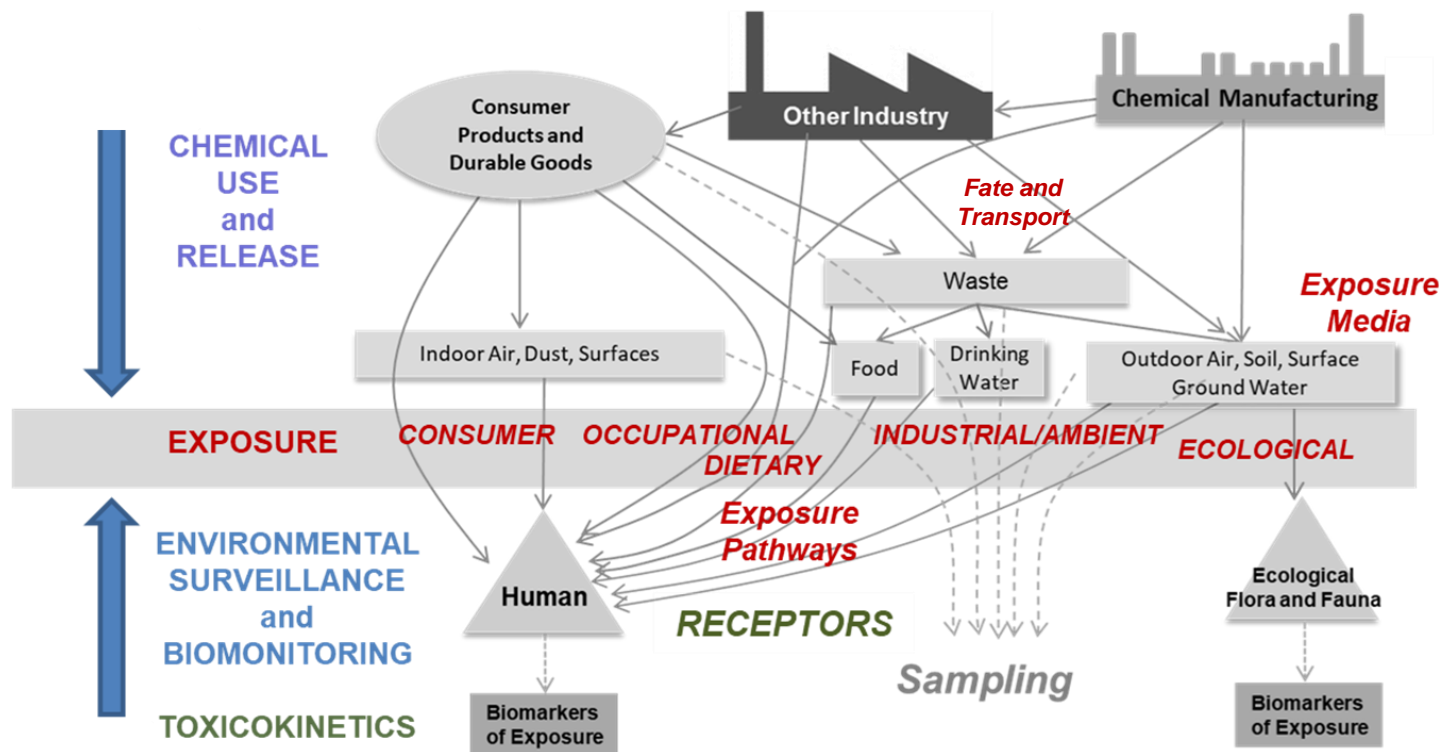


- Evaluating chemicals for risk to humans or the environment requires information on hazard and exposure potential
- Exposure potential quantifies the degree of contact between a chemical and a receptor
- Toxicokinetic information is required to bridge hazard and exposure (what real-world exposure is required to produce an internal concentration consistent with a potential hazard?)



Forward Models
Use/Release → Exposure
Often pathway-specific

Reverse Models
Biomarker Concentrations → Exposure
Useful for generating evaluation data



Traditional use, release, monitoring, and toxicokinetic data are still unavailable for 1000s of chemicals in commerce.



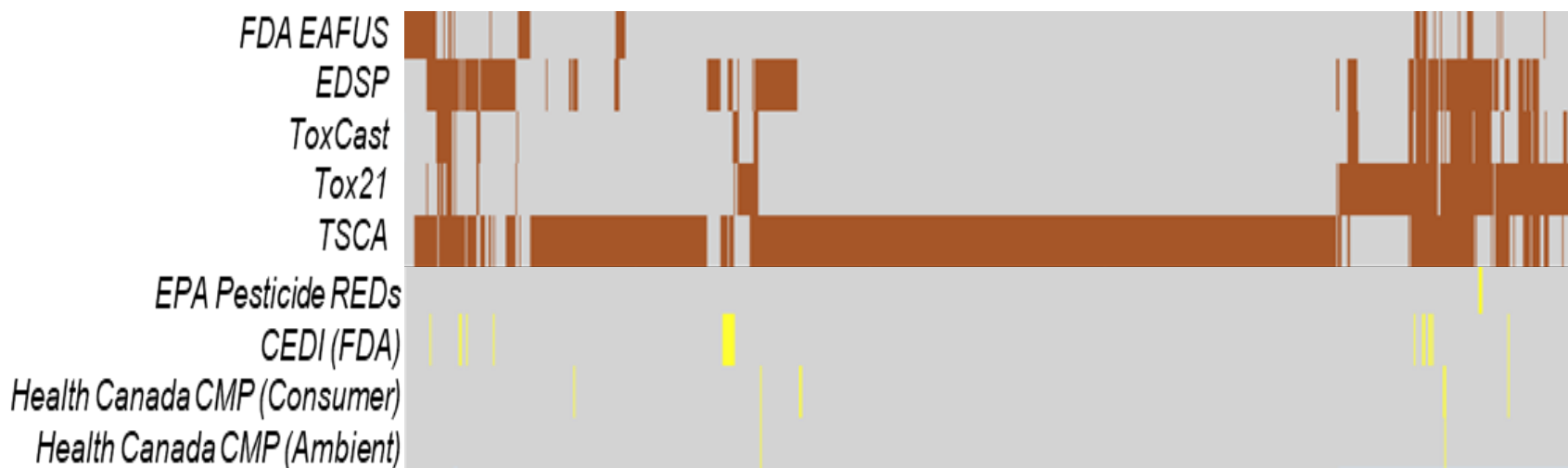


38,344 Inventory Chemicals

- Examined coverage of chemical inventories
- Regulatory lists
 - EPA Toxic Substance Control Act Non-Confidential Active Inventory
 - EPA Endocrine Disruptor Screening Program
 - FDA Everything Added to Food in the US (EAFUS)
- Chemicals tested in high-throughput screening
 - ToxCast
 - Tox21



How Have ML Models Improved the Exposure-Relevant Data Landscape?



38,344 Inventory Chemicals

- Examined coverage of chemical inventories
- Regulatory lists
 - EPA Toxic Substance Control Act Non-Confidential Active Inventory
 - EPA Endocrine Disruptor Screening Program
 - FDA Everything Added to Food in the US (EAFUS)
- Chemicals tested in high-throughput screening
 - ToxCast
 - Tox21



National Academies Workshop, June 2019

Proceedings of a Workshop IN BRIEF

August 2019

Leveraging Artificial Intelligence and Machine Learning to Advance
Environmental Health Research and Decisions

Proceedings of a Workshop—in Brief

- *“Machine learning algorithms can analyze large volumes of complex data to find patterns and make predictions, often exceeding the accuracy and efficiency of people who are attempting the same task.”*
- Highlighted several areas of environmental health for which AI and machine learning could play an integral role in research, including
 - Predicting the toxicology of chemicals
 - Characterizing the exposome



National Academies Workshop, June 2019

Proceedings of a Workshop IN BRIEF

August 2019

Leveraging Artificial Intelligence and Machine Learning to Advance
Environmental Health Research and Decisions

Proceedings of a Workshop—in Brief

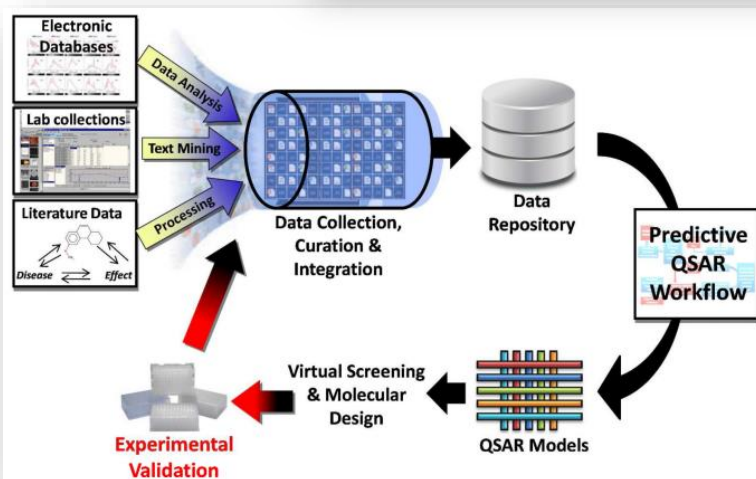


- *“Machine learning algorithms can analyze large volumes of complex data to find patterns and make predictions, often exceeding the accuracy and efficiency of people who are attempting the same task.”*
- Highlighted several areas of environmental health for which AI and machine learning could play an integral role in research, including
 - Predicting the toxicology of chemicals
 - Characterizing the exposome
- Defined eight classes of NAMs for exposure, including
 - **Chemical descriptors** that provide information on chemicals in an exposure context (e.g., how chemicals are used)
 - **Machine-learning approaches** that use these descriptors to fill gaps in existing data



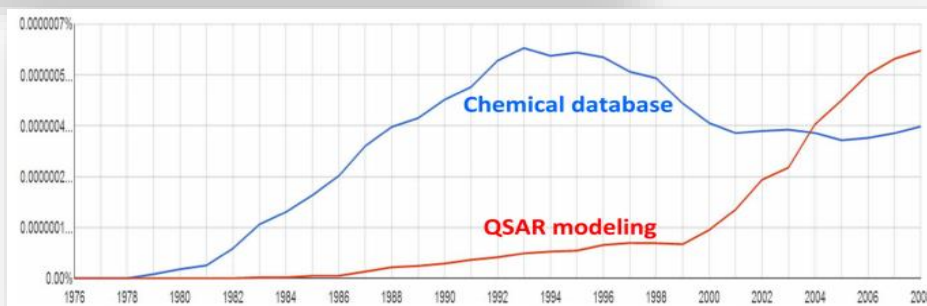
QSAR Modeling: Where Have You Been? Where Are You Going To?

Artem Cherkasov[†], Eugene N. Muratov^{‡§}, Denis Fourches[‡], Alexandre Varnek[‡], Igor I. Baskin[‡], Mark Cronin[‡], John Dearden[‡], Paola Gramatica[‡], Yvonne C. Martin[‡], Roberto Todeschini[‡], Viviana Consonni[‡], Victor E. Kuz'min[§], Richard Cramer[‡], Romualdo Benigni[‡], Chihai Yang[‡], James Rathman^{‡Δ}, Lothar Terfloth[‡], Johann Gasteiger[‡], Ann Richard[‡], and Alexander Tropsha^{‡‡}

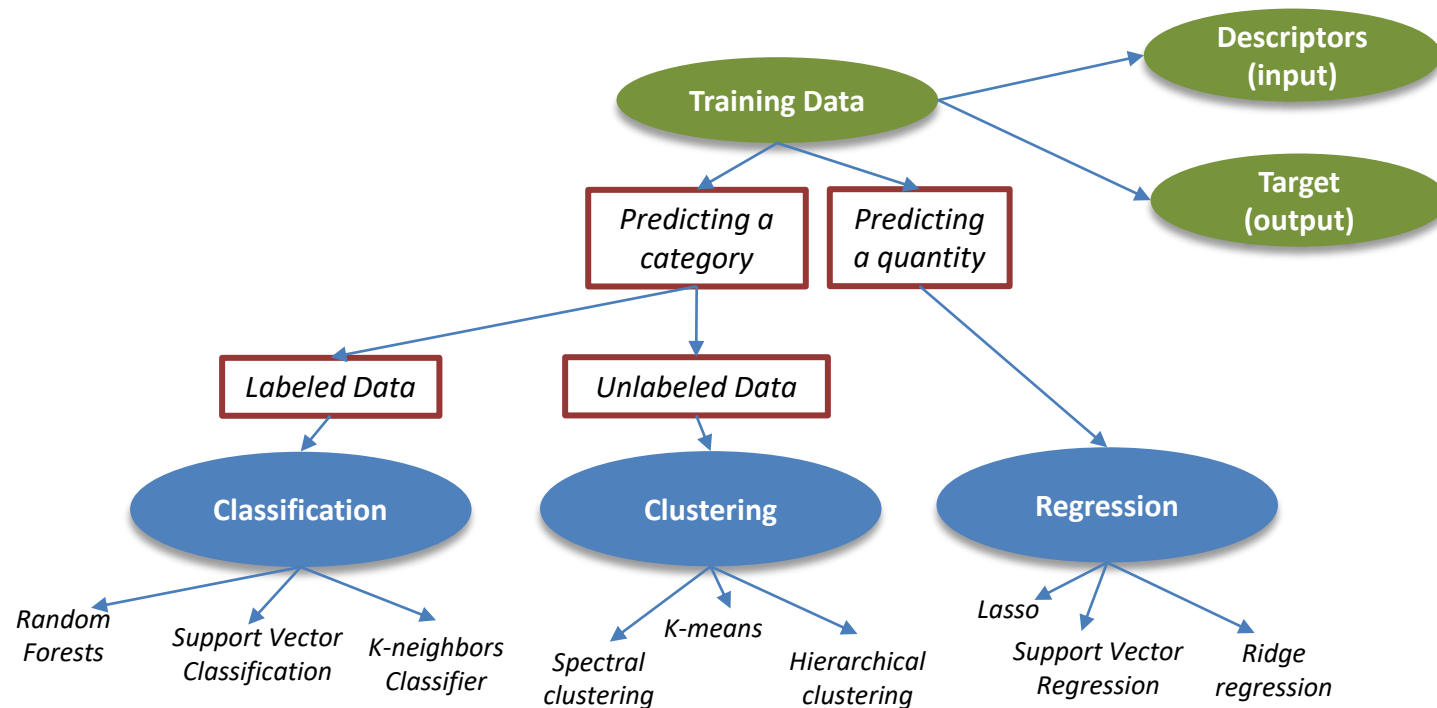


- Quantitative Structure-Activity Relationships (QSAR) models have been used for over 50 years to predict the physical and biological properties of chemicals.
- The field has advanced from simple regression methods to sophisticated machine learning techniques for the analysis of very large datasets comprising thousands of diverse molecular structures.
- The scientific QSAR community has been on the forefront of the use of machine learning methods, having developed:
 - New chemical structural descriptor sets
 - Many recommendations for best practices, including model validation

Mentions in Google Books Database



- Machine learning is ideally suited to look at many factors simultaneously.
- It can identify patterns in large datasets and build corresponding predictive models.
- A major challenge is determining the most appropriate method for the problem.



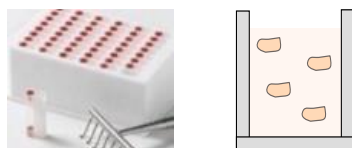
Portions adapted from <https://scikit-learn.org>



SCIENTIFIC DATA

OPEN Data Descriptor: The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products

Received: 16 October 2017
Accepted: 30 April 2018
Published: 10 July 2018
Kathie L. Dionisio¹, Katherine Phillips¹, Paul S. Price¹, Christopher M. Gulke², Anthony Williams², Derya Biryol^{1,3}, Tao Hong⁴ & Kristin K. Isaacs²



SCIENTIFIC DATA

OPEN Database of pharmacokinetic time-series data and parameters for 144 environmental chemicals

Risa R. Sayre^{1,2,3,4}, John F. Wambaugh⁵ & Christopher M. Gulke⁶

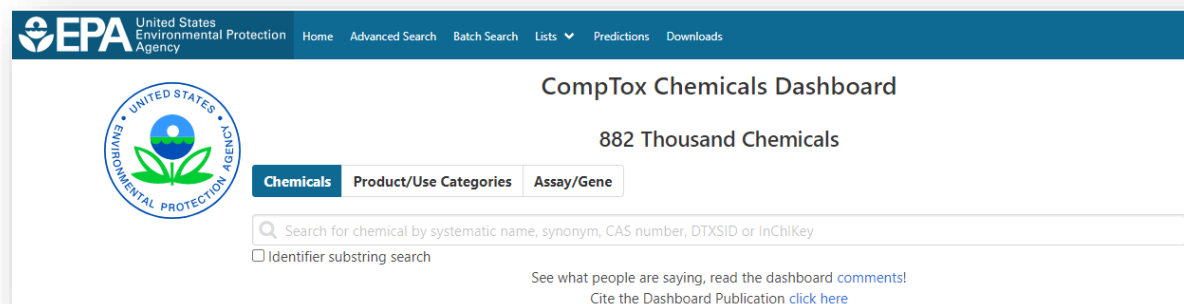
New quantitative and qualitative chemical use descriptors from EPA's Chemicals and Products Database (CPDat, Dionisio et al. 2018)

Traditional (targeted) monitoring data for various environmental media from publicly available monitoring databases

In-vitro protein binding and clearance (Wetmore et al. 2015, Pearce et al. 2017, Wambaugh et al. 2019a.)

In-vivo toxicokinetic parameters collected from the literature (Sayre et al. 2020)

- **Models are only as good as the underlying data!**
- In EPA-ORD's ExpoCast project, we are compiling the datasets that enable extrapolation of target information to data-poor chemicals using machine learning.
- Also currently developing IT infrastructure for automated and manual curation, QA, provenance tracking, and dissemination of these data (**Poster P-117**).
- Our goal is to be able to ultimately provide all these data publicly via the CompTox Chemicals Dashboard.
(<https://comptox.epa.gov/dashboard>)



Use Descriptors

**Environmental
Science & Technology**

High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals

John F. Wambaugh,^{*,†} Anran Wang,^{†,§,||} Kathie L. Richard Judson,[†] and R. Woodrow Setzer[†]

SCIENTIFIC DATA

OPEN Data Descriptor: The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products

Received: 16 October 2017
Accepted: 30 April 2018
Published: 10 July 2018

Kathie L. Dionisio¹, Katherine Phillips¹, Paul S. Price¹, Christopher M. Grulke¹, Antony Williams², Derya Biryol^{1,3}, Tao Hong⁴ & Kristin K. Isaacs¹

Property Descriptors

OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri^{1,2,3*}, Chris M. Grulke¹, Richard S. Judson¹ and Antony J. Williams¹

<https://github.com/kmansouri/OPERA>

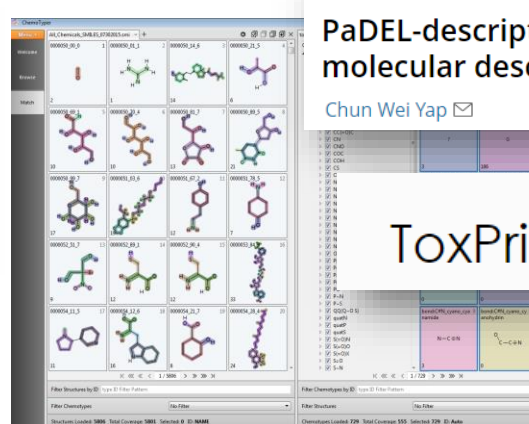
- We use a variety of chemical descriptor sets for our exposure models.
- Different descriptor sets contain unique information that can inform predictive models.
- OPERA and ToxPrint descriptors can be easily downloaded for thousands of substances using the batch search utility of the CompTox Chemicals Dashboard.

Structural/Molecular Descriptors

PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints

Chun Wei Yap

ToxPrint - A Public Set of Chemotypes



- Ensemble average over many decision trees
- Randomly select subset of descriptors and grow 'unpruned' tree, repeat many times
- Model returns a probability equal to fraction of trees returning a positive classification
- Importance of descriptors can be quantified

Package 'randomForest'

October 7, 2015

Title Breiman and Cutler's Random Forests for Classification and Regression

Version 4.6-12

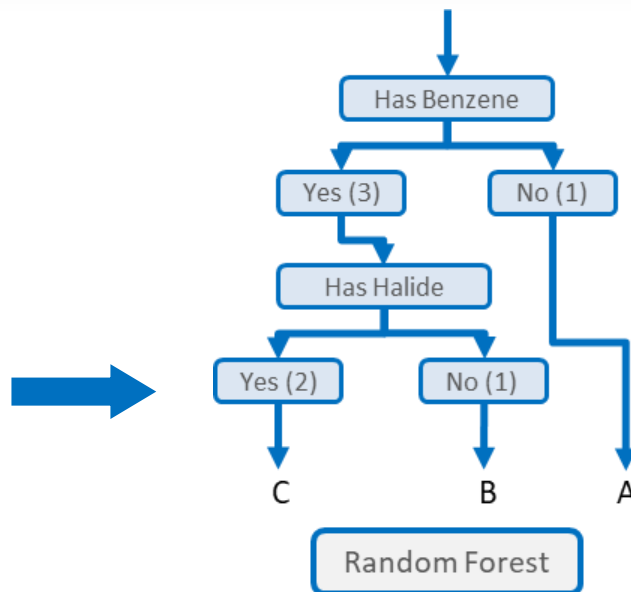
Date 2015-10-06



Train Model

		Has Metal	Has Halide	Has Benzene	Has Alkyne
Chemical 1	A	0	1	0	0
Chemical 2	B	0	0	1	1
Chemical 3	C	0	1	1	0
Chemical 4	C	1	1	1	1

Target Descriptors



Validate models

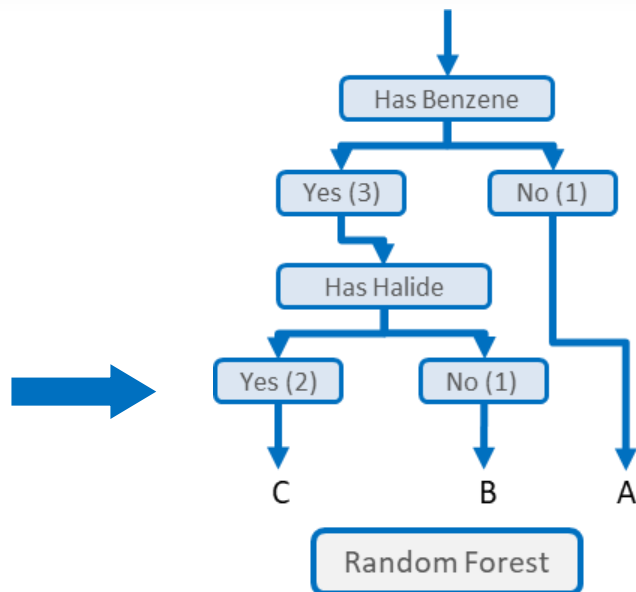
- Does model work beyond the training set?
- Does the model perform better than one built using random data?



Train Model

		Has Metal	Has Halide	Has Benzene	Has Alkyne
Chemical 1	A	0	1	0	0
Chemical 2	B	0	0	1	1
Chemical 3	C	0	1	1	0
Chemical 4	C	1	1	1	1

Target Descriptors



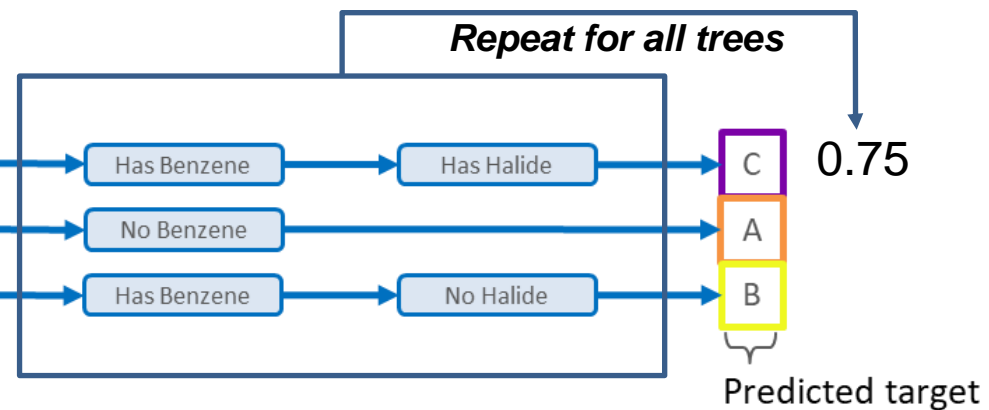
Validate models

- Does model work beyond the training set?
- Does the model perform better than one built using random data?

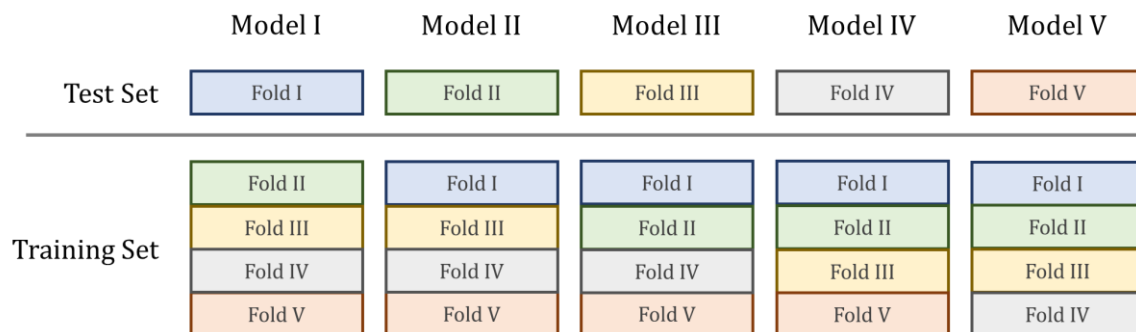
Predict Model

Chemical 5	0	1	1	0
Chemical 6	0	1	0	0
Chemical 7	0	0	1	1

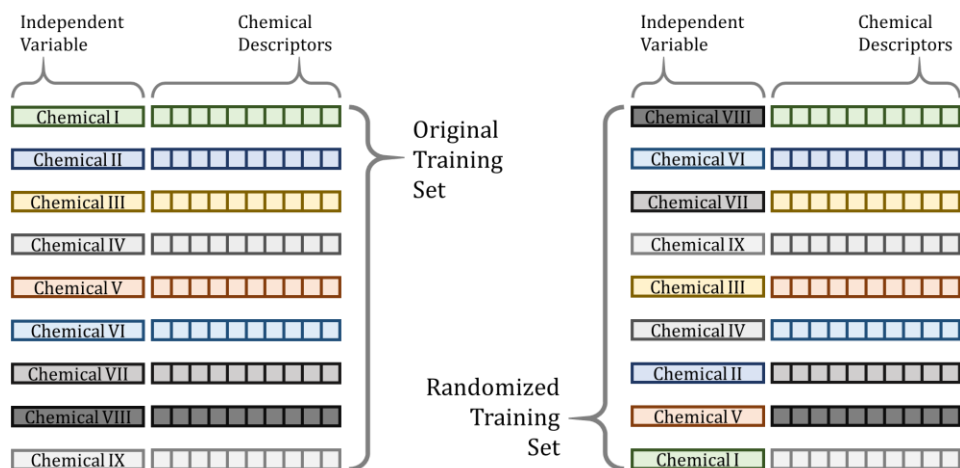
Descriptors



5-fold cross validation



Y-randomization

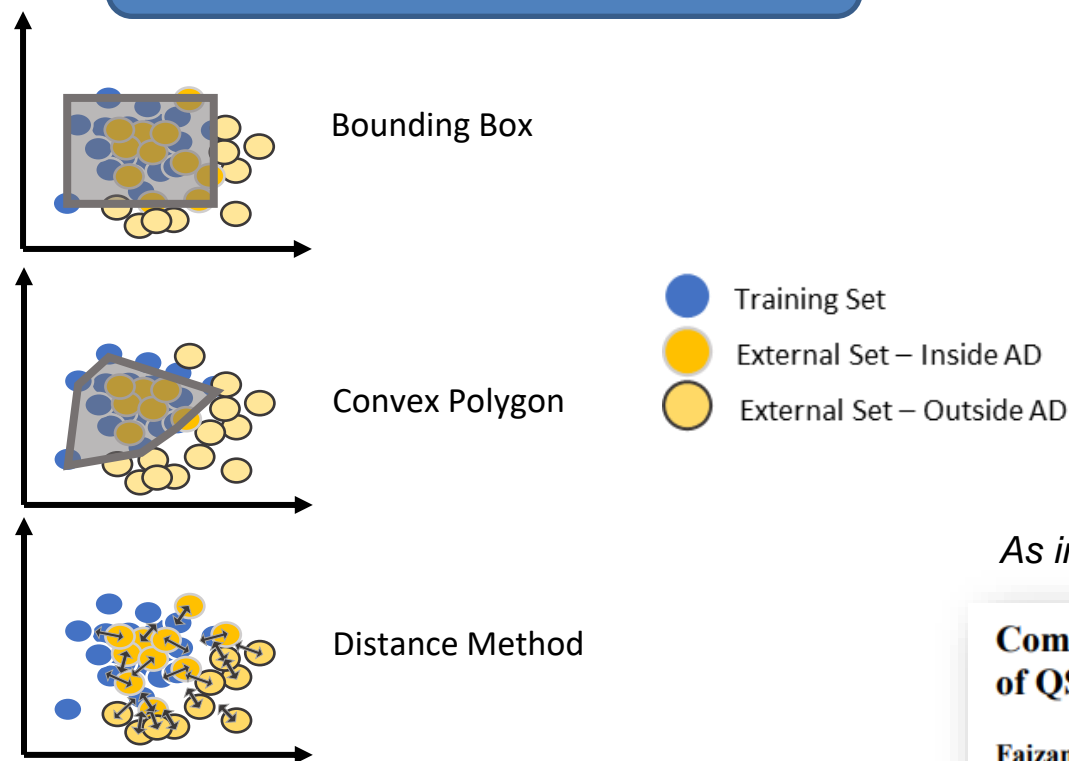


- Validation approaches:
 - 5-fold cross validation (build the model 5 times withholding a different subset of the data each time for testing)
 - Y-randomization (build the model using randomized target assignment to descriptors - does the true model outperform the randomized version?)
 - Evaluation with true external training sets



AD: The response and chemical structure space in which the model makes predictions with a given reliability

Methods for Assessing AD in Chemical Space



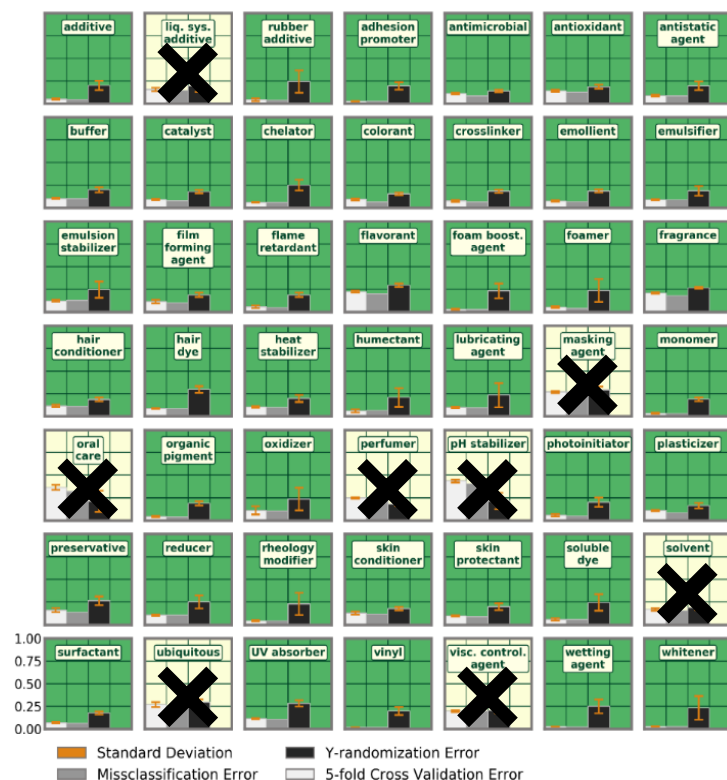
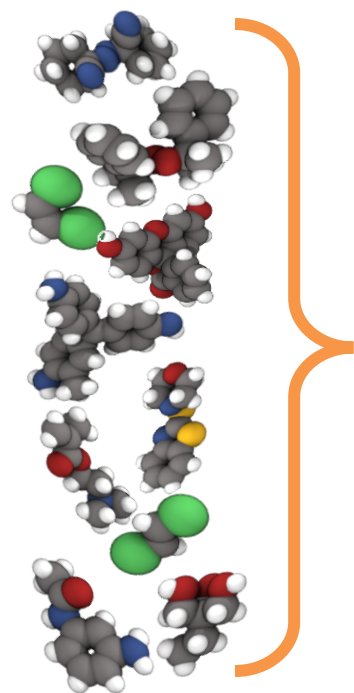
- QSAR/Machine learning best-practices include an emphasis on model validation and the need to define **model applicability domain** (AD) in the chemistry space (Tropsha et al. 2007)
- Knowledge of the AD is required for assessing confidence in predictions for new chemicals and quantifying the utility of additional data.

As in Sahigara et al., Molecules (2012):

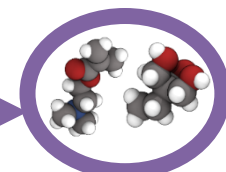
Comparison of Different Approaches to Define the Applicability Domain of QSAR Models

Faizan Sahigara, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni and Roberto Todeschini *

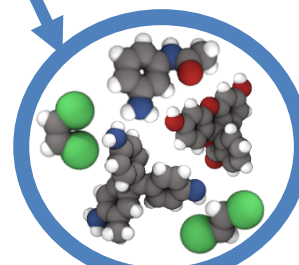




Catalysts



Crosslinkers



Plasticizer

- Quantitative Structure-Use Relationships (QSURs)
- Built from a training set of over 15,000 reported functions
- Used to inform screening of chemical libraries for potential alternatives with lower toxicity
- Used in non-targeted analysis workflows for ground truthing tentatively identified chemicals

Green Chemistry

PAPER

[View Article Online](#)
[View Journal](#) | [View Issue](#)



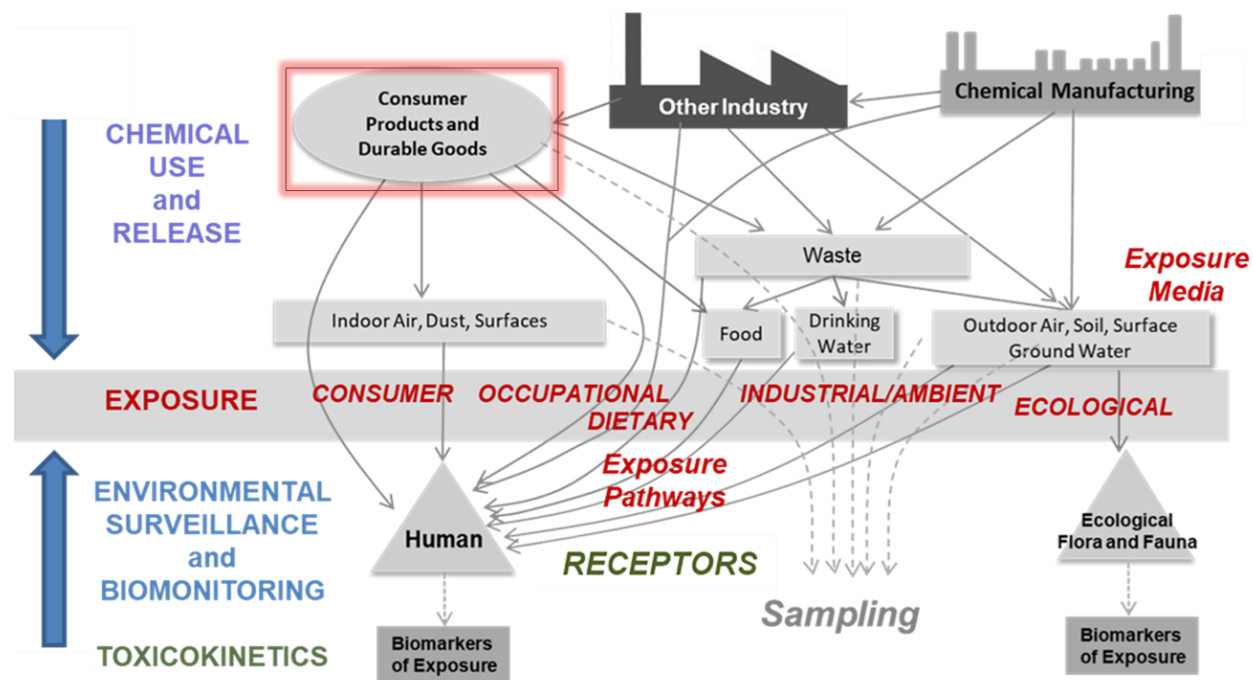
Cite this: *Green Chem.*, 2017, **19**, 1063

High-throughput screening of chemicals as functional substitutes using structure-based classification models†

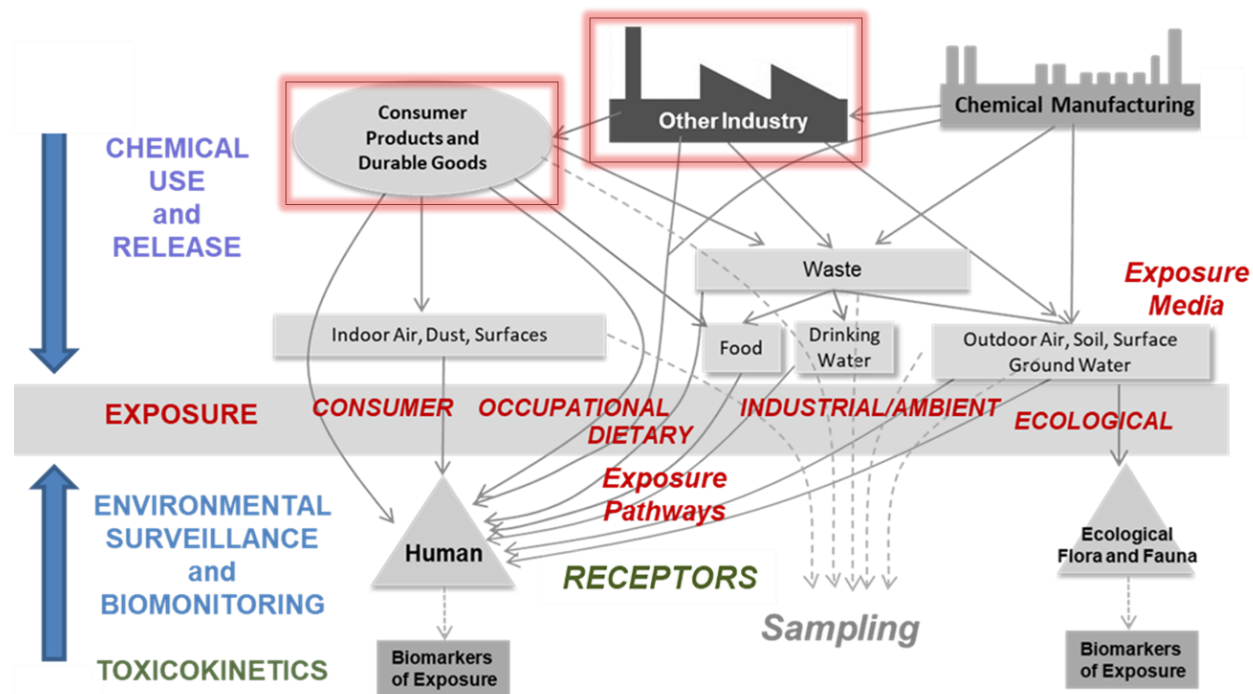
Katherine A. Phillips,^{a,c} John F. Wambaugh,^b Christopher M. Grulke,^b Kathie L. Dionisio^c and Kristin K. Isaacs^c



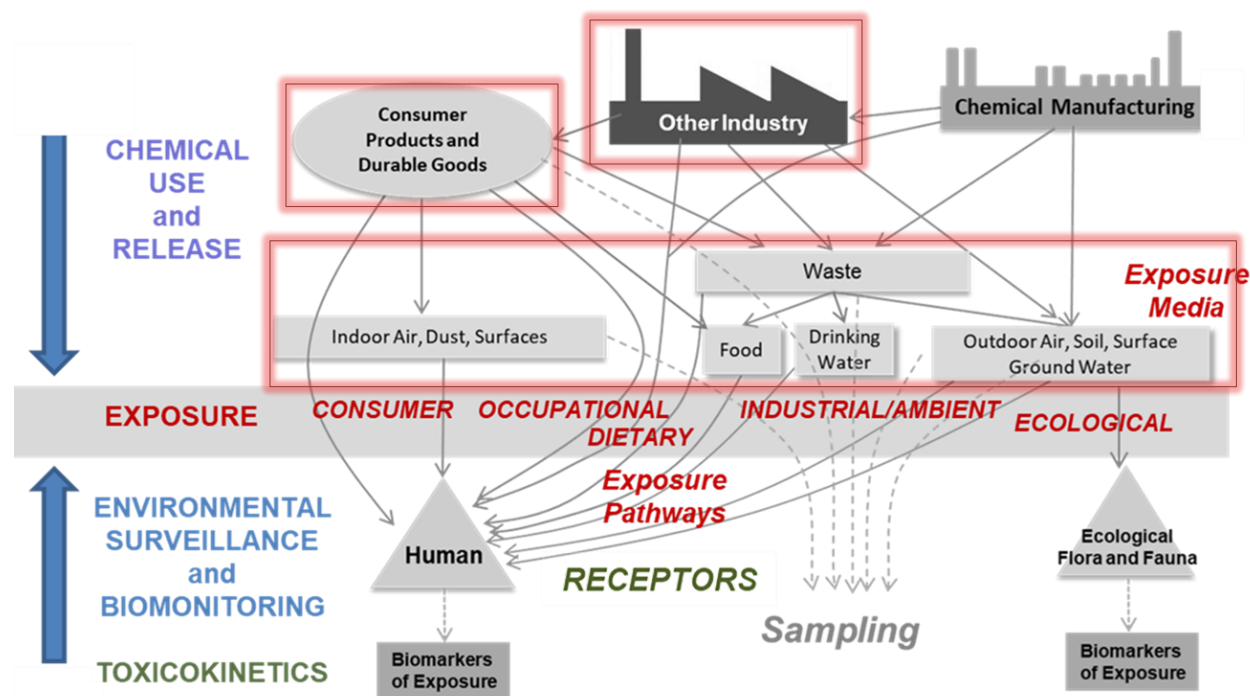
- Using functional use predictions to estimate quantitative chemical weight fractions in consumer formulations (Isaacs et al., 2016) and articles
- Using Natural Language Processing Support Vector Classification models to assign 100,000 consumer product ingredient documents to harmonized categories for modeling



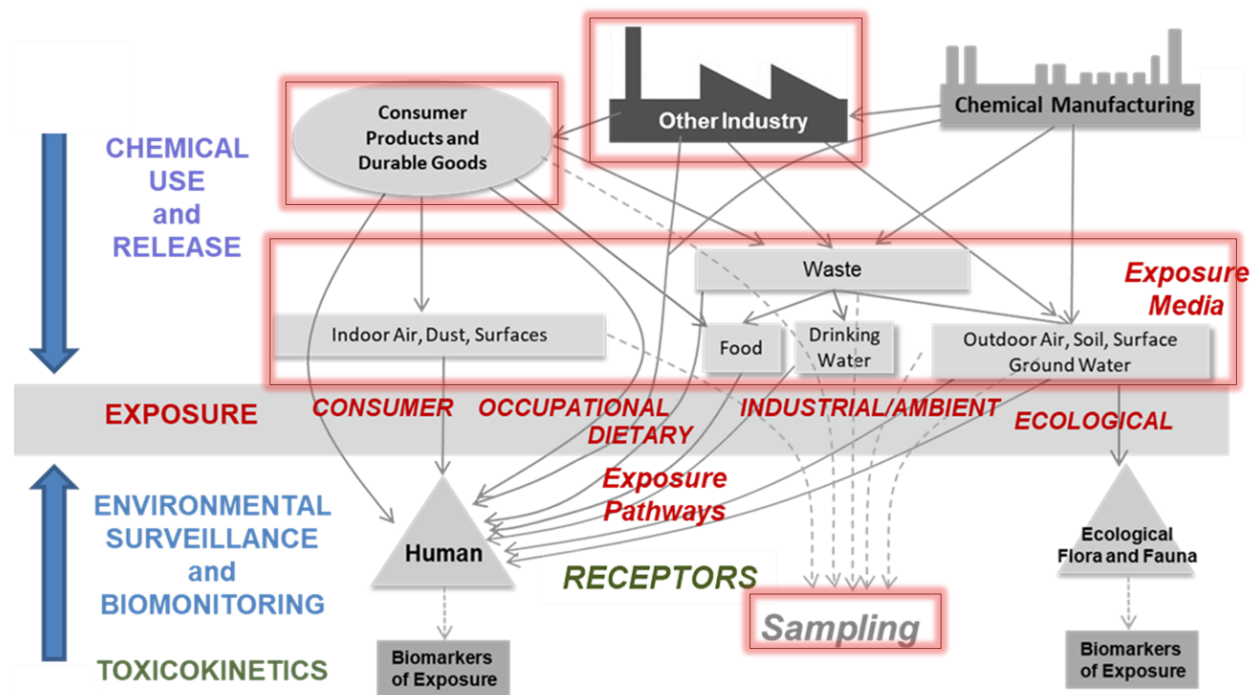
- Using functional use predictions to estimate quantitative chemical weight fractions in consumer formulations (Isaacs et al., 2016) and articles
- Using Natural Language Processing Support Vector Classification models to assign 100,000 consumer product ingredient documents to harmonized categories for modeling
- Prediction of chemical releases associated with industrial scenarios and processes (ORD Center for Environmental Solutions & Emergency Response)



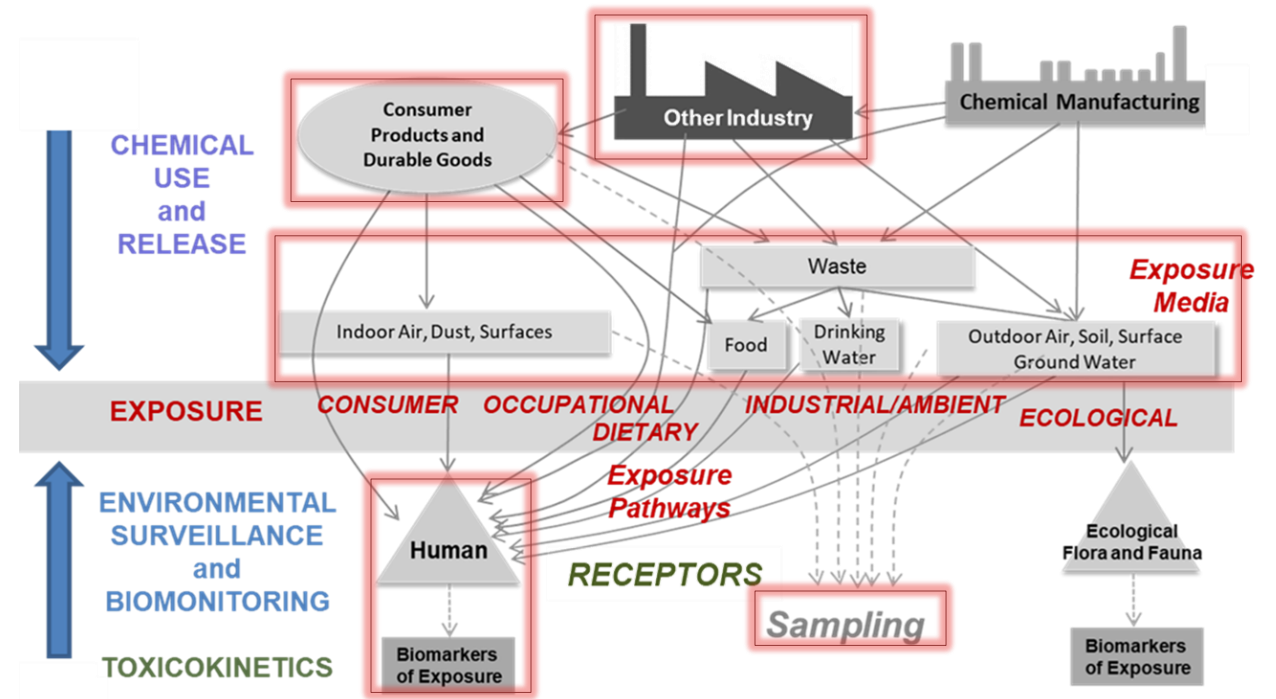
- Using functional use predictions to estimate quantitative chemical weight fractions in consumer formulations (Isaacs et al., 2016) and articles
- Using Natural Language Processing Support Vector Classification models to assign 100,000 consumer product ingredient documents to harmonized categories for modeling
- Prediction of chemical releases associated with industrial scenarios and processes (ORD Center for Environmental Solutions & Emergency Response)
- Prediction of chemical occurrence in 26 different types of environmental and biological media.



- Using functional use predictions to estimate quantitative chemical weight fractions in consumer formulations (Isaacs et al., 2016) and articles
- Using Natural Language Processing Support Vector Classification models to assign 100,000 consumer product ingredient documents to harmonized categories for modeling
- Prediction of chemical releases associated with industrial scenarios and processes (ORD Center for Environmental Solutions & Emergency Response)
- Prediction of chemical occurrence in 26 different types of environmental and biological media.
- Prediction of method amenability in high-resolution mass spectrometry (ORD Center for Computational Toxicology and Exposure)



- Using functional use predictions to estimate quantitative chemical weight fractions in consumer formulations (Isaacs et al., 2016) and articles
- Using Natural Language Processing Support Vector Classification models to assign 100,000 consumer product ingredient documents to harmonized categories for modeling
- Prediction of chemical releases associated with industrial scenarios and processes (ORD Center for Environmental Solutions & Emergency Response)
- Prediction of chemical occurrence in 26 different types of environmental and biological media.
- Prediction of method amenability in high-resolution mass spectrometry (ORD Center for Computational Toxicology and Exposure)
- In-silico machine learning models for toxicokinetic parameters: protein binding and clearance for environmental chemicals



Systematic Empirical Evaluation of Models

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

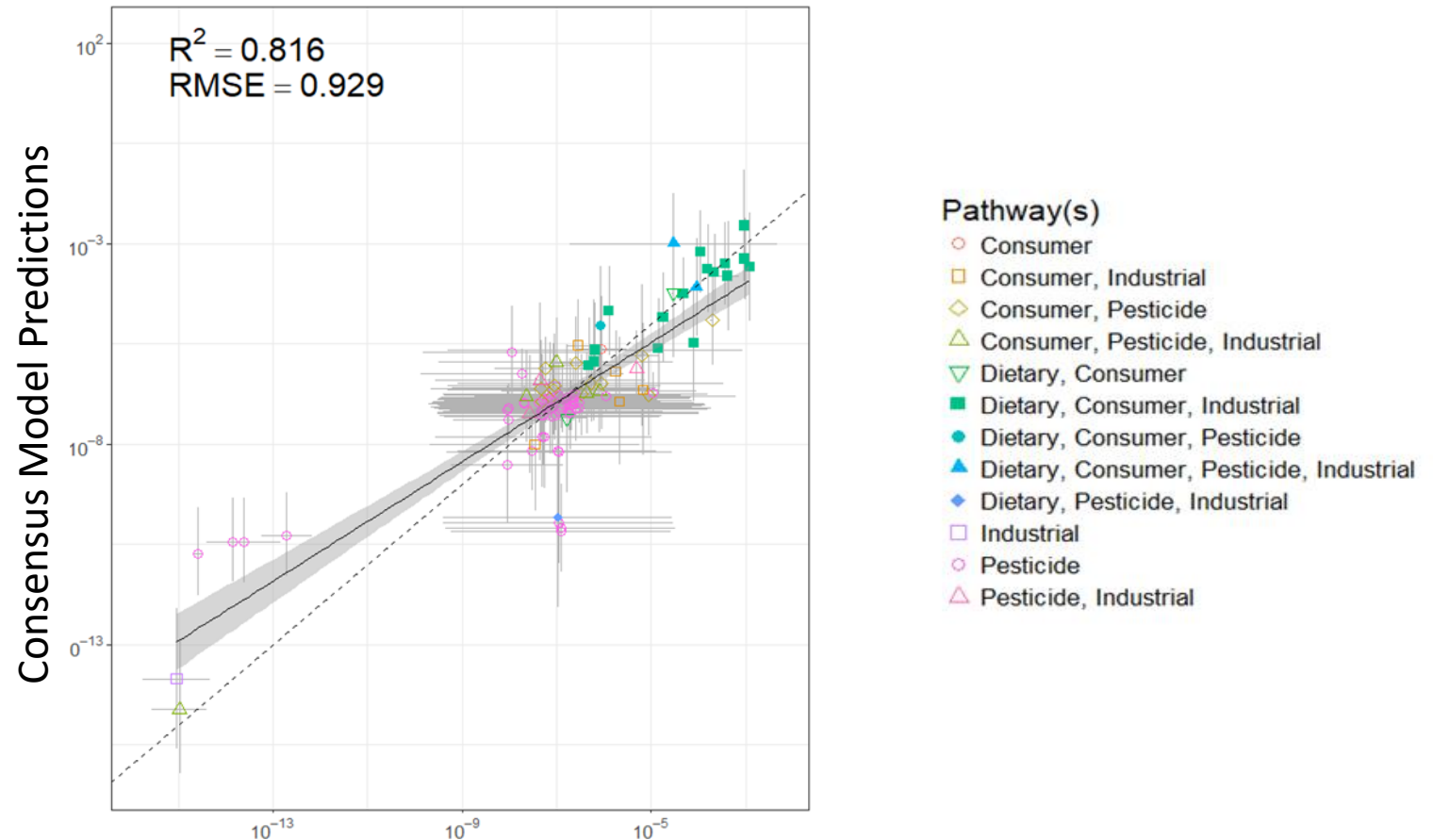


Predictor (including Models)	Reference(s)	Chemicals	Pathways
EPA Inventory Update Reporting and Chemical Data Reporting (CDR) (2015)	US EPA (2018)	7856	All
Stockholm Convention of Banned Persistent Organic Pollutants (2017)	Lallas (2001)	248	Far-Field Industrial and Pesticide
EPA Pesticide Reregistration 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), Ernststoff et al. (2017)	8167	Dietary



Ring et al., 2019

- Consumer (Near-Field), Industrial, Pesticide, Dietary
- Each chemical may have exposure by multiple pathways
- **Machine learning models were built for each of four exposure pathways**



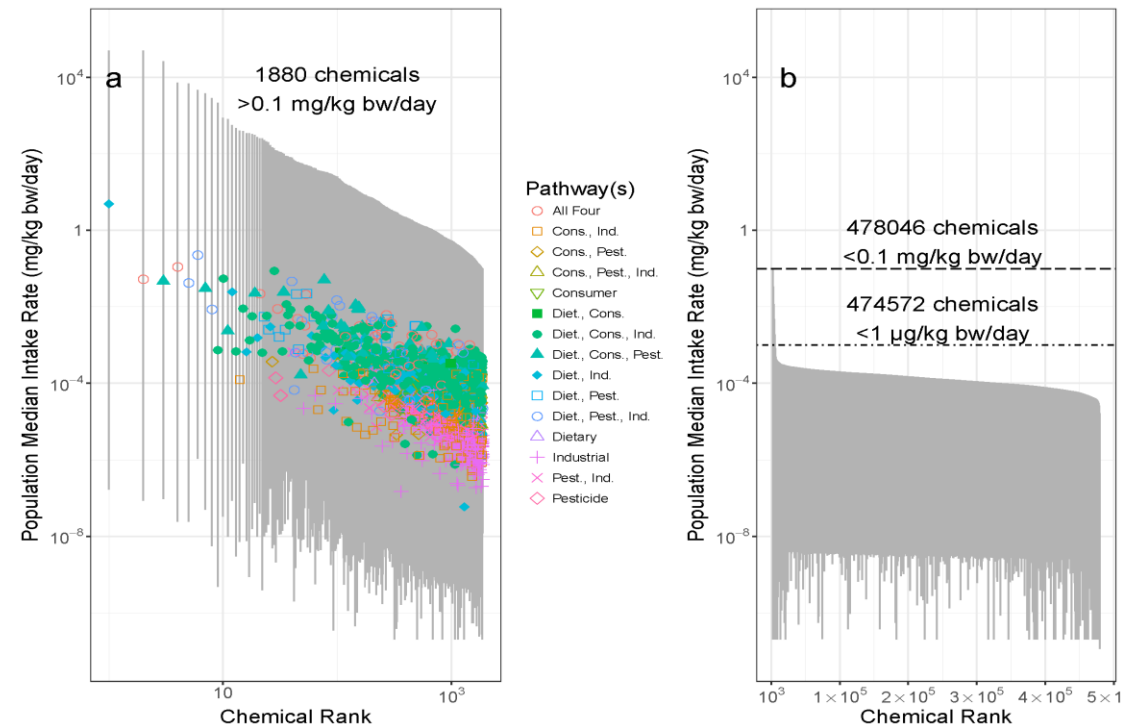
We use the method of Random Forests to relate chemical structure and properties to exposure pathway

	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



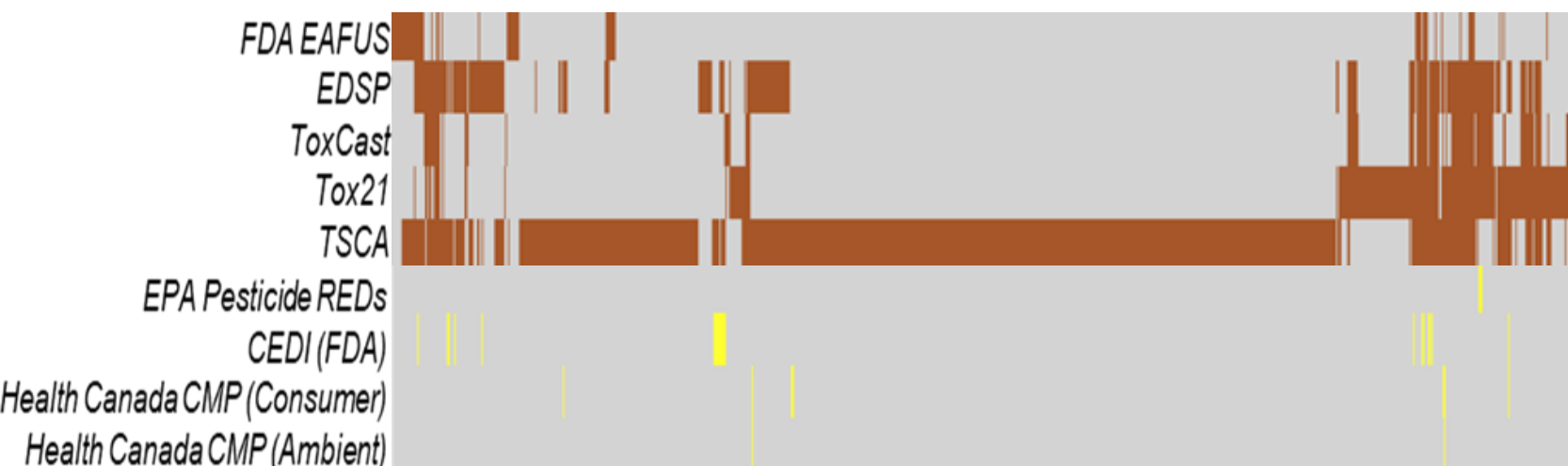
Ring et al., 2019

- 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 any of the four pathways and are considered outside the applicability domain.

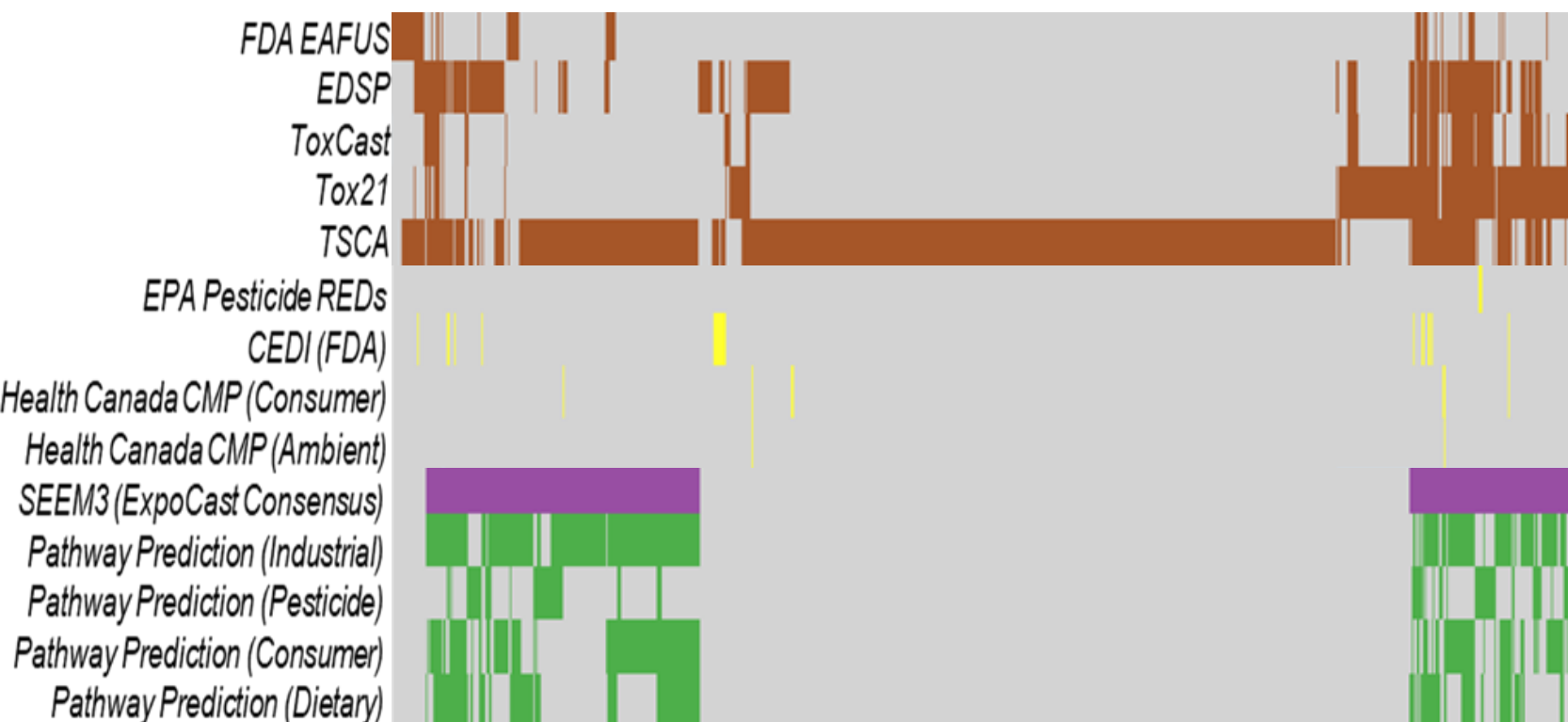




38,344 Inventory Chemicals

- Examined coverage of chemical inventories
- Regulatory lists
 - EPA Toxic Substance Control Act Non-Confidential Active Inventory
 - EPA Endocrine Disruptor Screening Program
 - FDA Everything Added to Food in the US (EAFUS)
- Chemicals tested in high-throughput screening
 - ToxCast
 - Tox21



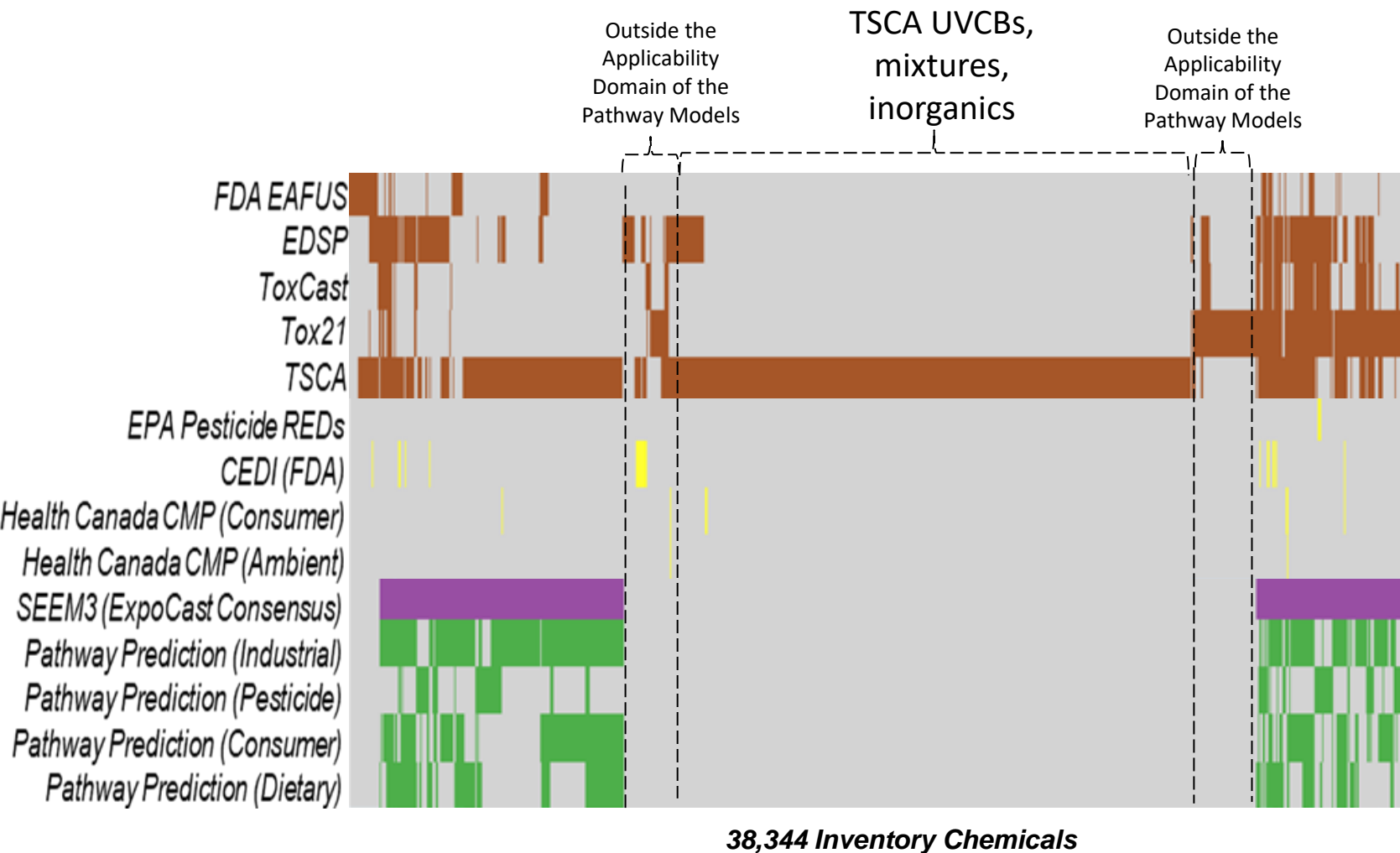


38,344 Inventory Chemicals

- Examined coverage of chemical inventories
- Regulatory lists
 - EPA Toxic Substance Control Act Non-Confidential Active Inventory
 - EPA Endocrine Disruptor Screening Program
 - FDA Everything Added to Food in the US (EAFUS)
- Chemicals tested in high-throughput screening
 - ToxCast
 - Tox21



How Have ML Models Improved the Exposure-Relevant Data Landscape?



- Examined coverage of chemical inventories
- Regulatory lists
 - EPA Toxic Substance Control Act Non-Confidential Active Inventory
 - EPA Endocrine Disruptor Screening Program
 - FDA Everything Added to Food in the US (EAFUS)
- Chemicals tested in high-throughput screening
 - ToxCast
 - Tox21

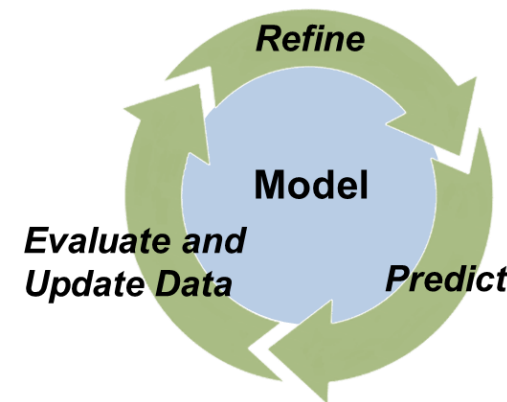


- **Challenges:**

- Transparency and reproducibility
- Determination of fit-for-purpose: How is suitability assessed? What criteria?

- **Strategies:**

- Learning from QSAR: Development of documentation and reporting standards
 - Training data, modeling methods, AD, results (predictions), performance metrics
 - Data accessibility
 - Versioning
 - Iterative development frameworks
- Integration into tiered workflow case studies (demonstration of value added when no other data are available)
- Continued external validation (with datasets of regulatory relevance)
- Characterization of uncertainty



- Machine learning is a powerful tool for extrapolating existing information to chemicals lacking data.
- We are building the training sets and machine-learning based predictive tools to estimate exposure-relevant information from chemical descriptors.
- We aim to develop workflows that allow for validation of model performance, characterization of chemical domain of applicability, and incorporation of new information as data become available.
- These new approach methodologies are improving our coverage of key chemical inventories.
- The predictions from these models provide defensible methods for filling knowledge gaps in process-based models, analytical workflows, chemical prioritization, and other risk-based evaluations.



ExpoCast Project (Exposure Forecasting)

Collaborators

CCTE

Linda Adams
Miyuki Breen*
Alex Chao*
Dan Dawson*
Mike Devito
Kathie Dionisio
Christopher Ecklund
Marina Evans
Peter Egeghy
Michael-Rock Goldsmith
Chris Grulke
Mike Hughes
Kristin Isaacs
Richard Judson
Jen Korol-Bexell*
Anna Kreutz*
Charles Lowe*
Seth Newton

Katherine Phillips
Paul Price
Tom Purucker
Ann Richard
Caroline Ring
Marci Smeltz*
Jon Sobus
Risa Sayre*
Mark Sfeir*
Mark Strynar
Zach Stanfield*
Rusty Thomas
Mike Tornero-Velez
Elin Ulrich
Dan Vallero
John Wambaugh
Barbara Wetmore
Antony Williams

CEMM

Xiaoyu Liu

CPHEA

Jane Ellen Simmons

CESER

David Meyer
Gerardo Ruiz-Mercado
Wes Ingwersen

*Trainees

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
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
Miyoun 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

