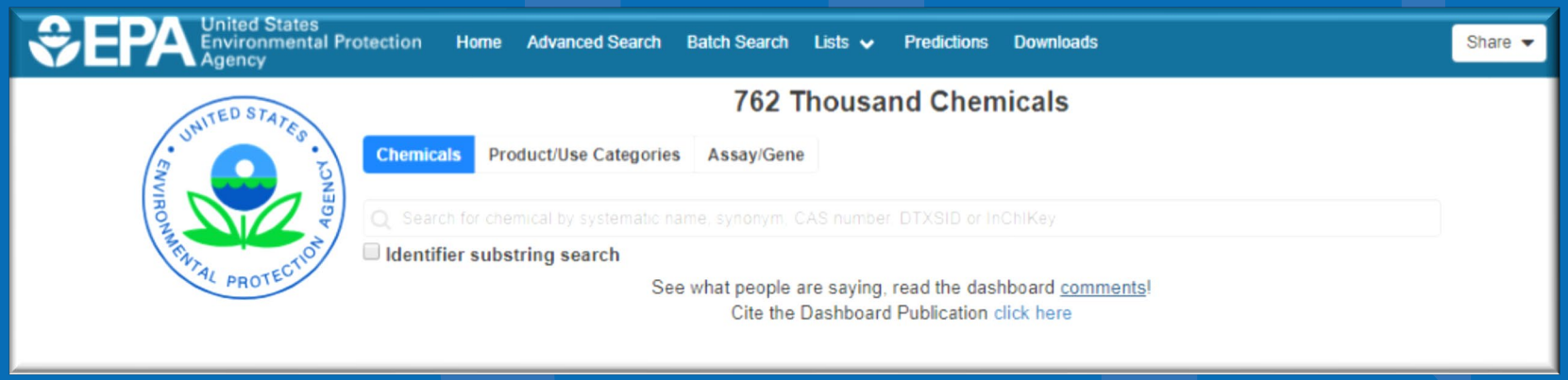
Public Exposure Information on the CompTox Dashboard

John Wambaugh[1], Kristin Isaacs[2], Katherine Phillips[2], Antony Williams[1]

1. National Center for Computational Toxicology, Office of Research and Development, U.S. EPA
2. Computational Exposure Division, National Exposure Research Laboratory, Office of Research and Development, U.S. EPA



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

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

International Society of Exposure Science
Pre-Conference Course: Application of New Approach Methodologies for Exposure Assessment and Prioritization – Tools for Researchers and Regulators Including use of Quantitative Structure Use Relationships (QSUR)
Ottawa, Canada
August 25, 2018

Chemical Regulation in the United States

- Park *et al.* (2012): At least 3221 chemicals present in pooled human blood samples, many appear to be exogenous albeit at low levels
- 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)

November 29, 2014

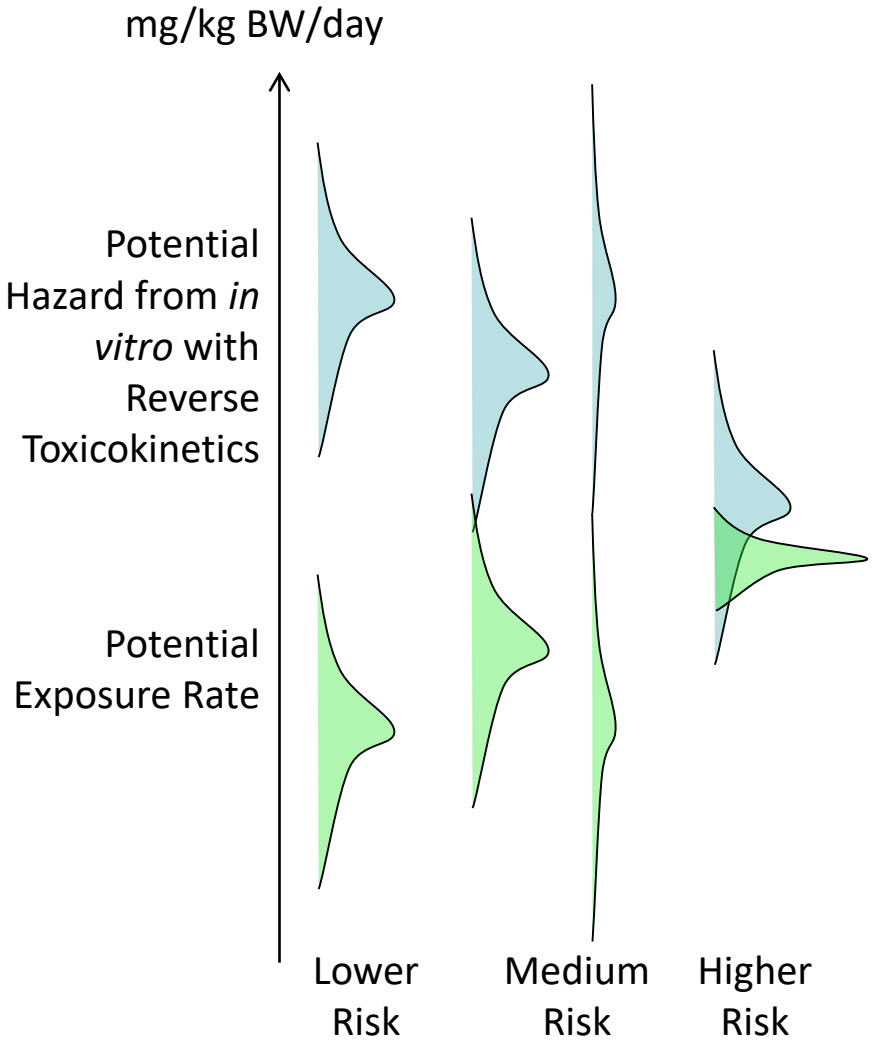
Chemical Regulation in the United States

- Most other chemicals, ranging from industrial waste to dyes to packing materials, are covered by the Toxic Substances Control Act (TSCA)
  - Thousands of chemicals on the market were either "grandfathered" in or were allowed without experimental assessment of hazard, toxicokinetics, or exposure
  - Thousands of new chemical use submissions are made to the EPA every year
- TSCA was updated in June, 2016 to allow evaluation of these and other chemicals
  - New alternative methodologies (NAMs) are being developed to prioritize these existing and new chemicals for testing

November 29, 2014

Chemical Risk = Hazard x Exposure

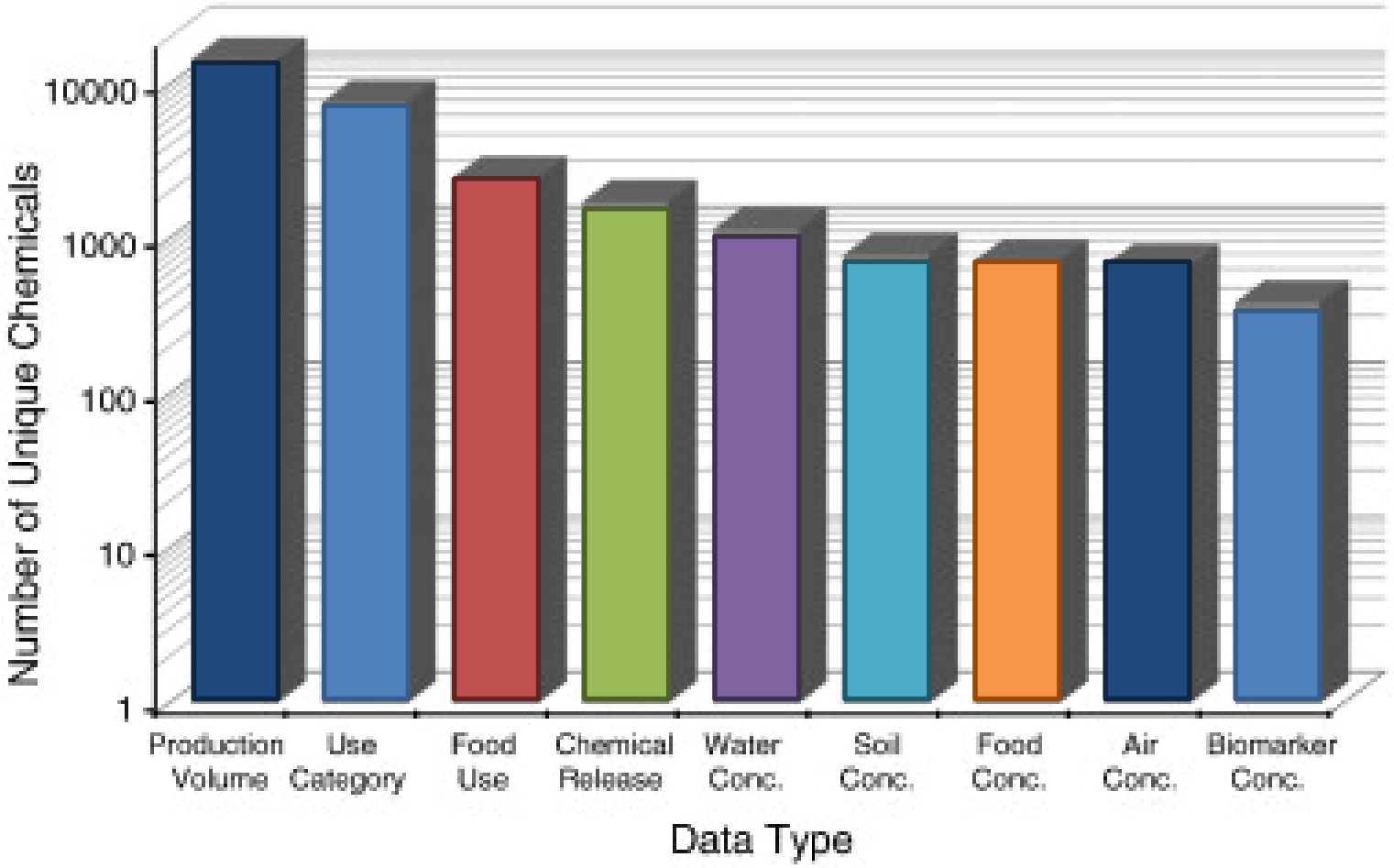
- National Research Council (1983) identified chemical risk as a function of both inherent hazard and exposure
- To address thousands of chemicals, we need new approach methodologies that can prioritize those chemicals most worthy of additional study
- **High throughput risk prioritization** needs:
  1. high throughput **hazard** characterization (Dix et al., 2007, Collins et al., 2008)
  2. high throughput **exposure** forecasts (Wambaugh et al., 2013, 2014)
  3. high throughput **toxicokinetics** (*i.e.*, dose-response relationship) linking hazard and exposure



Rotroff et al. (2010)
Wetmore et al. (2012, 2014, 2015)

Limited Available Data for Exposure Estimation

Most chemicals lack public exposure-related data beyond production volume (Egeghy et al., 2012)



Can we develop new tools to generate the exposure information we need?

Understanding Exposure to Chemicals

Rudel et al., 2003 Environmental Science & Technology

Research

Phthalates, Alkylphenols, Pesticides, Polybrominated Diphenyl Ethers, and Other Endocrine-Disrupting Compounds in Indoor Air and Dust

RUTHANN A. RUDEL,*,† DAVID E. CAMANN,‡ JOHN D. SPENGLER,§ LEO R. KORN,∥ AND JULIA G. BRODY†

Silent Spring Institute, 29 Crafts Street, Newton, Massachusetts 02458, Southwest Research Institute, 6220 Culebra Road, P.O. Box 28510, San Antonio, Texas 78228-0510, Environmental Science and Engineering Program, Harvard University School of Public Health, Landmark Center, 401 Park Drive, Boston, Massachusetts 02115, and Division of Biometrics, University of Medicine and Dentistry of New Jersey, School of Public Health, 335 George Street, Liberty Plaza, Suite 2200, New Brunswick, New Jersey 08903-2688

Chemicals identified as endocrine-disrupting compounds (EDCs) have widespread consumer uses, yet little is known about indoor exposure. We sampled indoor air and dust in 120 homes, analyzing for 89 organic chemicals identified as EDCs. Fifty-two compounds were detected in air and 66 were detected in dust. These are the first reported measures in residential environments for over 30 of the compounds, including several detected at the highest concentrations. The number of compounds detected per home ranged from 13 to 28 in air and from 6 to 42 in dust. The most abundant compounds in air included phthalates (plasticizers, emulsifiers), *o*-phenylphenol (disinfectant), 4-nonylphenol (detergent metabolite), and 4-*tert*-butylphenol (adhesive) with typical concentrations in the range of 50–1500 ng/m³. The penta- and tetrabrominated diphenyl ethers (flame retardants) were frequently detected in dust, and 2,3-dibromo-1-propanol, the carcinogenic intermediate of a flame retardant banned in 1977, was detected in air and dust. Twenty-three pesticides were detected in air and 27 were detected in dust, the most abundant being

Introduction

Current widespread interest in a range of health effects potentially associated with endocrine-disrupting compounds (EDCs) has made exposure assessment for these compounds a priority. Studies of potential health effects associated with EDCs have been hampered by lack of information about the major sources of exposure to EDCs. Furthermore, because many EDCs act additively through a common mechanism of action or have antagonistic or other interactive effects by operating at different points in cell signaling systems, consideration of exposure to mixtures is critical in studies of health effects (*1*–*7*). These questions are particularly important in relation to indoor environments, which have been identified as an important source of chemical exposures (*8*–*11*). People spend a large fraction of their time indoors, and indoor sources of chemicals, coupled with limited ventilation and slow chemical degradation processes, cause increased pollutant concentrations indoors. In fact, indoor air specifically has been described as "one of the most serious environmental risks to human health" (*8*).

Many high production volume chemicals—including some already identified as EDCs—have consumer uses (e.g., in plastics, detergents, and other household and consumer products) that make them potentially important indoor contaminants. While a number of comprehensive exposure studies have been conducted or are underway to characterize residential exposures to selected contaminants, particularly volatile organic compounds, pesticides, and polyaromatic hydrocarbons (PAHs), these studies have been limited to a small number of compounds and have focused on characterizing exposure pathways and sources (*12*–*18*). We were unable to locate exposure data for many of our compounds of interest, including alkylphenols, parabens, polybrominated diphenyl ethers (PBDEs), and many of the estrogenic phenolic compounds such as bisphenol A. We located only one (unpublished) study of substantial size that has characterized phthalate concentrations in indoor air (*18*).

The primary objective of this study is to provide an assessment of household exposure to a broad suite of organic chemicals that have been identified as EDCs. Indoor air and dust were selected for analysis because many EDCs are used in consumer products and building materials (*6*, *19*), so these chemicals would be expected indoors. Indoor air has been identified as an important source of chemical exposure, while house dust has been demonstrated to be an important exposure pathway in young children (*20*). Dust also provides

TABLE 4. Most Abundant Chemicals

Ten Chemicals with Highest 90th Percentile Concentrations

air (ng/m³)[a]	dust (μg/g)[a]
diethyl phthalate (1,600) *100*	bis(2-ethylhexyl) phthalate (854) *100*
o-phenylphenol (440) *100*	benzyl butyl phthalate (277) *100*
di-*n*-butyl phthalate (430) *100*	di-*n*-butyl phthalate (43.9) *98*
4-nonylphenol (230) *100*	nonylphenol diethoxylate (18.9) *86*
bis(2-ethylhexyl) phthalate (210) *68*	bis(2-ethylhexyl) adipate (16.6) *100*
diisobutyl phthalate (150) *100*	*trans*-permethrin (16.5) *53*
benzyl butyl phthalate (68) *44*	piperonyl butoxide (15.1) *66*
4-*tert*-butylphenol (43) *100*	diethyl phthalate (10.8) *89*
nonylphenol monoethoxylate (41) *95*	nonylphenol monoethoxylate (8.55) *86*
bis(2-ethylhexyl) adipate (22) *99*	*cis*-permethrin (7.04) *45*

10 Pesticides with Highest 90th Percentile Concentrations

air (ng/m³)[a]	dust (μg/g)[a]
o-phenylphenol (440) *100*	*trans*-permethrin (16.5) *53*
heptachlor[b] (19) *44*	piperonyl butoxide (15.1) *66*
propoxur (16) *49*	*cis*-permethrin (7.04) *45*
γ-chlordane[b] (12) *53*	methoxychlor[b] (3.38) *54*
chlorpyrifos (12) *38*	4,4′-DDT[b](3.19) *65*
pentachlorophenol[b] (10) *58*	pentachlorophenol[b] (2.42) *86*
diazinon (9.0) *40*	chlorpyrifos[b] (1.87) *18*
α-chlordane[b] (8.8) *51*	carbaryl (1.72) *43*
chlorothalonil (3.4) *17*	propoxur (1.70) *42*
3,5,6-trichloro-2-pyridinol (1.1) *13*	bendiocarb (1.11) *12*

[a] Percent detection in italics. [b] Indicates banned or restricted-use pesticide (at time of sample collection).

Rudel et al., 2008 Epidemiology

Correlations Between Urinary Phthalate Metabolites and Phthalates, Estrogenic Compounds 4-Butyl phenol and o-Phenyl phenol, and Some Pesticides in Home Indoor Air and House Dust

Rudel, R A*; Dodson, R E*; Newton, E†; Zota, A R*; Brody, J G*

Epidemiology: November 2008 - Volume 19 - Issue 6 - p S332
doi: 10.1097/01.ede.0000340529.83416.do
Abstracts: ISEE 20th Annual Conference, Pasadena, California, October 12–16, 2008: Contributed Abstracts

"Overall, these data show that concentrations of many EDCs in biological samples and indoor air and dust co-vary, suggesting that some EDC mixtures may originate from common exposure sources and highlighting potential confounding by other EDCs in health effect studies of phthalates. Future work will utilize factor analysis to **identify source profiles** of EDC mixtures that are associated with urinary phthalate levels."

2013: ExpoCast Team "Helps" With Correlating Exposure and Biomarkers

- The EPA's Exposure Forecaster (ExpoCast) team works to develop new methods for exposure

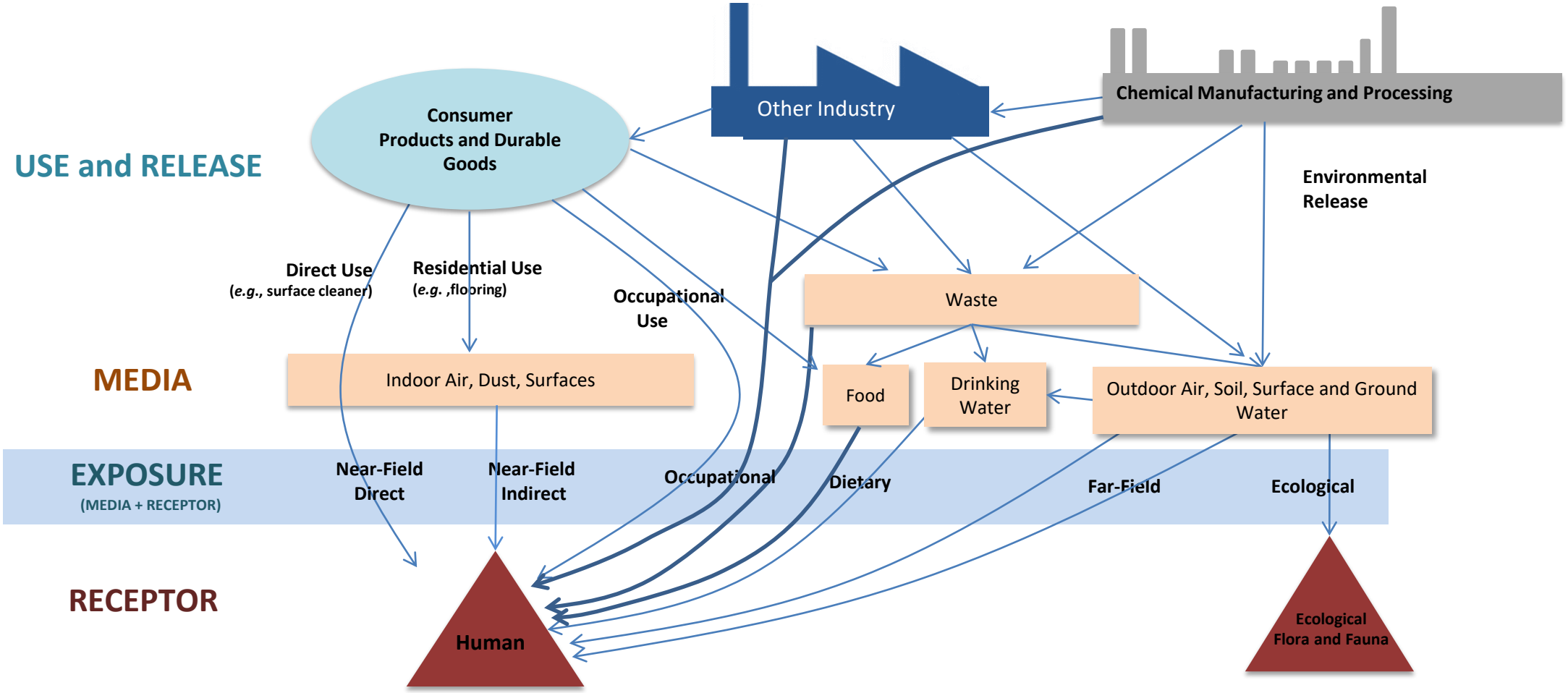
- To correlate metabolites in urine with parent chemical exposures from dust and consumer products we need to know (at least):

    - What chemicals are we talking about? CAS are not unique! Names are certainly not unique.
    - Are there any toxicity data?
    - Which metabolites link to which parent chemicals?
    - What consumer products contain which chemicals?
    - What does a given concentration of a chemical in urine (or plasma) imply about total body burden and exposure?
    - What are the relevant physicochemical properties?

- In 2013 when we set out to do this analysis **we could not answer most of these questions**

- Now you can with the CompTox dashboard: https://comptox.epa.gov/dashboard

Understanding Exposure is a Systems Problem

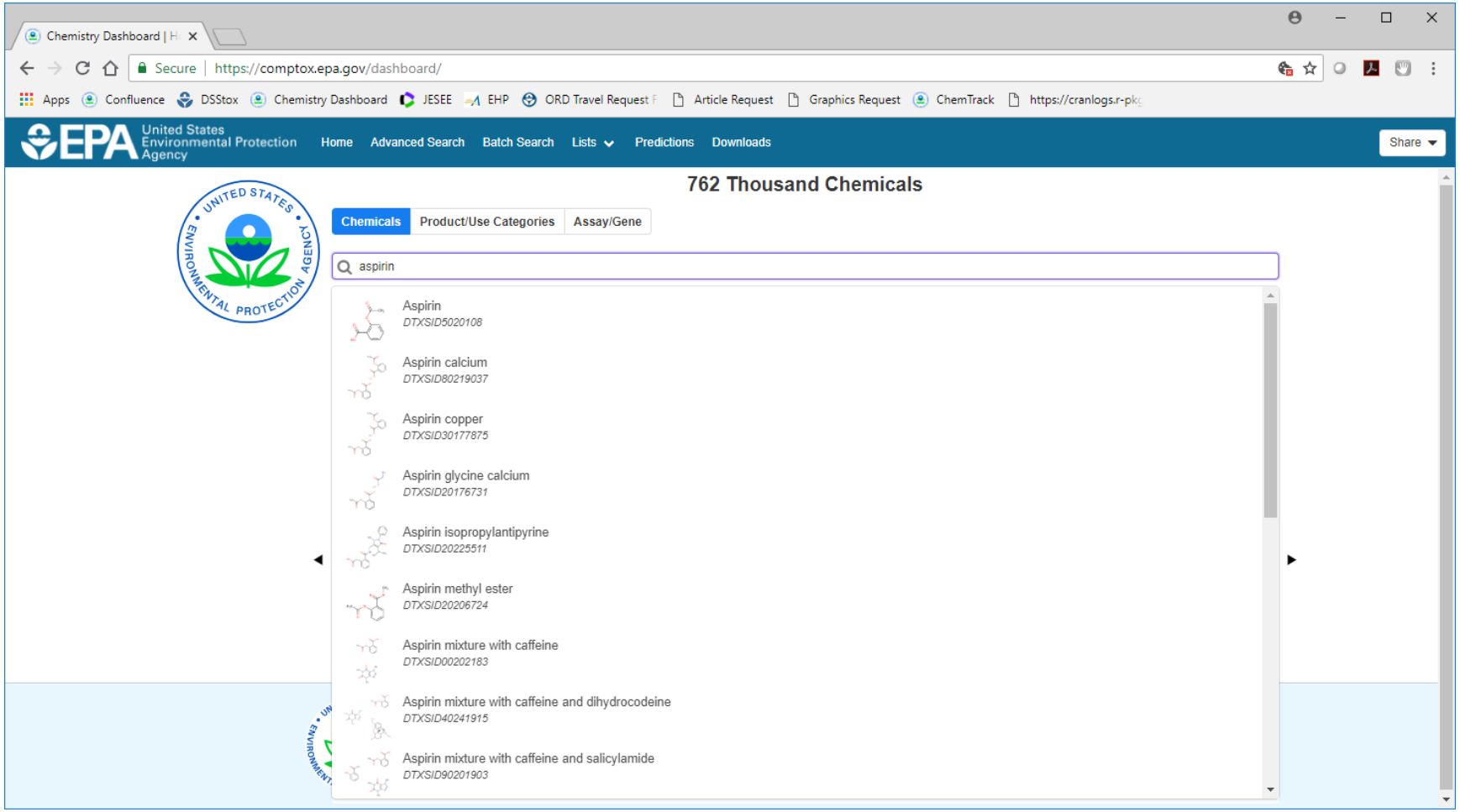


- **Exposure event unobservable:** Can try to predict exposure by characterizing pathway
- Some pathways have much higher average exposures: In home "Near field" sources significant (Wallace, *et al.,* 1987)

Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

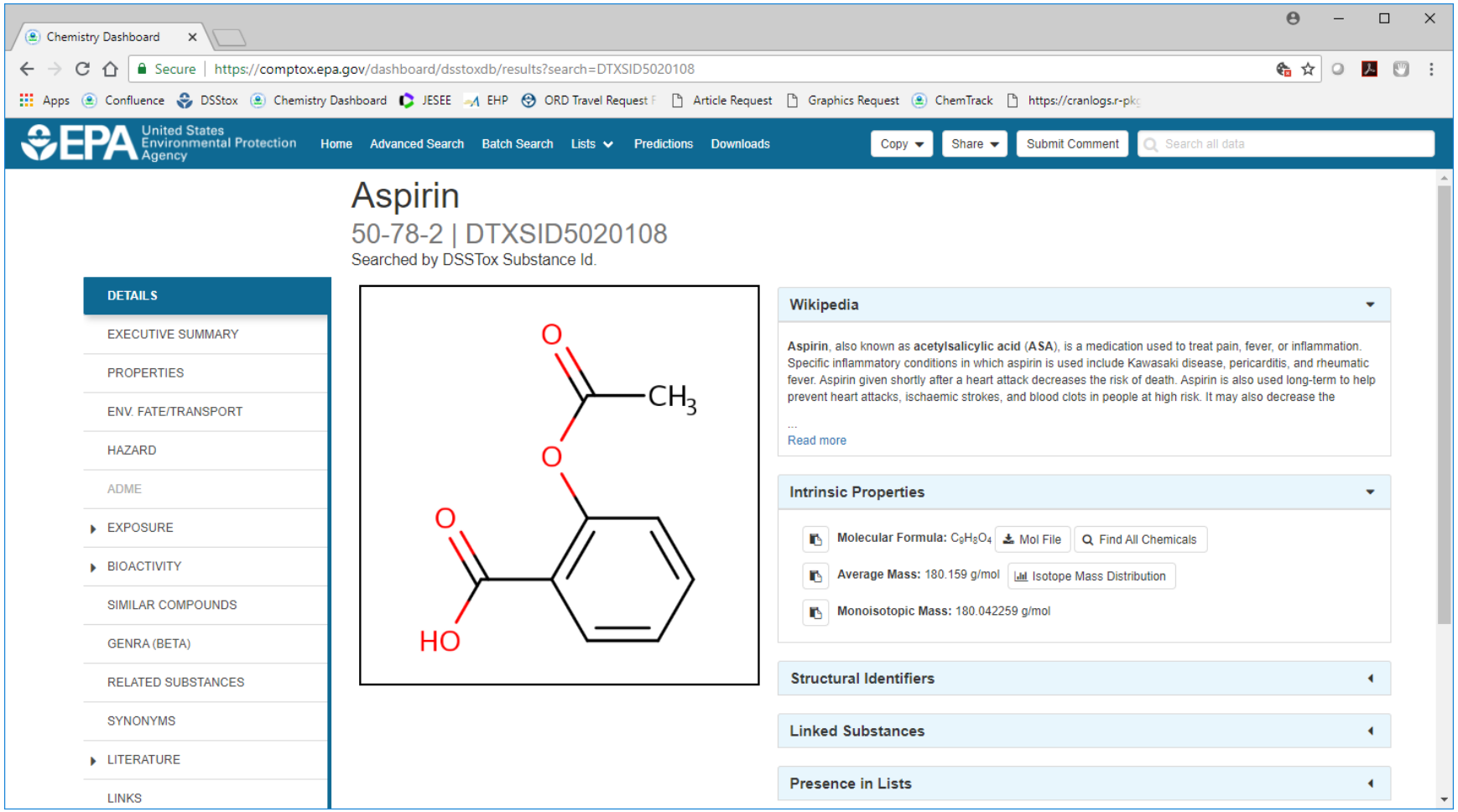
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

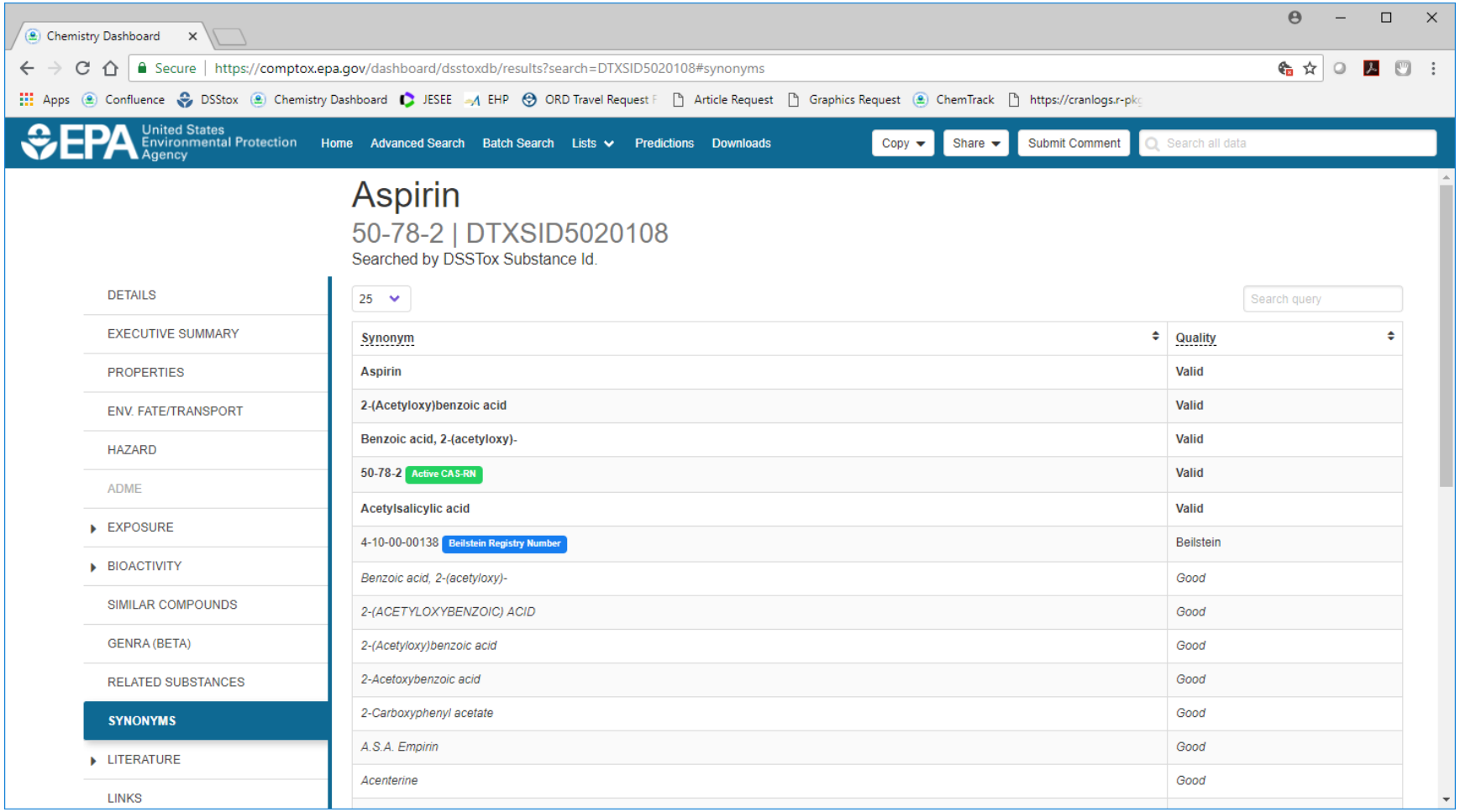
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

https://comptox.epa.gov/dashboard/

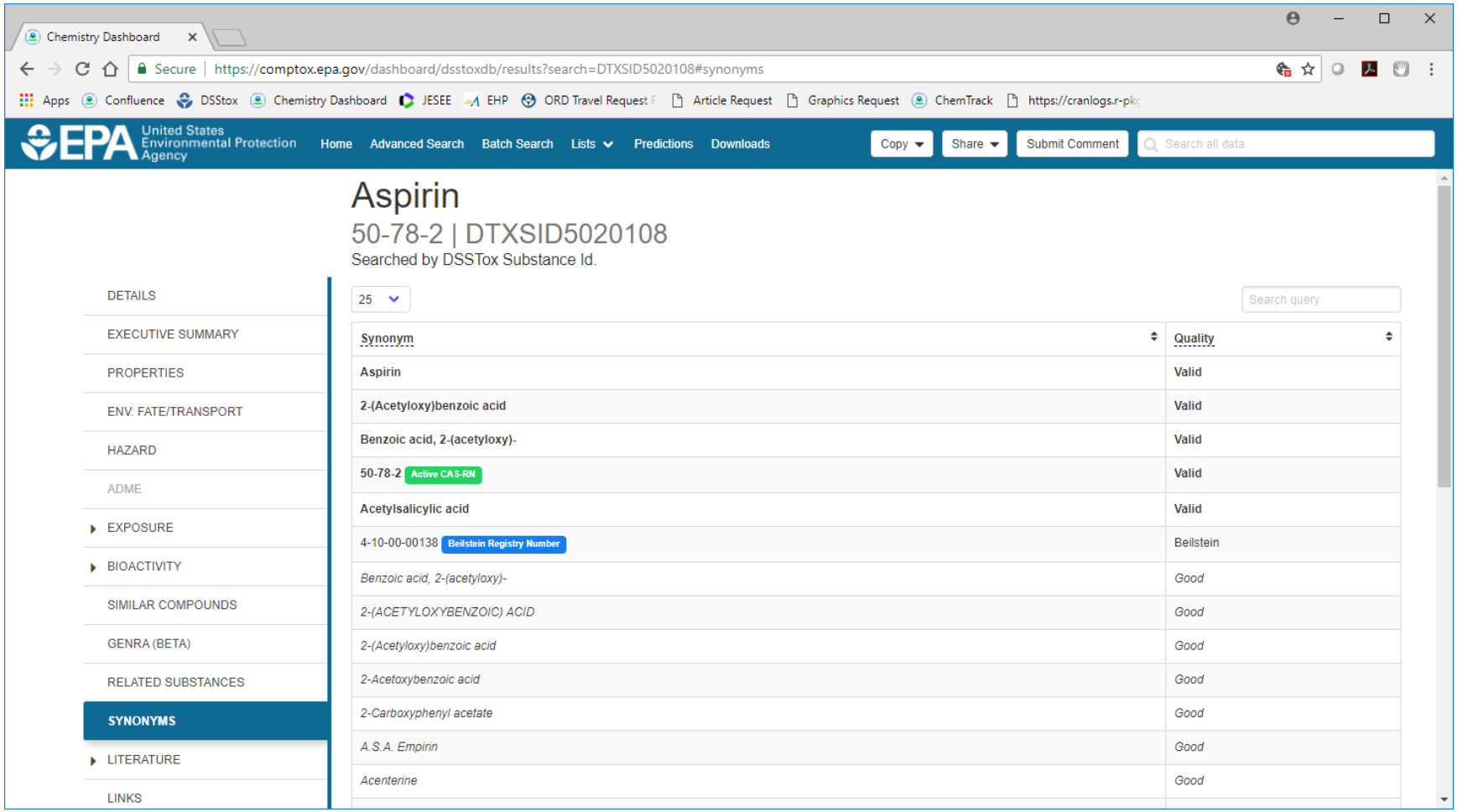


Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

https://comptox.epa.gov/dashboard/

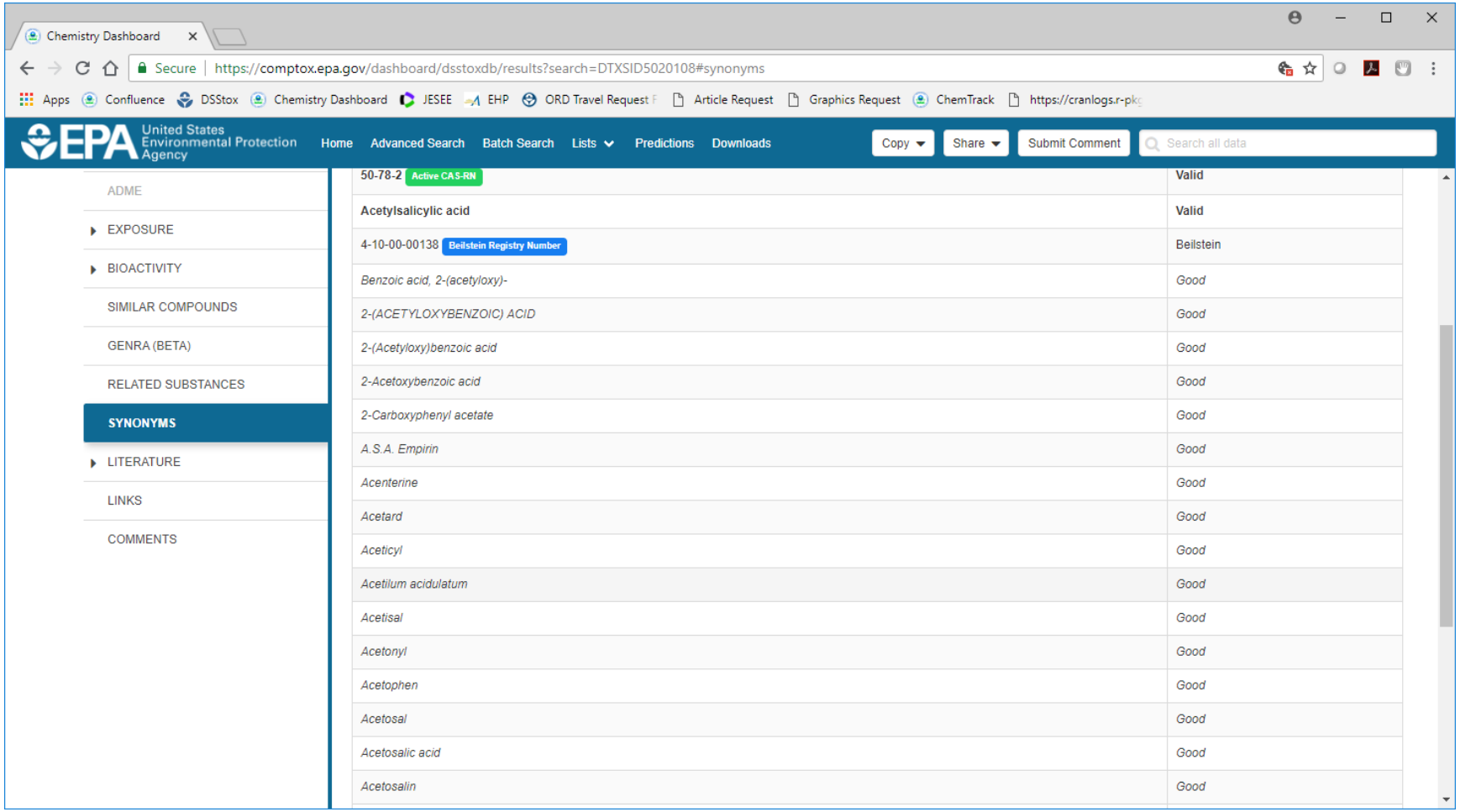
50-78-2 Active CAS-RN



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

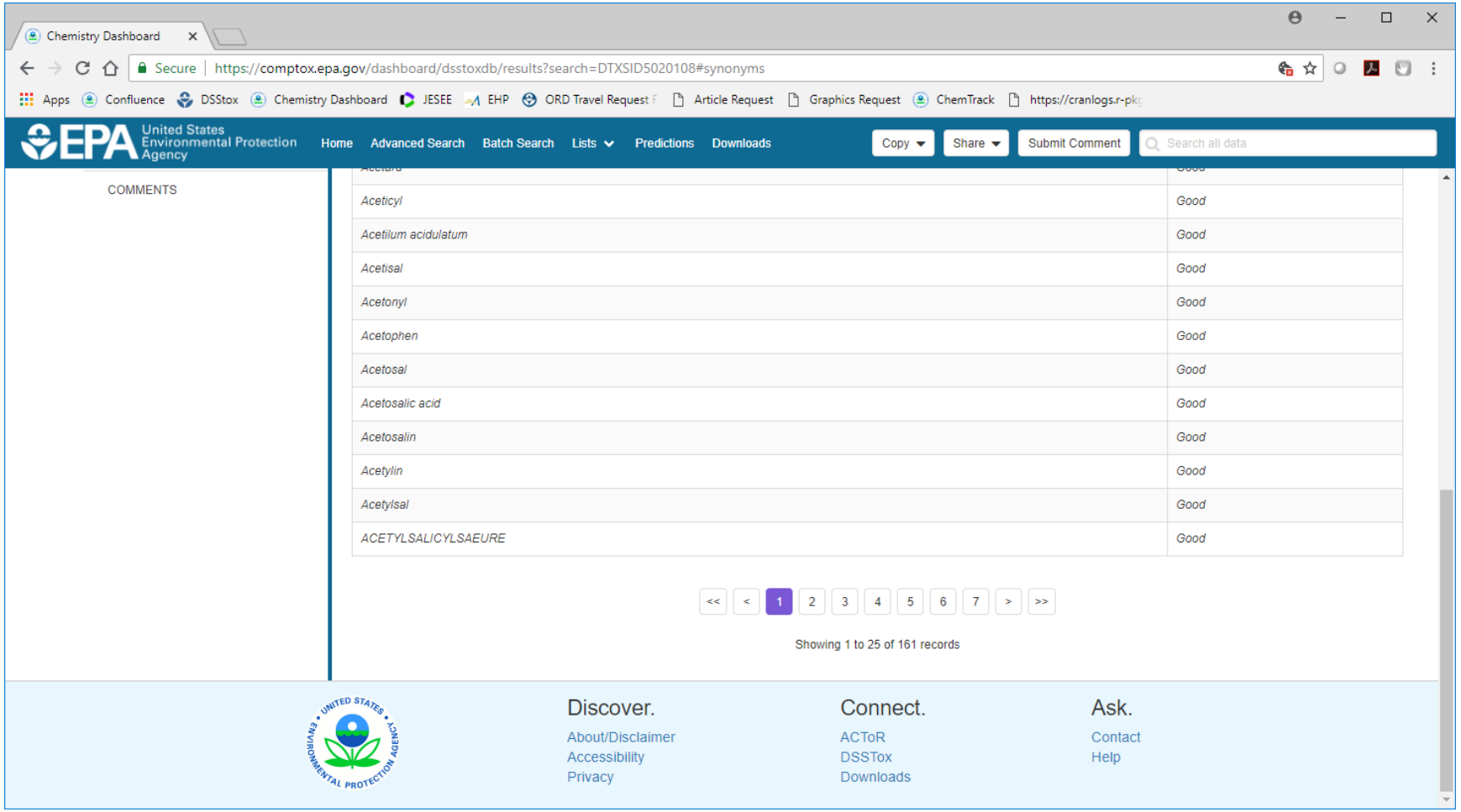
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

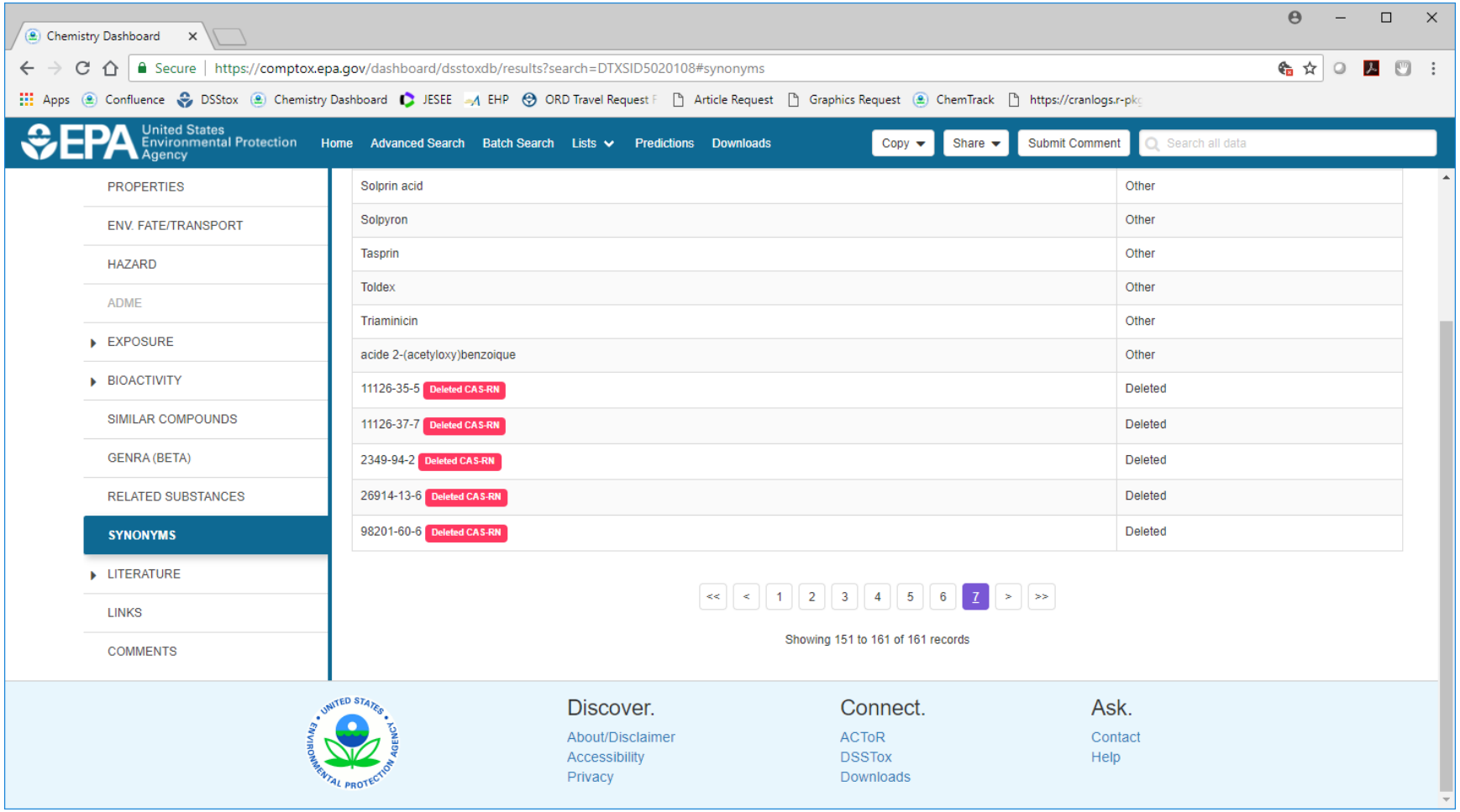
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

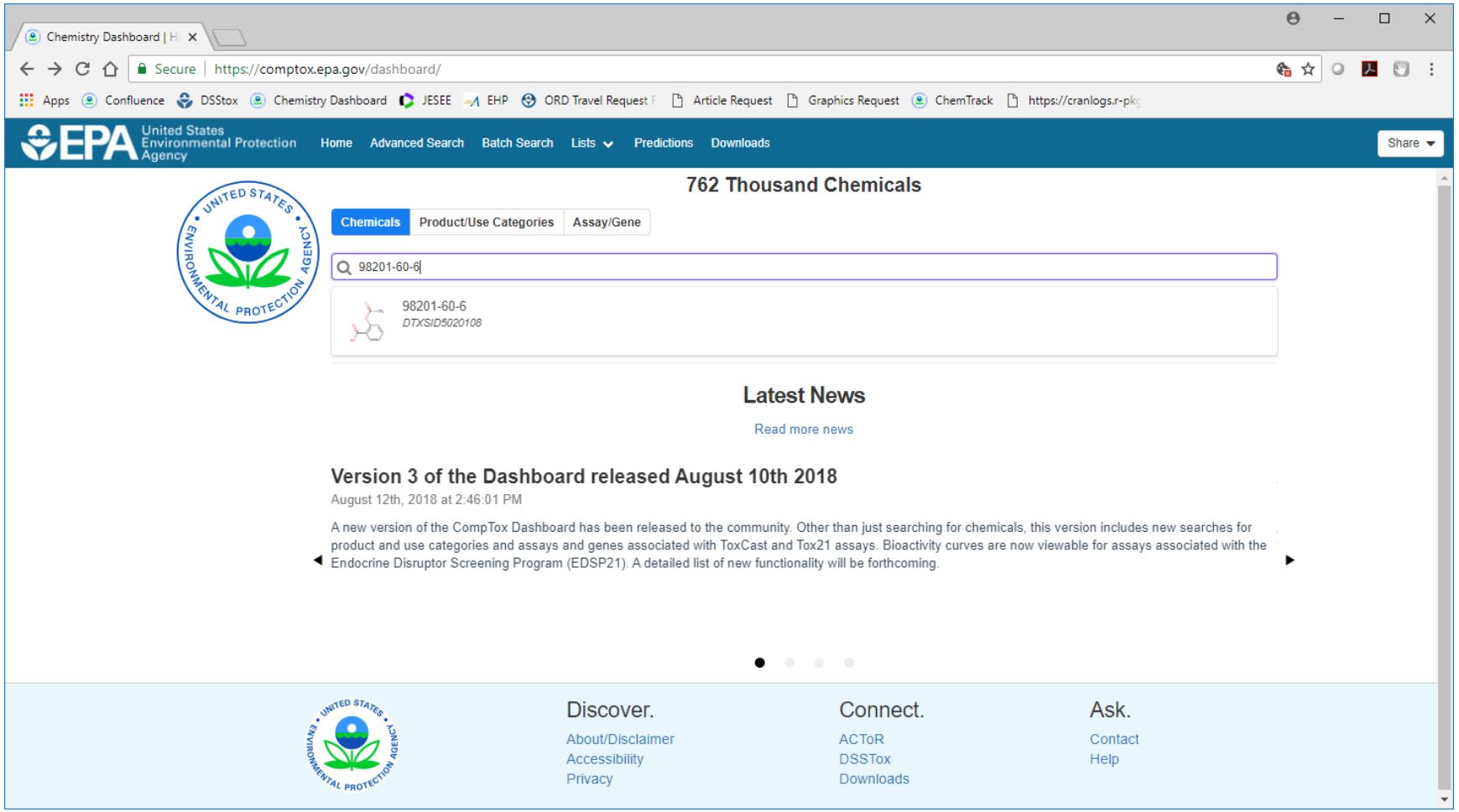
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Chemical Abstracts Service (CAS) Registry Numbers are not unique:
  - Aspirin (CAS #: 50-78-2)
  - Deleted CAS #: 2349-94-2, 11126-35-5, 11126-37-7, 26914-13-6, 98201-60-6

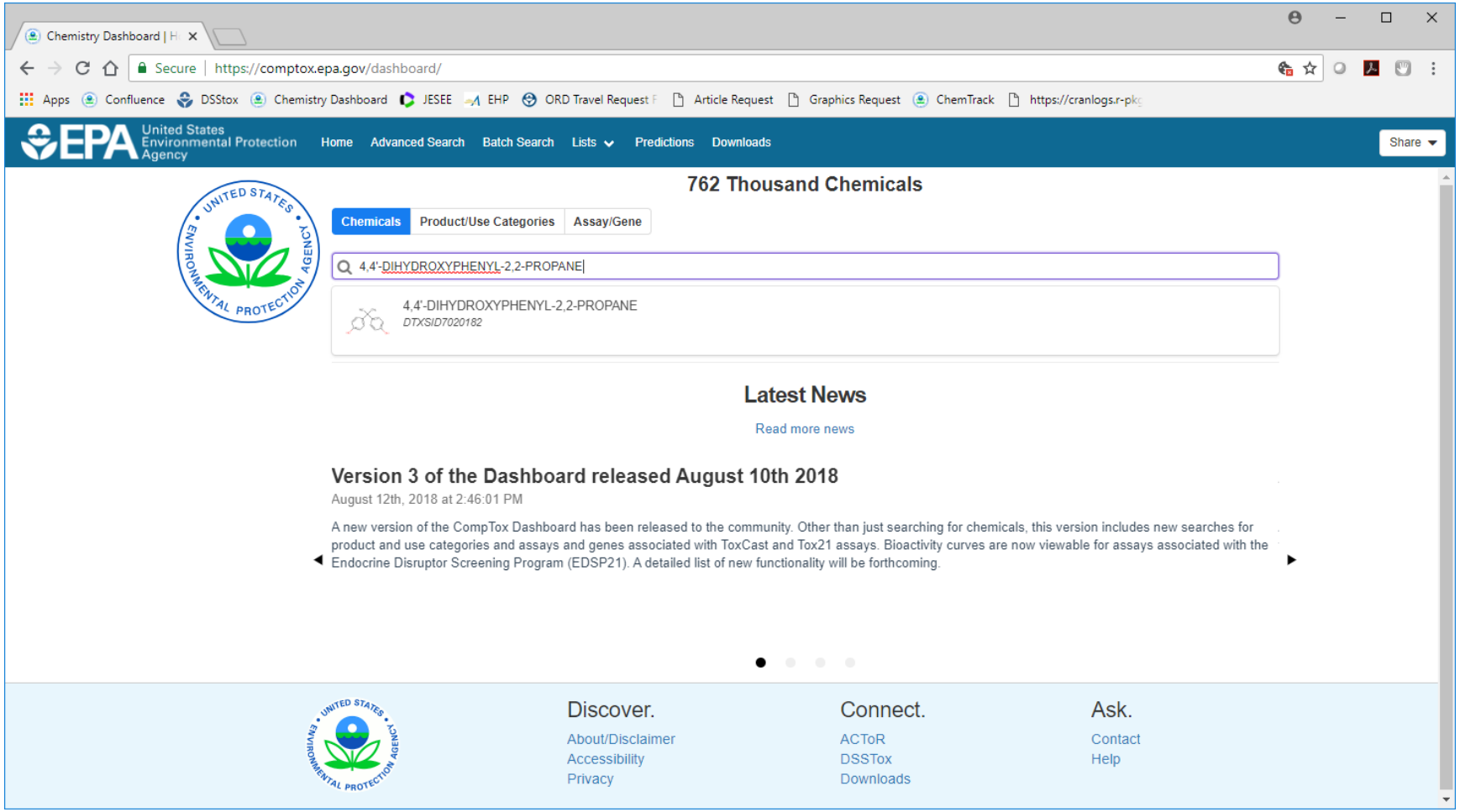
https://comptox.epa.gov/dashboard/



Chemical Identity

- What chemicals are we talking about? Names are certainly not unique: *4,4'-DIHYDROXYPHENYL-2,2-PROPANE*

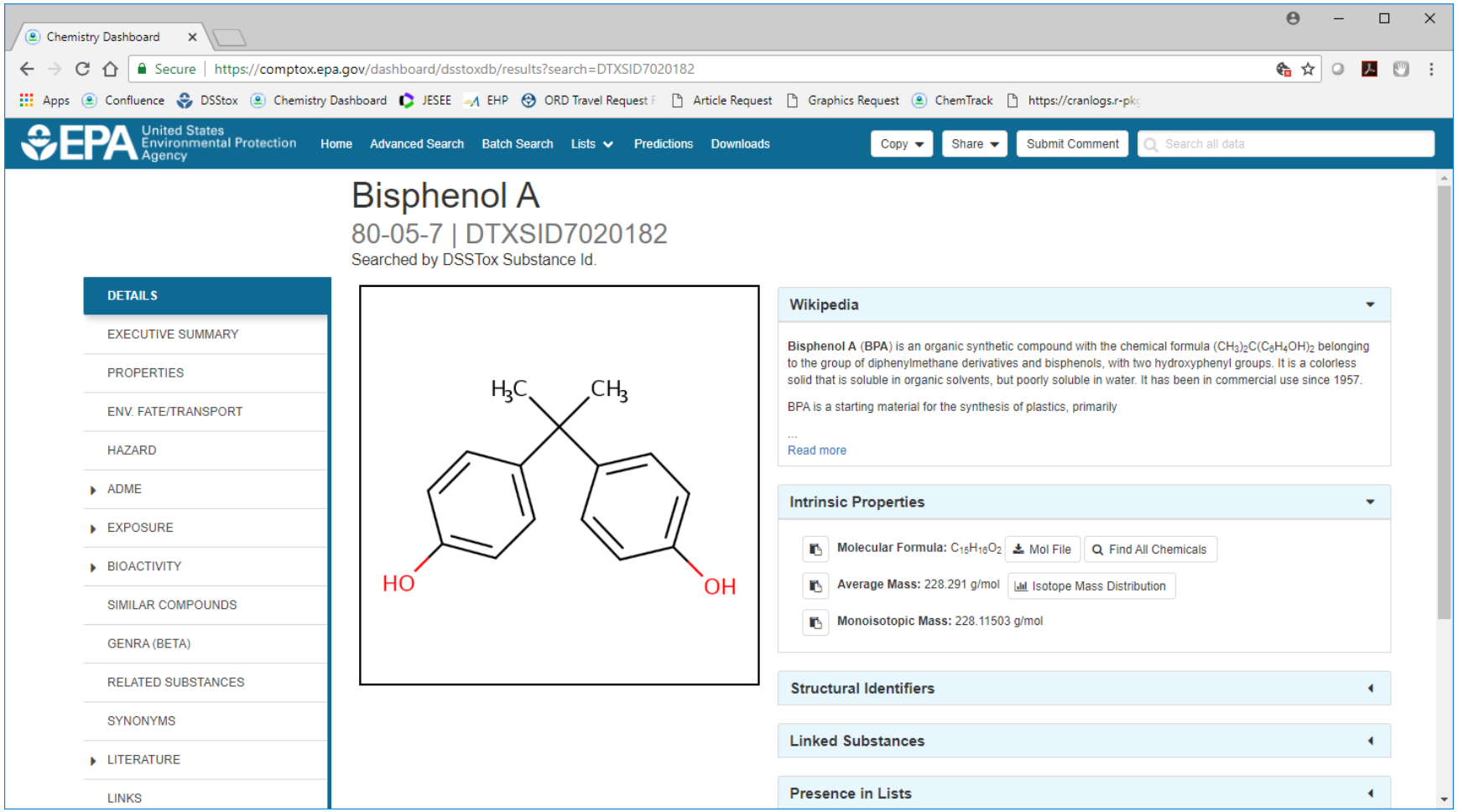
https://comptox.epa.gov/dashboard/



Chemical Identity

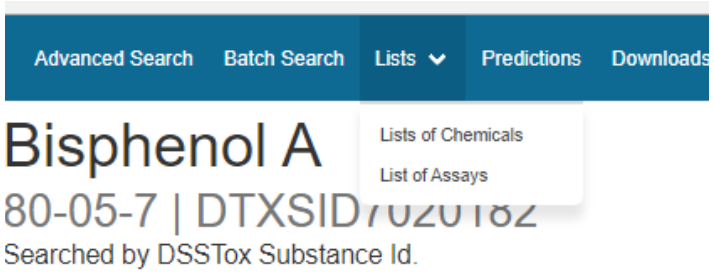
- What chemicals are we talking about? Names are certainly not unique: *4,4'-DIHYDROXYPHENYL-2,2-PROPANE*

https://comptox.epa.gov/dashboard/

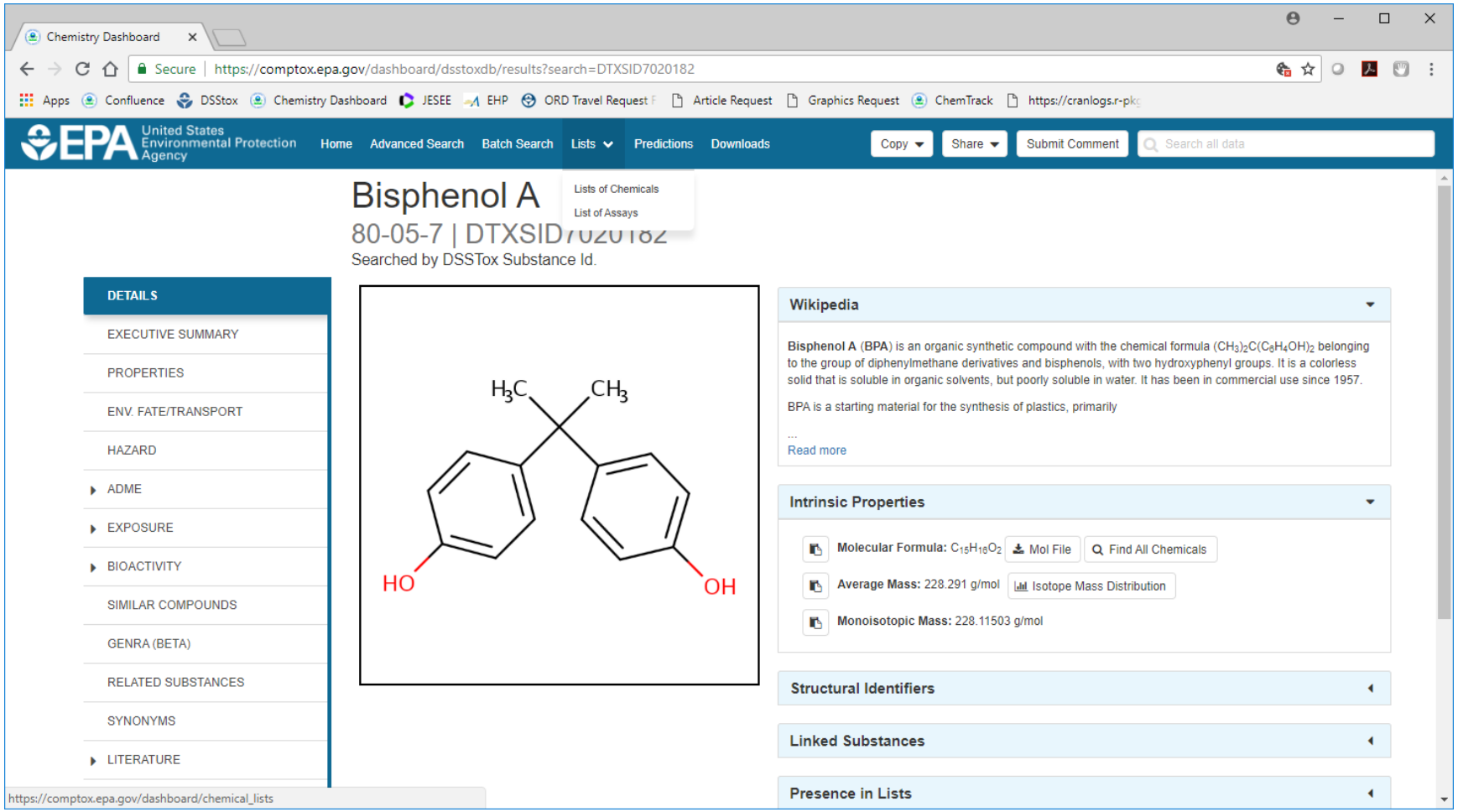


Chemical Lists

- Can find chemicals by lists



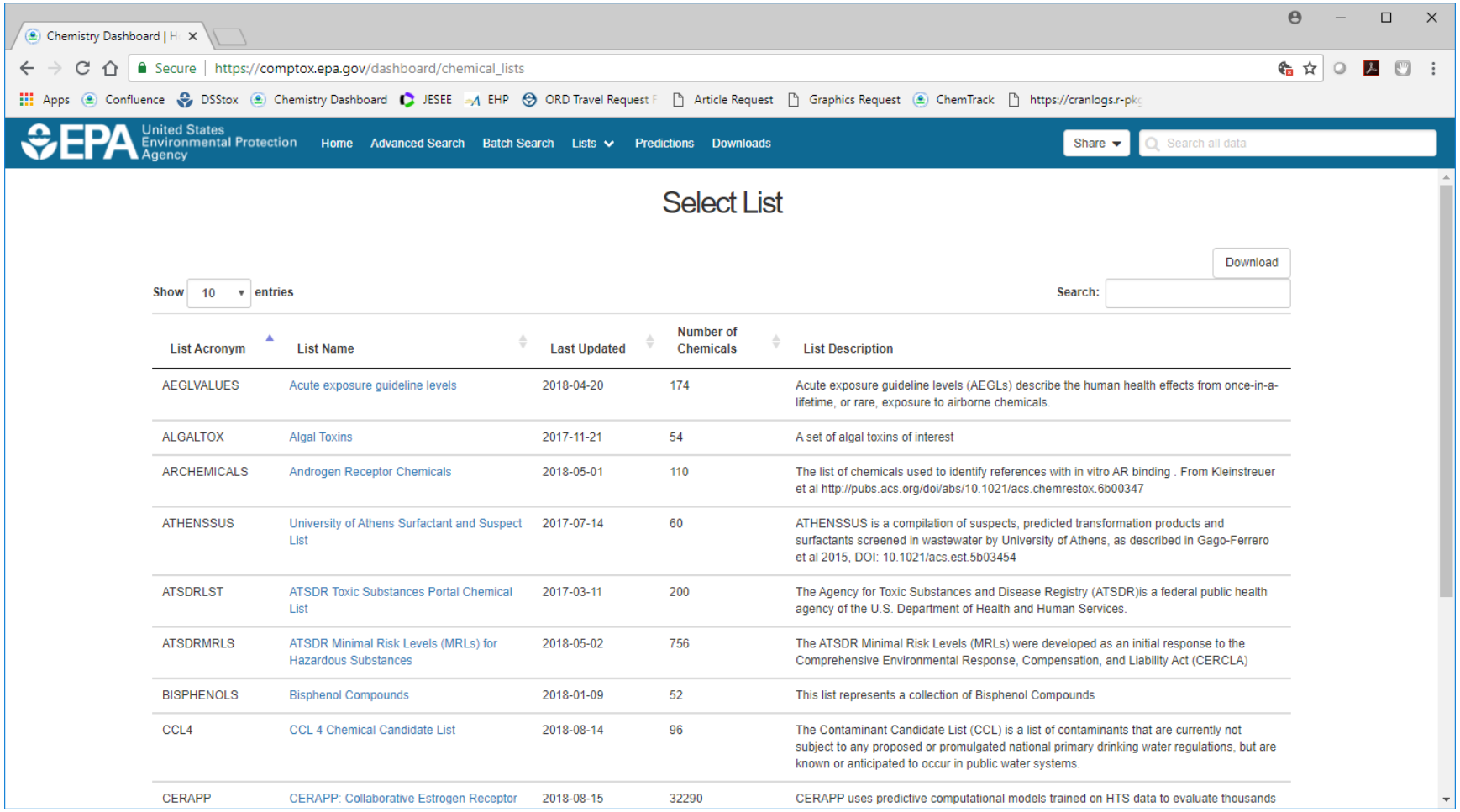
https://comptox.epa.gov/dashboard/



Chemical Lists

- Can find chemicals by lists

https://comptox.epa.gov/dashboard/

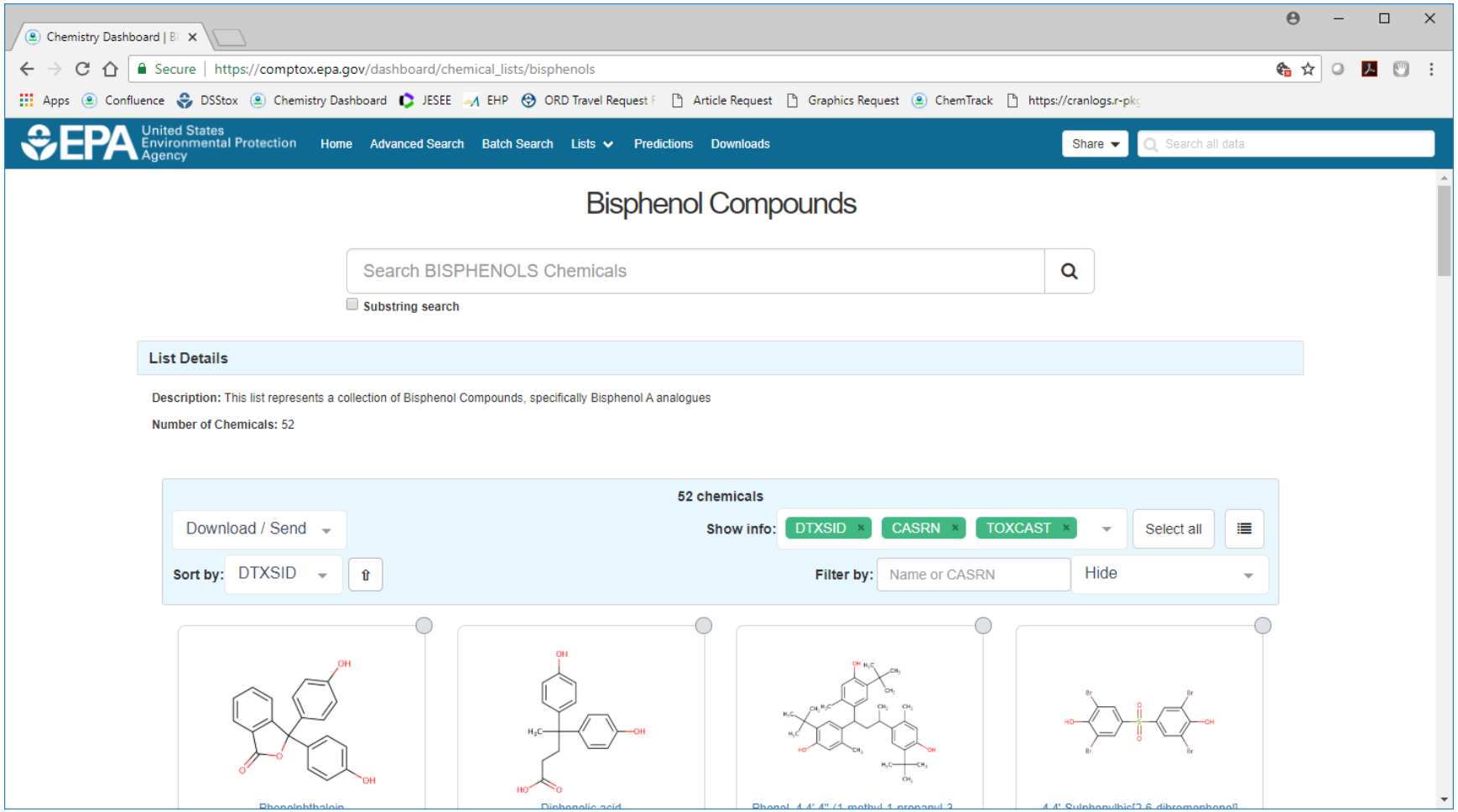


List Acronym	List Name	Last Updated	Number of Chemicals	List Description
AEGLVALUES	Acute exposure guideline levels	2018-04-20	174	Acute exposure guideline levels (AEGLs) describe the human health effects from once-in-a-lifetime, or rare, exposure to airborne chemicals.
ALGALTOX	Algal Toxins	2017-11-21	54	A set of algal toxins of interest
ARCHEMICALS	Androgen Receptor Chemicals	2018-05-01	110	The list of chemicals used to identify references with in vitro AR binding . From Kleinstreuer et al http://pubs.acs.org/doi/abs/10.1021/acs.chemrestox.6b00347
ATHENSSUS	University of Athens Surfactant and Suspect List	2017-07-14	60	ATHENSSUS is a compilation of suspects, predicted transformation products and surfactants screened in wastewater by University of Athens, as described in Gago-Ferrero et al 2015, DOI: 10.1021/acs.est.5b03454
ATSDRLST	ATSDR Toxic Substances Portal Chemical List	2017-03-11	200	The Agency for Toxic Substances and Disease Registry (ATSDR)is a federal public health agency of the U.S. Department of Health and Human Services.
ATSDRMRLS	ATSDR Minimal Risk Levels (MRLs) for Hazardous Substances	2018-05-02	756	The ATSDR Minimal Risk Levels (MRLs) were developed as an initial response to the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA)
BISPHENOLS	Bisphenol Compounds	2018-01-09	52	This list represents a collection of Bisphenol Compounds
CCL4	CCL 4 Chemical Candidate List	2018-08-14	96	The Contaminant Candidate List (CCL) is a list of contaminants that are currently not subject to any proposed or promulgated national primary drinking water regulations, but are known or anticipated to occur in public water systems.
CERAPP	CERAPP: Collaborative Estrogen Receptor	2018-08-15	32290	CERAPP uses predictive computational models trained on HTS data to evaluate thousands

Chemical Lists

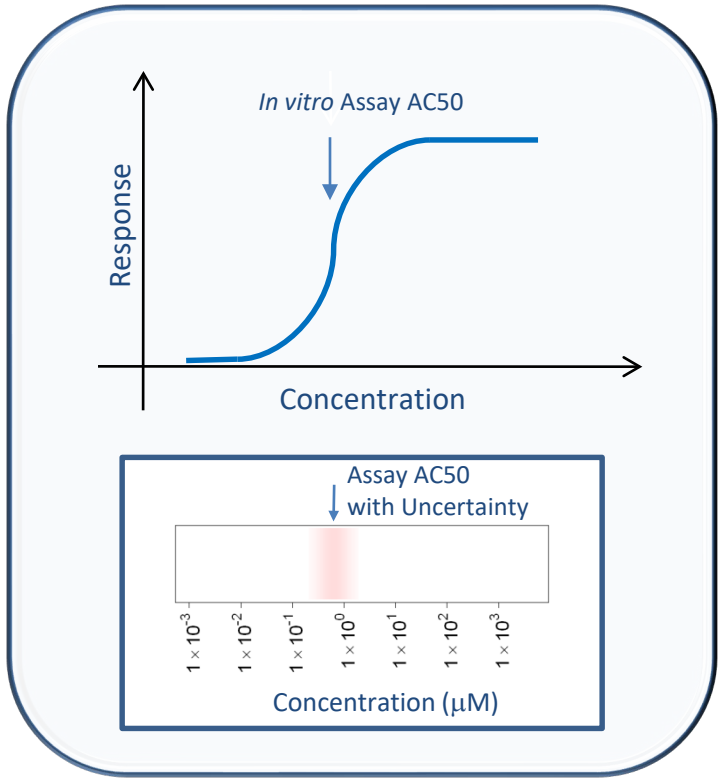
- Can find chemicals by lists

https://comptox.epa.gov/dashboard/



High-Throughput Bioactivity Screening

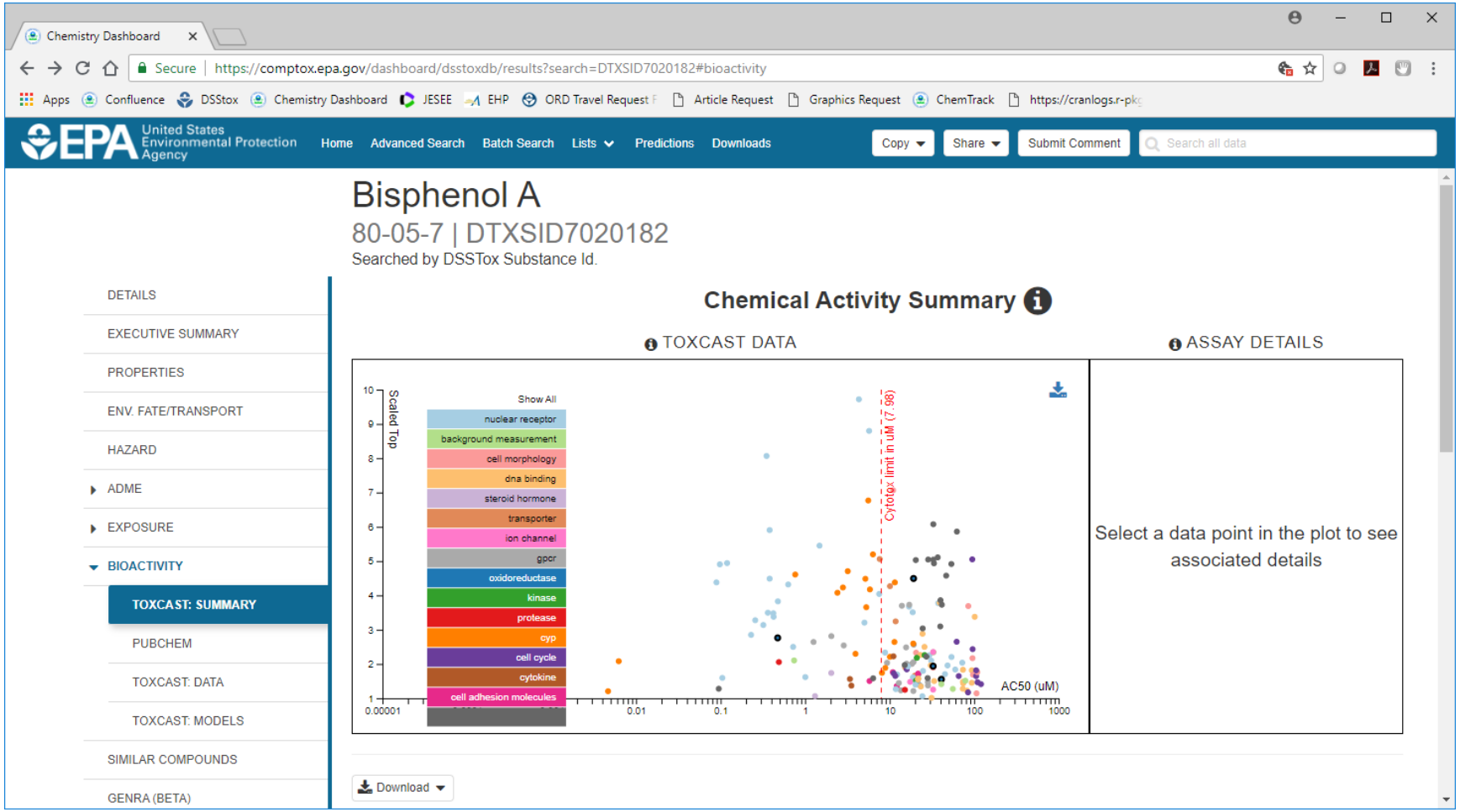
- The dashboard provides *in vivo* and *in vitro* toxicity data, where available
- **Most chemicals do not have *in vivo* toxicity data available (Judson, 2008)**
  - Bisphenol A vs. Bisphenol S
- **Tox21**: Examining >10,000 chemicals using ~50 assays intended to identify interactions with biological pathways (Schmidt, 2009)
- **EPA Toxicity Forecaster (ToxCast)**:
  - For a subset (>3000) of Tox21 chemicals run >1000 additional assay endpoints (Judson et al., 2010)
- Data are being revised, new chemicals tested, new assays added



Chemical Bioactivity Data

- Data from the ToxCast and Tox21 projects are available through the dashboard

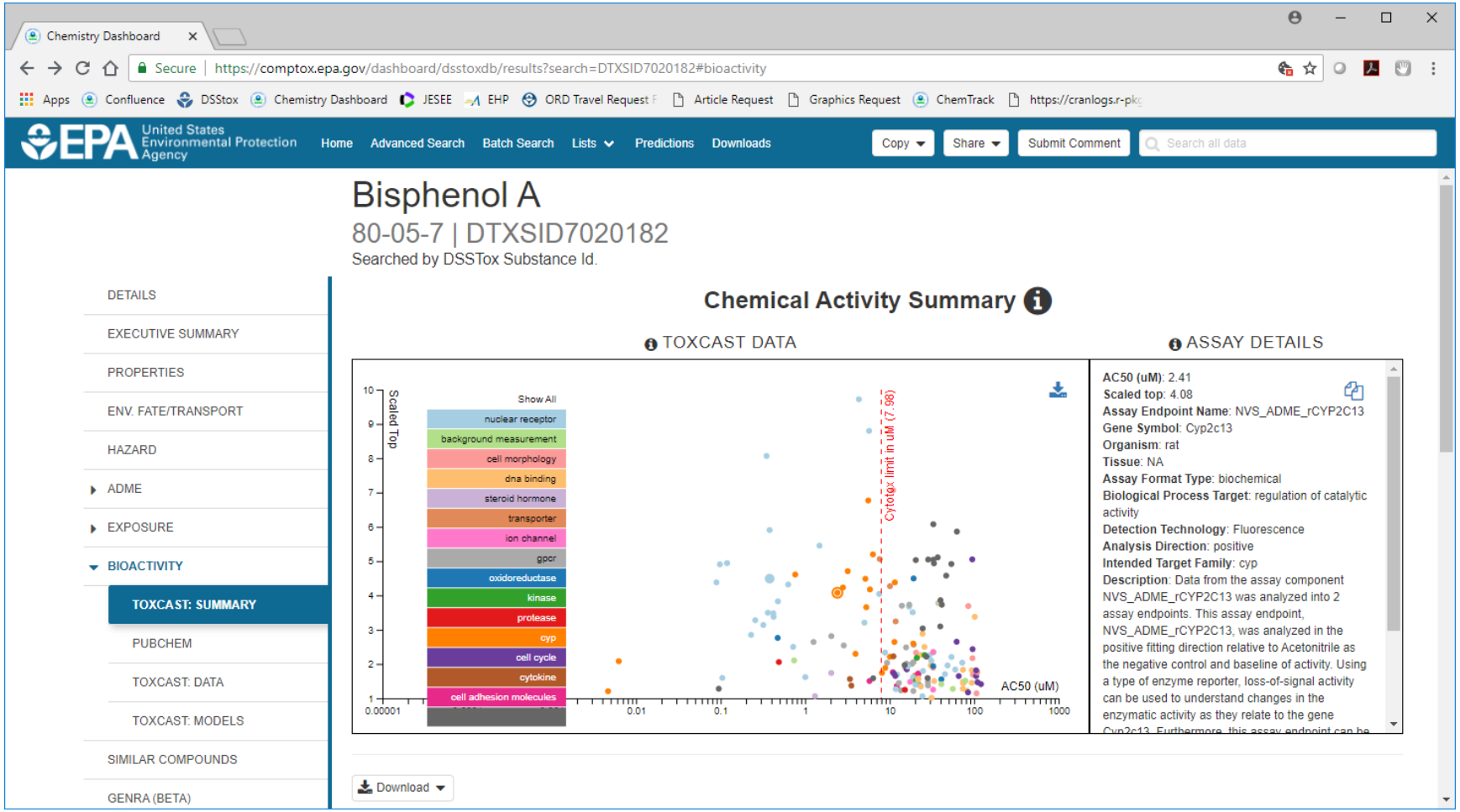
https://comptox.epa.gov/dashboard/



Chemical Bioactivity Data

- Data from the ToxCast and Tox21 projects are available through the dashboard

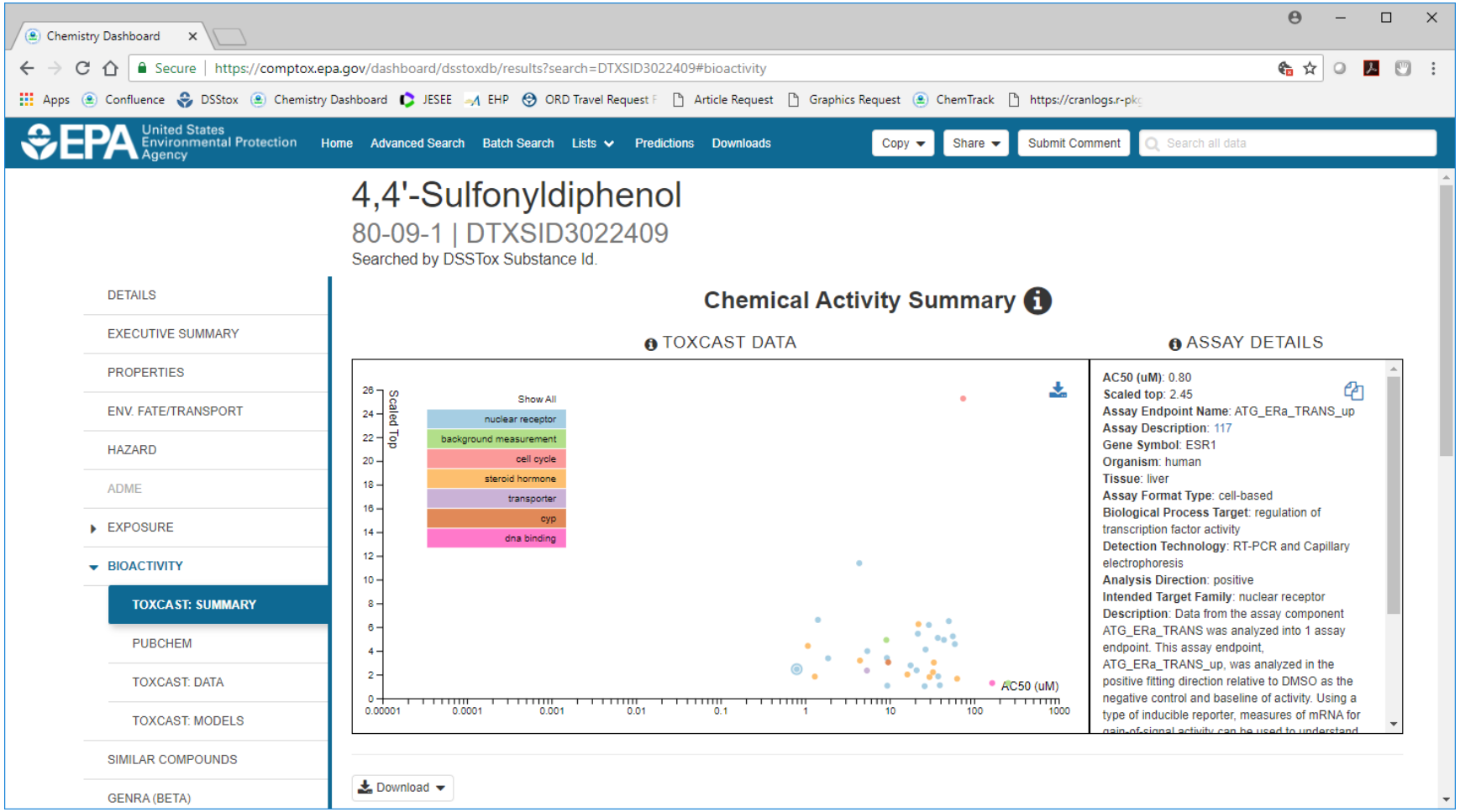
https://comptox.epa.gov/dashboard/



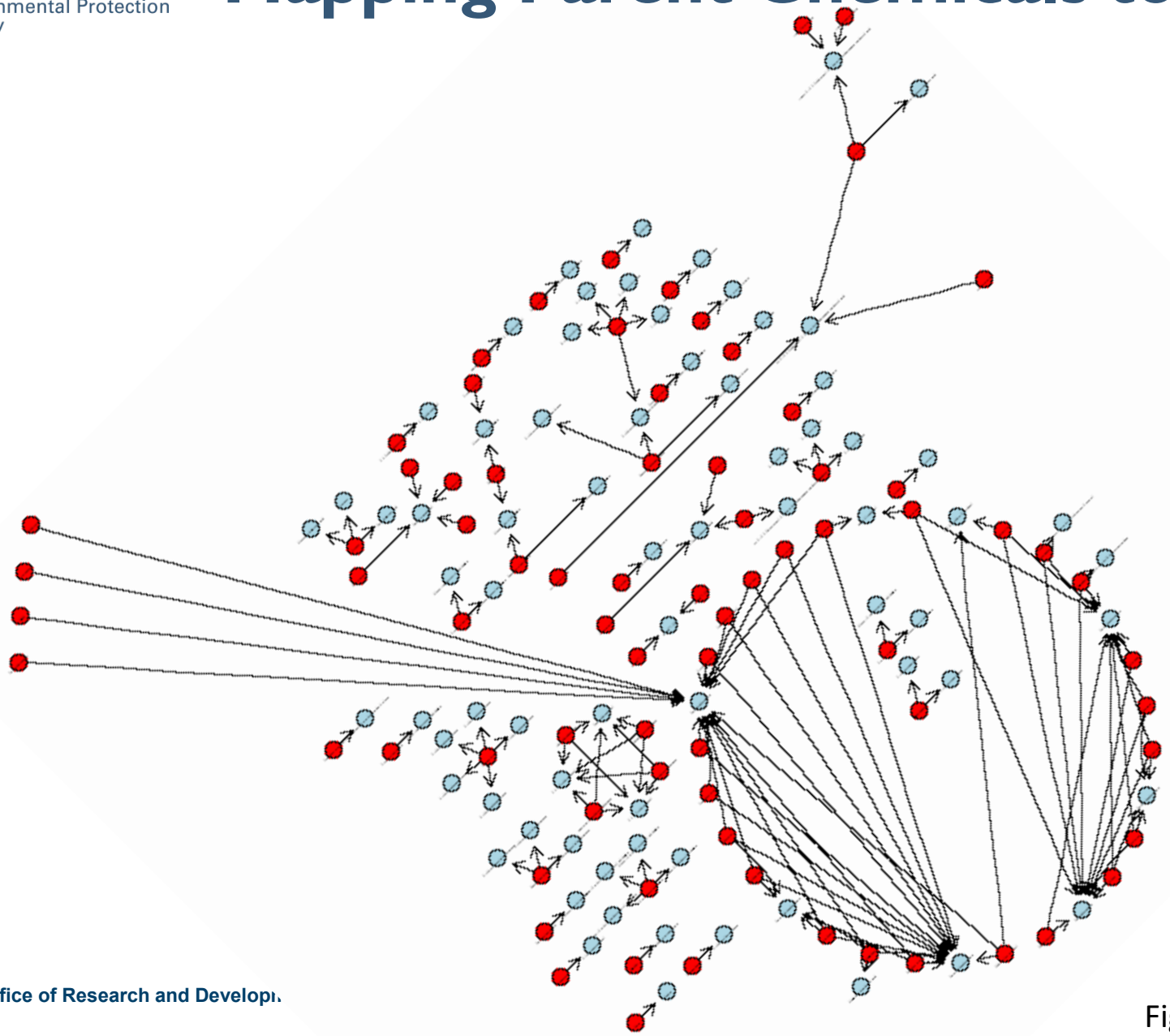
Chemical Bioactivity Data

- Data from the ToxCast and Tox21 projects are available through the dashboard

https://comptox.epa.gov/dashboard/



Mapping Parent Chemicals to NHANES Analytes



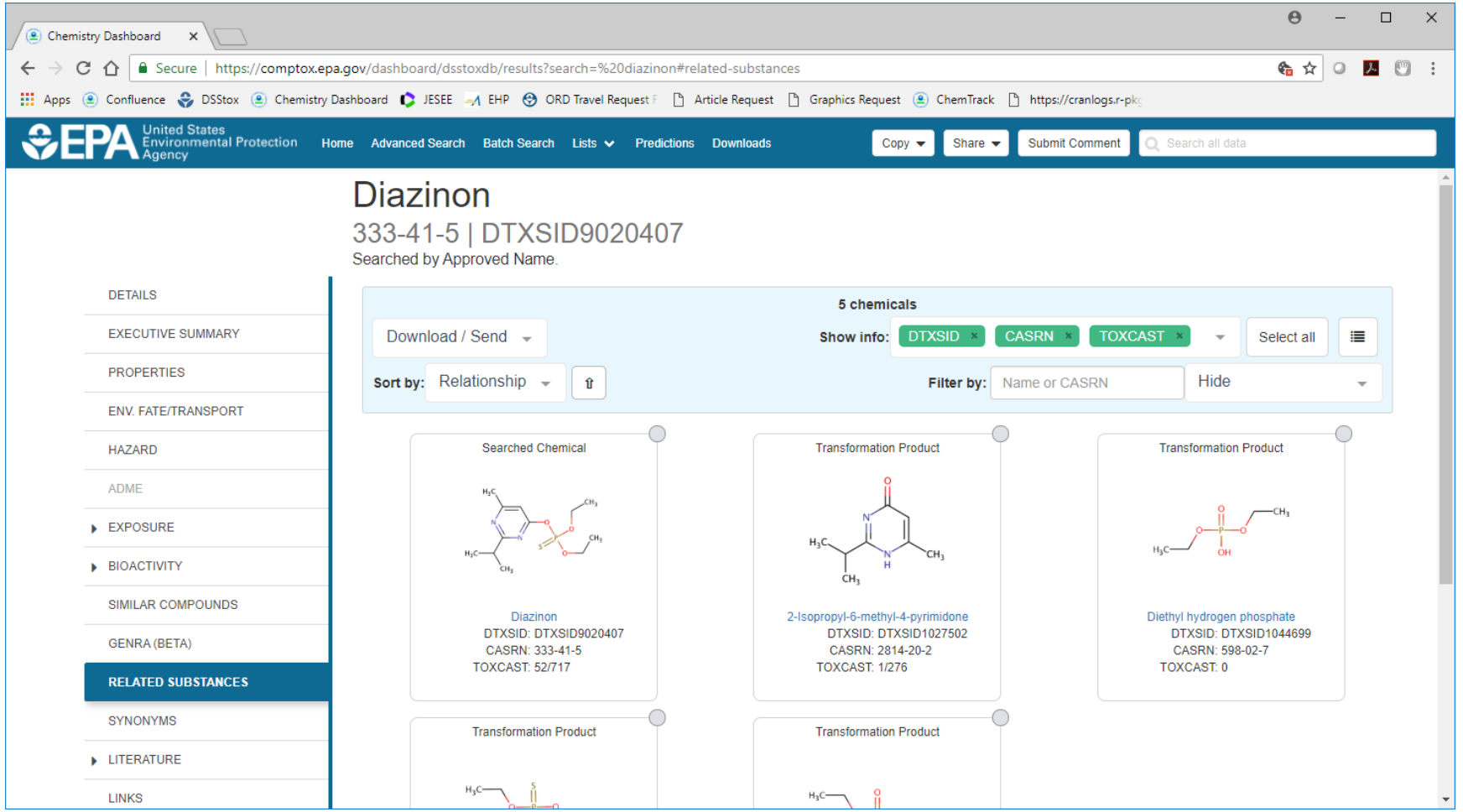
Parent chemical

Analyte measured in urine

Figure from Wambaugh *et al.* (2013)

Parent-Metabolite Linkage

https://comptox.epa.gov/dashboard/

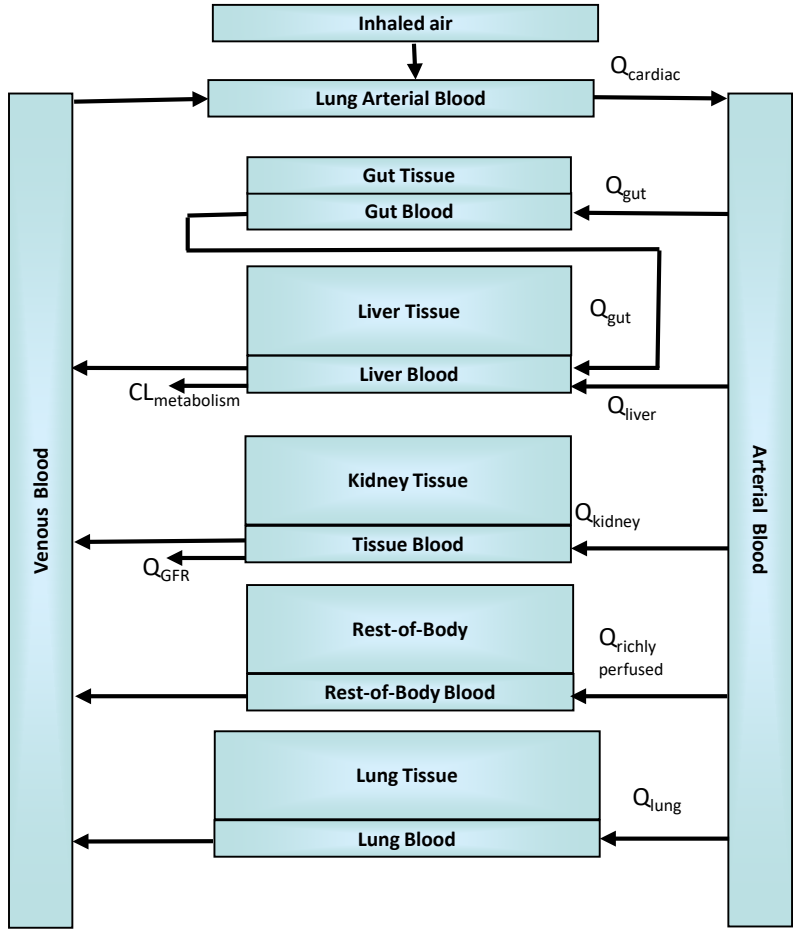


Chemical Toxicokinetics

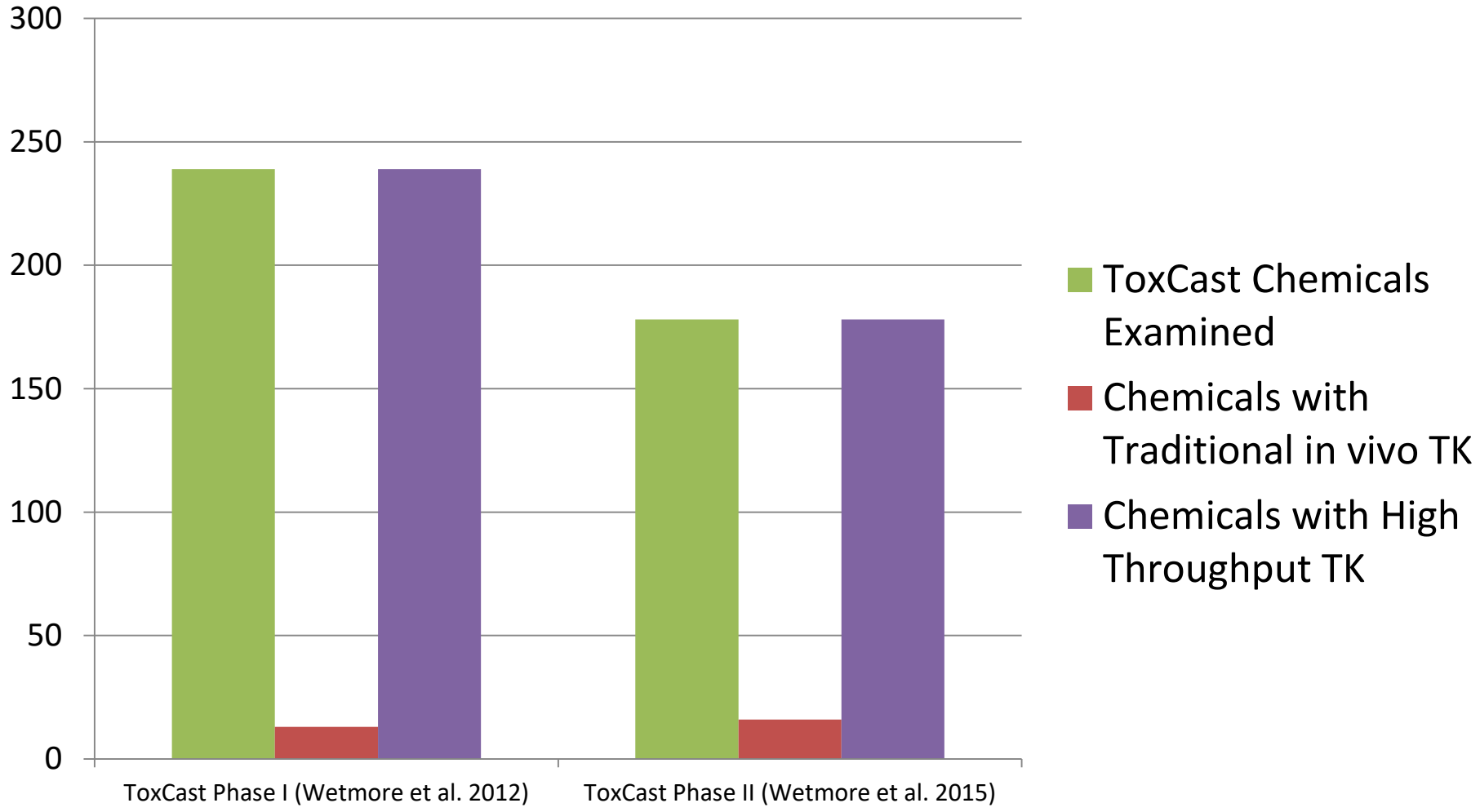
- Toxicokinetics (TK) provides a bridge between toxicity and exposure assessment by predicting tissue concentrations due to exposure
  - However traditional TK methods are resource intensive

- Need to understand what the human body does with a chemical:
  - absorption, distribution, metabolism, excretion (ADME)

- Can relate *in vitro* bioactive concentrations (µM) to steady-state human doses (mg/kg body weight/day) using reverse toxicokinetics (Wetmore et al., 2012 and 2015):
  - You divide by the steady-state plasma concentration, C_{ss}, to convert µM to mg/kg body weight/day
  - Dashboard will give you C_{ss} predicted by HTTK



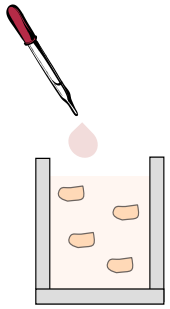
The Need for *In Vitro* Toxicokinetics



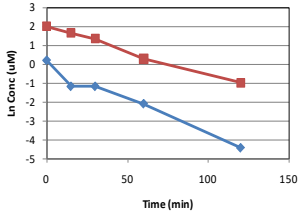
- Studies like Wetmore et al. (2012, 2015), addressed the need for TK data using *in vitro* methods

High-Throughput Toxicokinetics (HTTK) for *In Vitro-In Vivo* Extrapolation (IVIVE)

Cryopreserved hepatocyte suspension Shibata *et al.* (2002)



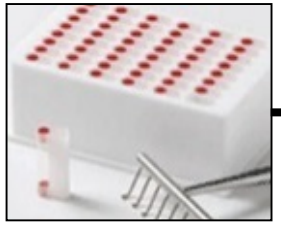
Human Hepatocytes (10 donor pool)



Hepatic Clearance

Rapid Equilibrium Dialysis (RED) Waters *et al.* (2008)

Human Plasma (6 donor pool)



Plasma Protein Binding

In Vitro – In Vivo Extrapolation

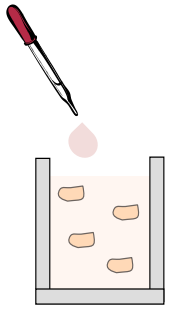
Plasma Concentrations

- **Most chemicals do not have TK data** – we use *in vitro* HTTK methods adapted from pharma to fill gaps
- In drug development, HTTK methods allow IVIVE to estimate therapeutic doses for clinical studies – predicted concentrations are typically on the order of values measured in clinical trials (Wang, 2010)

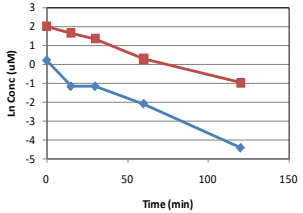
Figure from Barbara Wetmore

High-Throughput Toxicokinetics (HTTK) for *In Vitro-In Vivo* Extrapolation (IVIVE)

Cryopreserved hepatocyte suspension Shibata *et al.* (2002)



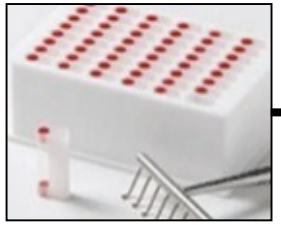
Human Hepatocytes (10 donor pool)



Hepatic Clearance

Rapid Equilibrium Dialysis (RED) Waters *et al.* (2008)

Human Plasma (6 donor pool)



Plasma Protein Binding

In Vitro – In Vivo Extrapolation

Plasma Concentrations

- **Most chemicals do not have TK data** – we use *in vitro* HTTK methods adapted from pharma to fill gaps
- In drug development, HTTK methods allow IVIVE to estimate therapeutic doses for clinical studies – predicted concentrations are typically on the order of values measured in clinical trials (Wang, 2010)

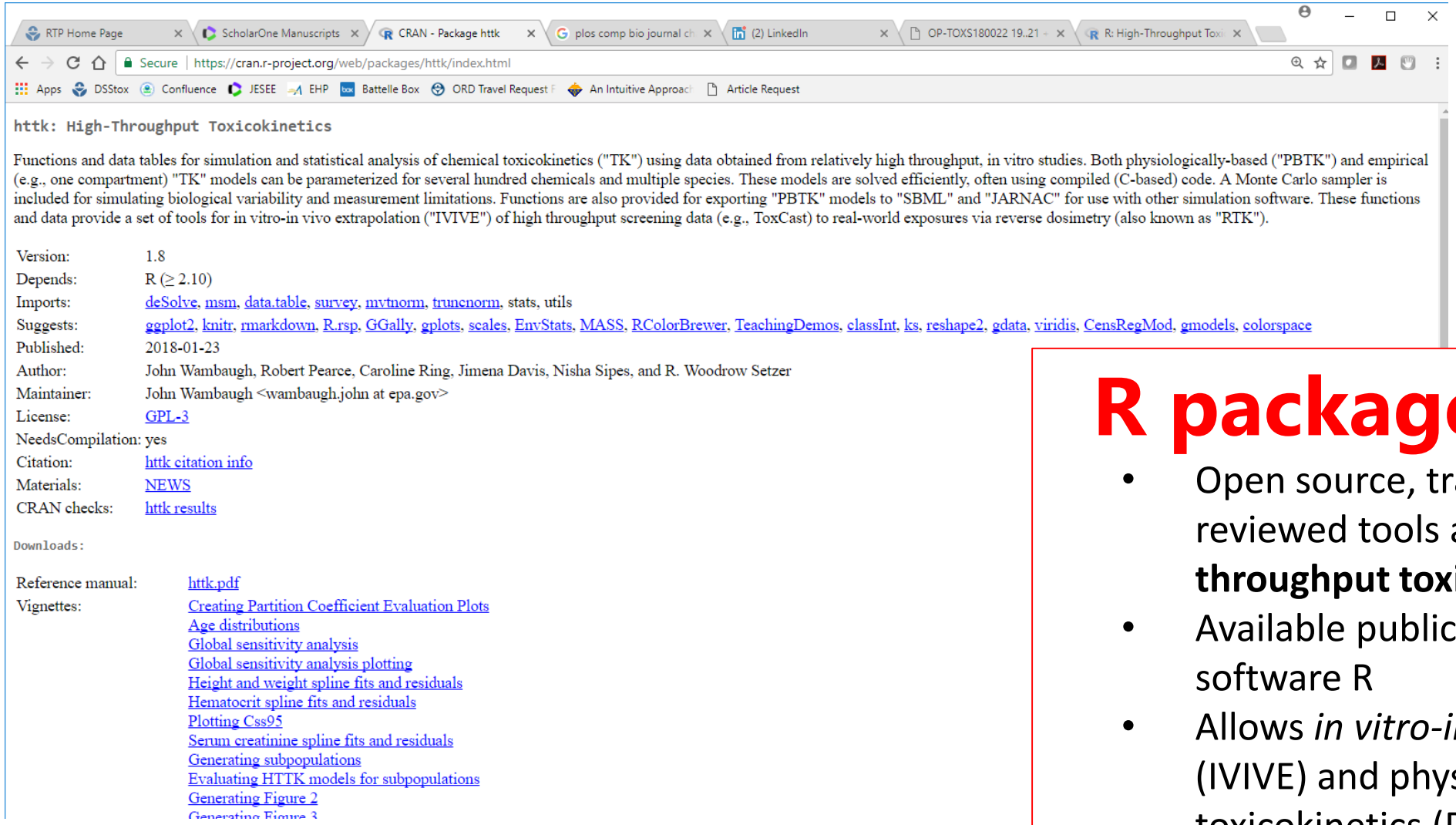
Environmental chemicals:

Rotroff et al. (2010) 35 chemicals
Wetmore et al. (2012) +204 chemicals
Wetmore et al. (2015) +163 chemicals
Wambaugh et al. (in prep.) + ~300 chemicals

Figure from Barbara Wetmore

Open Source Tools and Data for HTTK

https://CRAN.R-project.org/package=httk



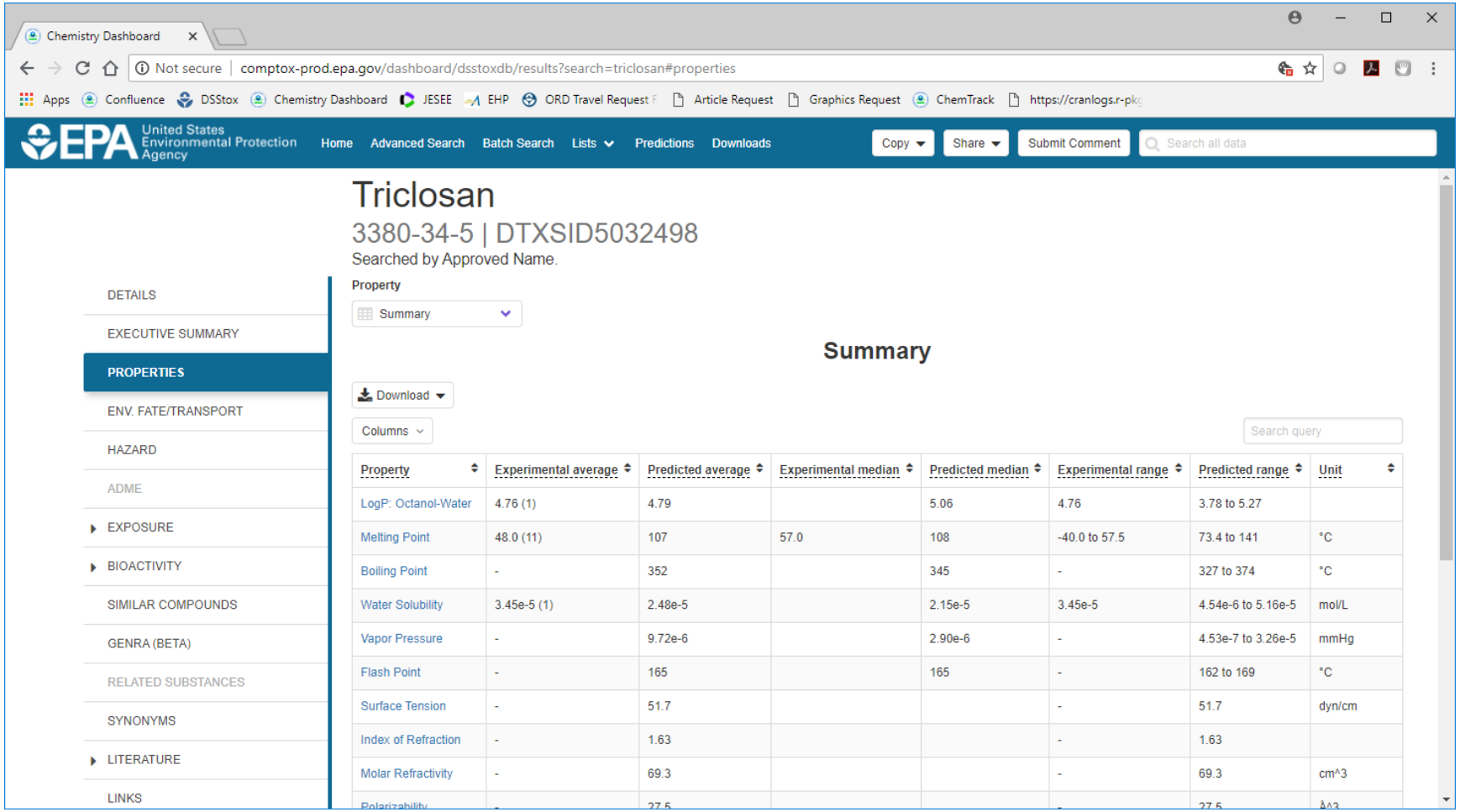
R package "httk"

- Open source, transparent, and peer-reviewed tools and data for **high throughput toxicokinetics (httk)**
- Available publicly for free statistical software R
- Allows *in vitro-in vivo* extrapolation (IVIVE) and physiologically-base toxicokinetics (PBTK)

Physicochemical Properties

- Measured and predicted physicochemical properties are available
  - OPEn structure–activity/property Relationship App (OPERA) by Mansouri, et al. (2018)

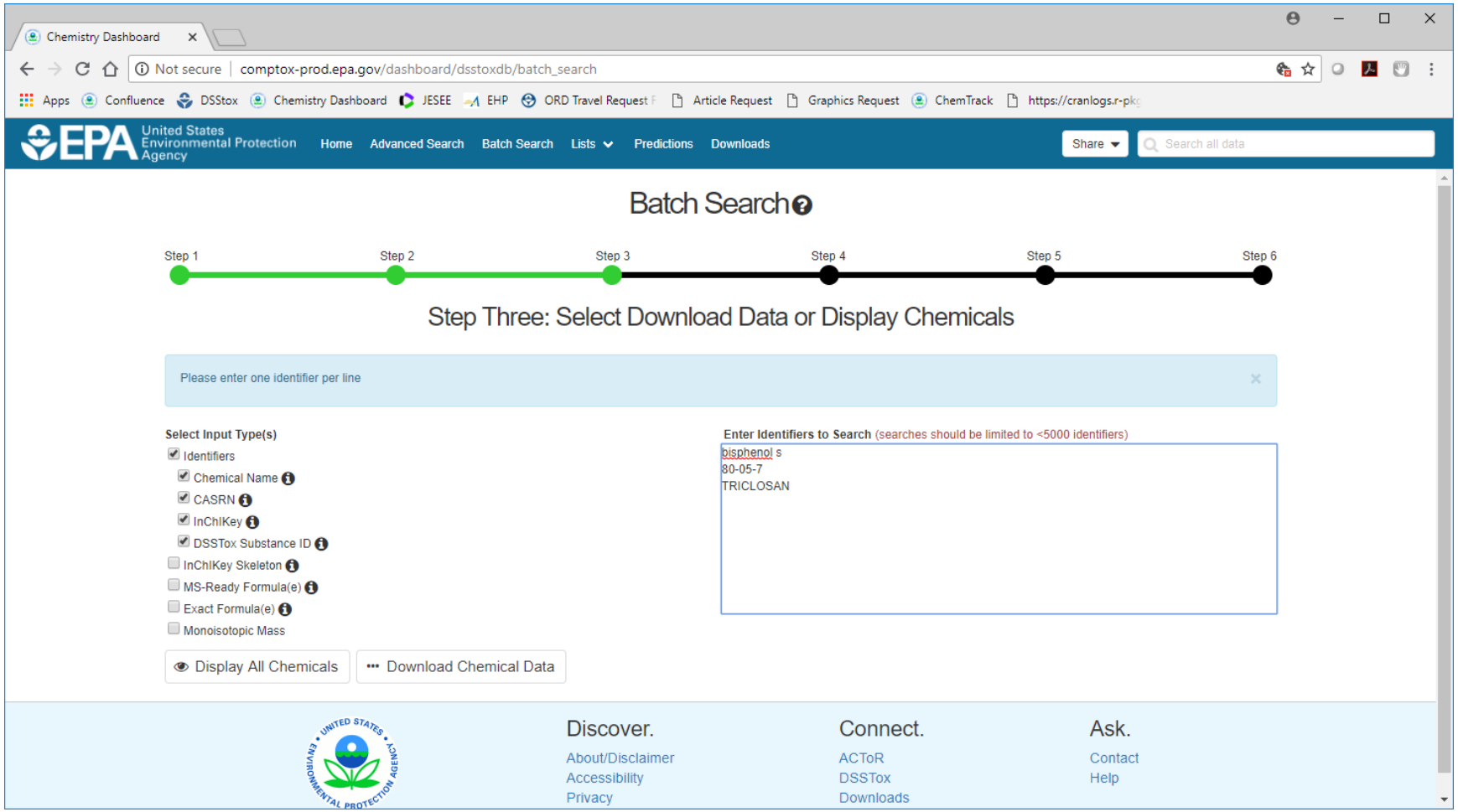
https://comptox.epa.gov/dashboard/



Bulk Download of Data

- Can download databases and spreadsheets of data using Batch Search

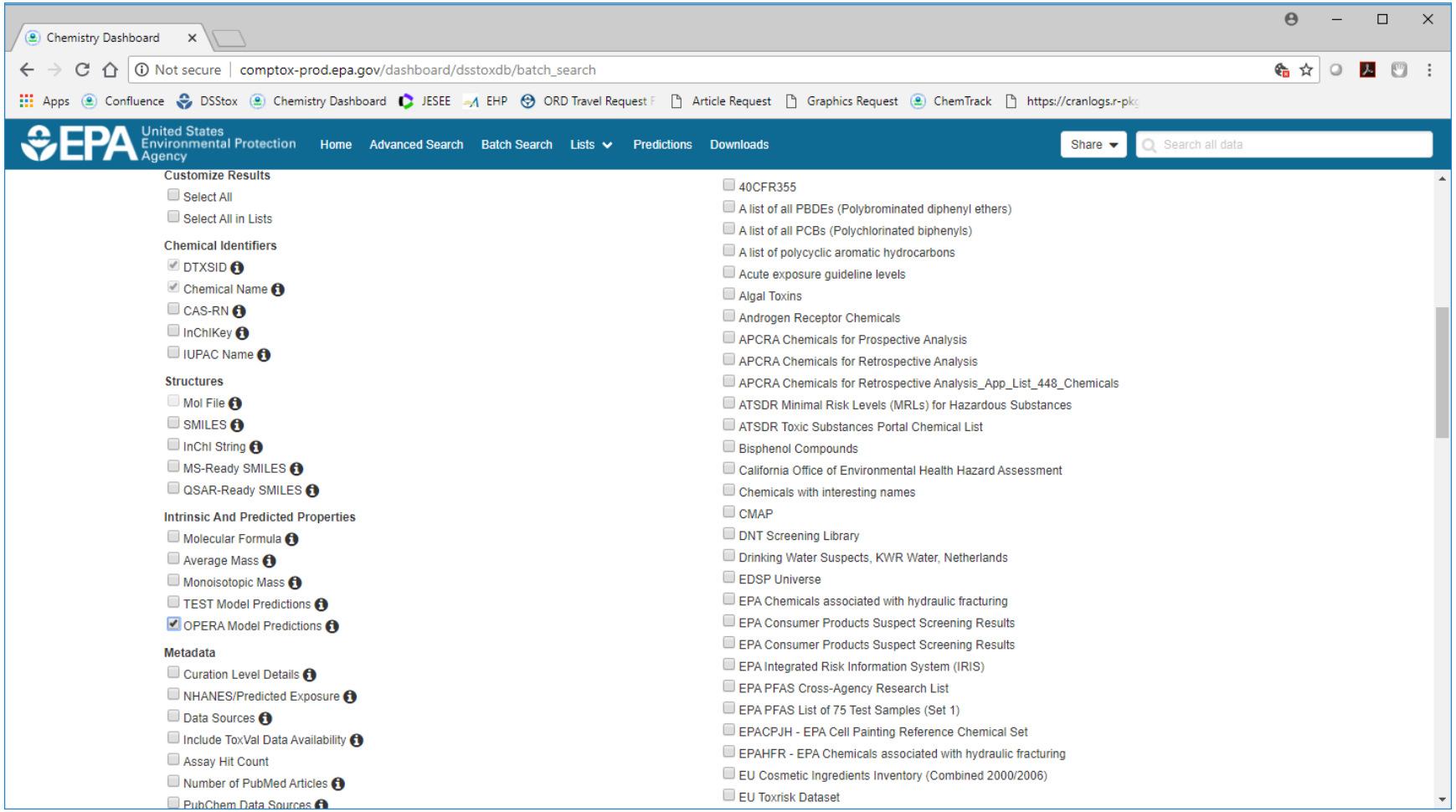
https://comptox.epa.gov/dashboard/



Bulk Download of Data

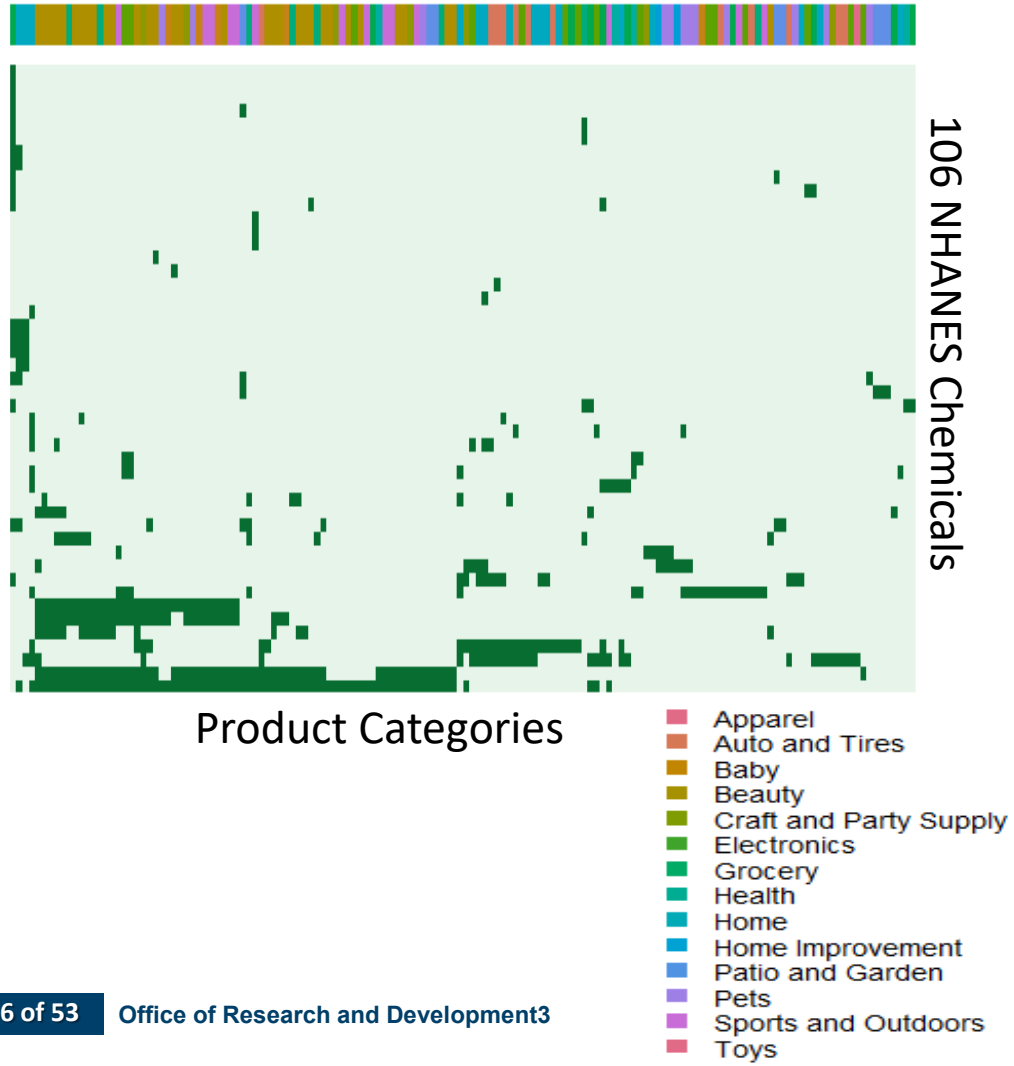
- Can download databases and spreadsheets of data using Batch Search

https://comptox.epa.gov/dashboard/

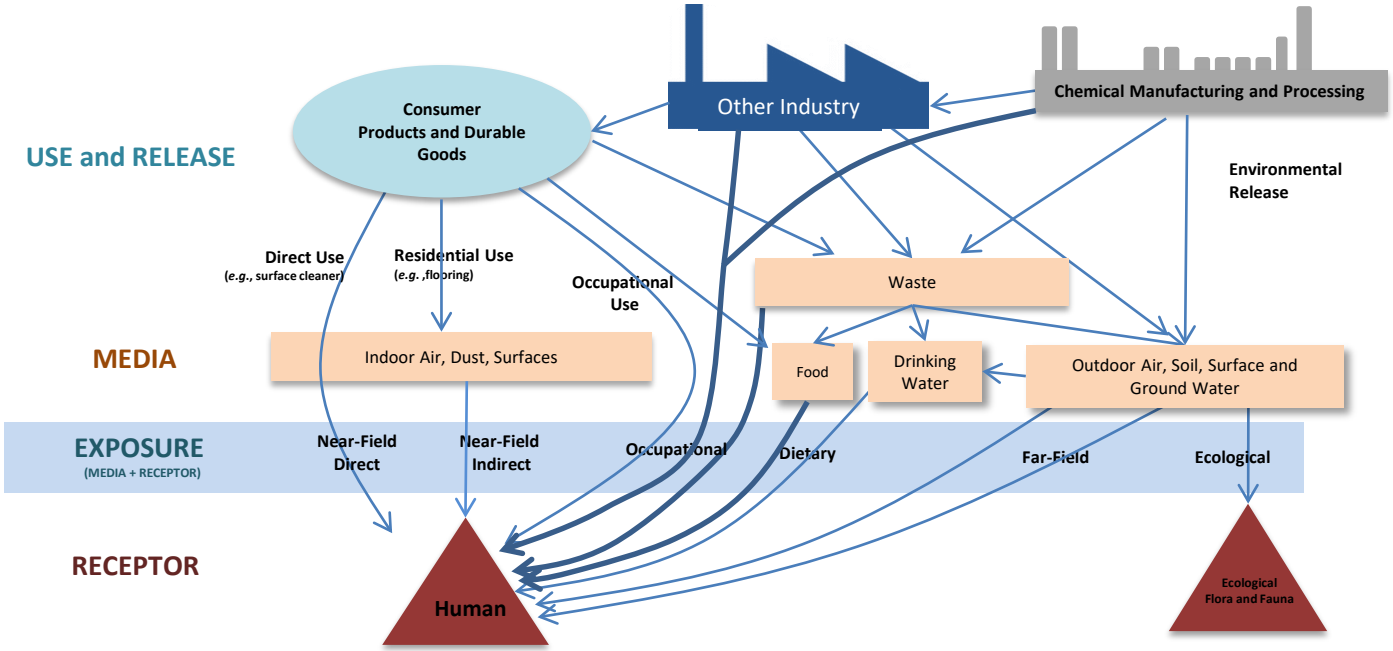


Chemical Sources

>2000 chemicals with Material Safety Data Sheets (MSDS) in Goldsmith *et al.*, 2014



- Near field sources have been known to be important at least since 1987 – see Wallace, *et al.*
- *Hard to know what chemicals are in which materials*
- Dashboard provides this information, and will be addressed in depth in subsequent lectures



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

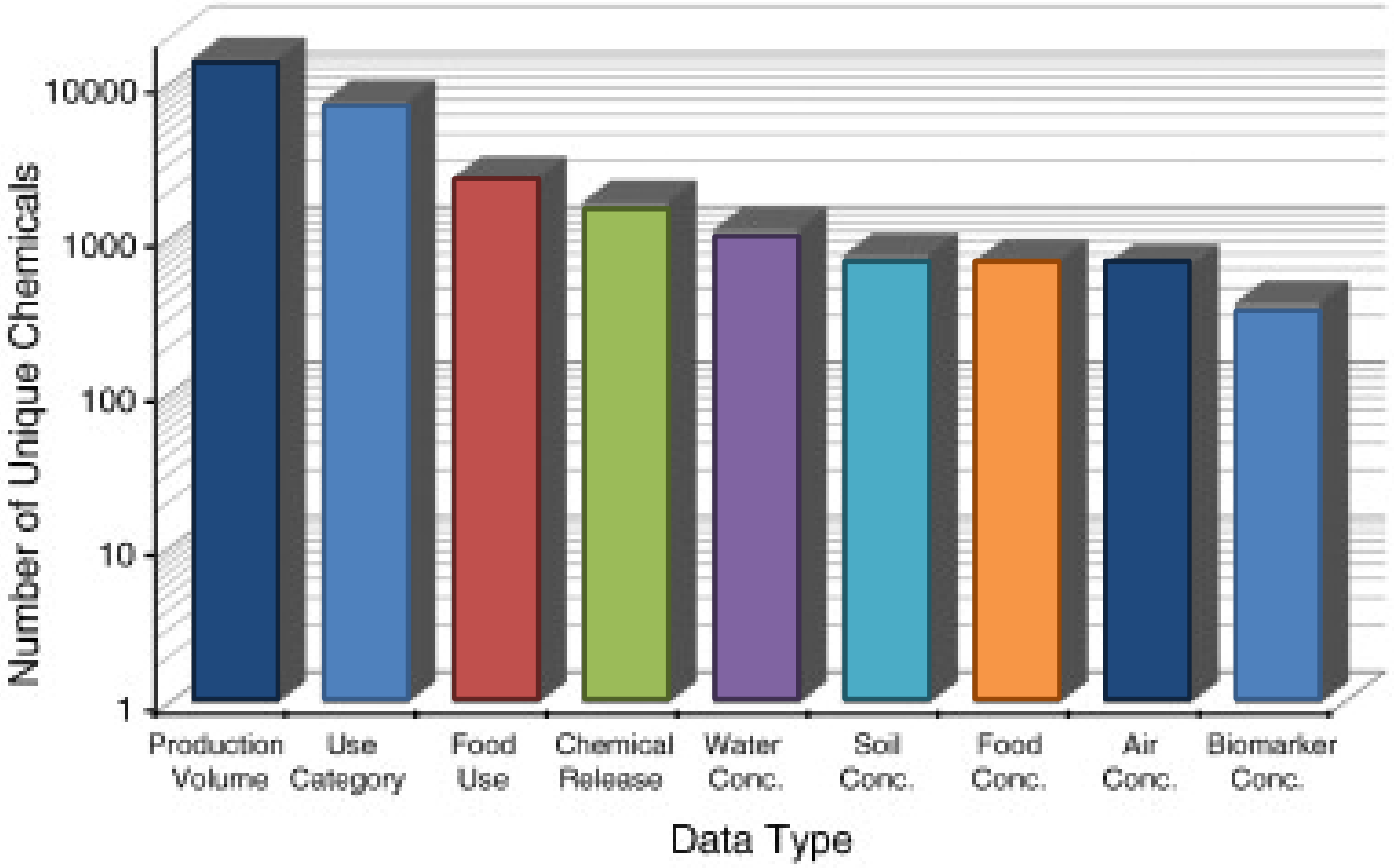
[¶]School of Public Health, University of California, Berkeley, California 94720, United States

Supporting Information

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

Limited Available Data for Exposure Estimation

Most chemicals lack public exposure-related data beyond production volume (Egeghy et al., 2012)



Can we develop new tools to generate the exposure information we need?

Decision Trees

- Decision trees are a useful tool for making predictions of how something should be classified
- Unfortunately, they are unstable Dietterich (2000)

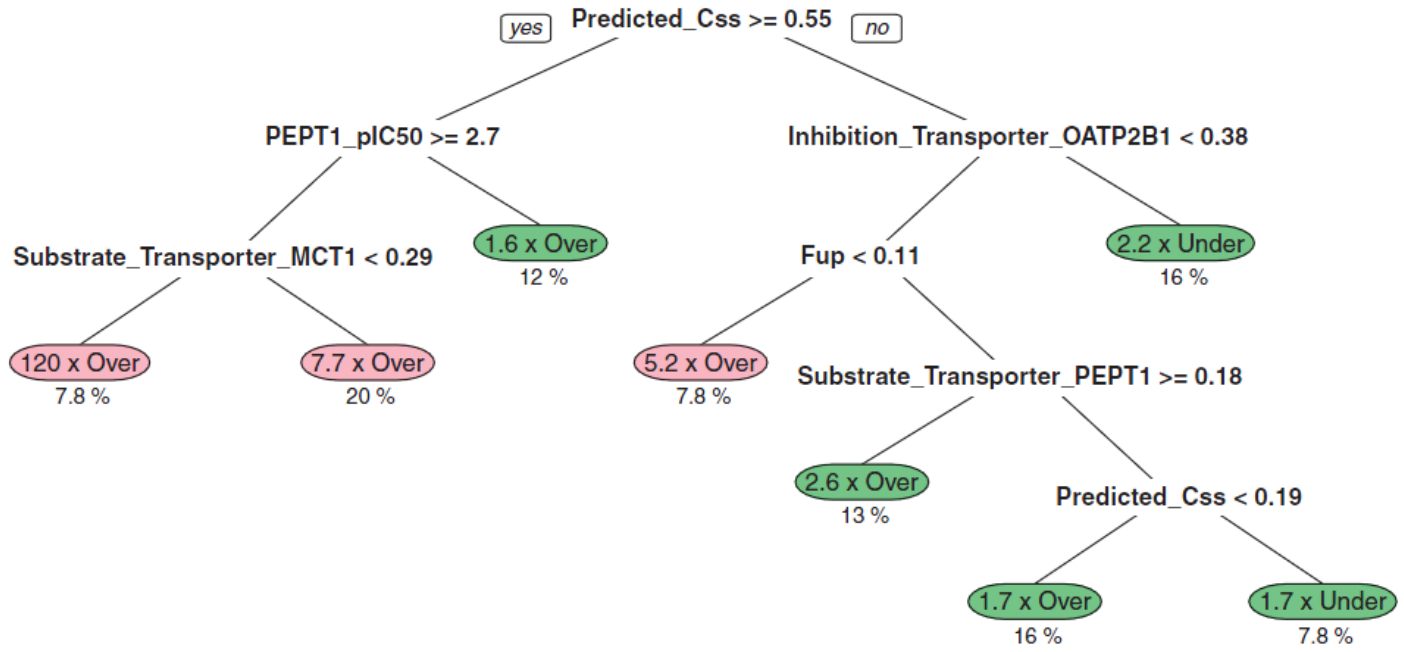


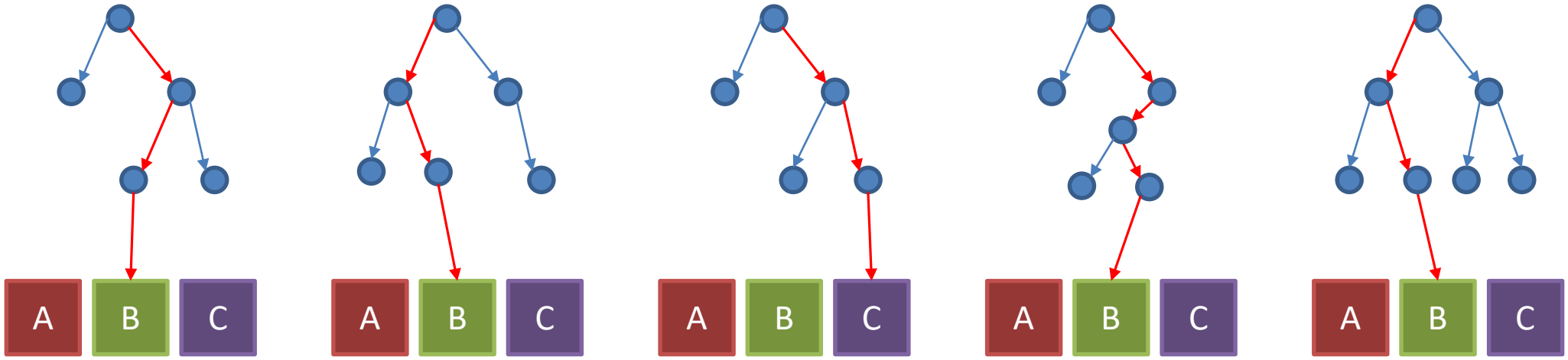
FIG. 5. A recursive partitioning regression tree was used to classify the discrepancy between the C_{ss} predicted from *in vitro* data and the *in vivo* C_{ss} (Obach *et al.*, 2008; Wetmore *et al.*, 2012). Each "leaf" of the tree shows a group of chemicals for which HTTK either overestimates C_{ss} (making conservative predictions) or underestimates C_{ss}. For all but 3 groups, the predictions are on the order of the observed C_{ss} (approximately within a factor of 3.2× greater or lesser). For the other 3 groups, the C_{ss} is 5.2×, 7.7×, and 120× overestimated. The dashed line indicates the identity (perfect predictor) line.

- In Wambaugh et al. (2015) various chemical properties, including Fraction unbound in plasma (F_{up}) and transporter affinities were used to predict whether C_{ss} would be over or underestimated

Ensemble Predictions

- “Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions.” Dietterich (2000)

- Every model gets a “vote” – can think of this probabilistically



Four votes for “B” – 80%

Bootstrap AGGregatING: Bagging

- How do we get multiple decision trees? We use the method of Random Forests (Brieman, 2001)
- Construct multiple training sets that are subsets of the available data
- The models corresponding to each data subset each get a vote
- Estimate the error of each tree using the data not in the subset
  - Out of bag (OOB) error

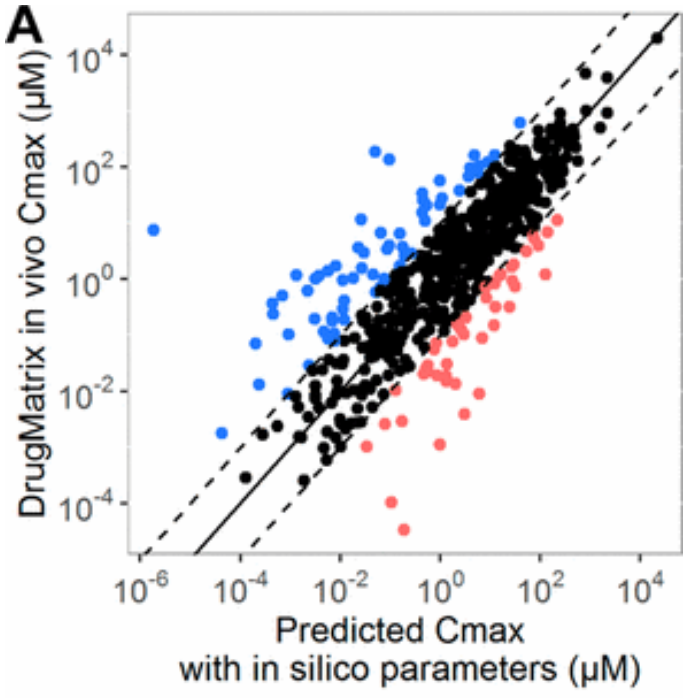
Image from Wikipedia article on “cross-validation”

Descriptor Subsets

- In Random Forests, each decision point only considers a random subset of the available predictors (Ho, 1995)
- Allows for large predictors sets, such as chemical structure chemotypes (thousands of structure features)
- Because only a subset of predictors are evaluated for each branching point in each tree, can evaluate how well do trees that include predictor X perform relative to trees that don't include predictor X
- This is a measure of predictor importance (Archer et al., 2008)
- Can tell us which parameters drive the model
- Need to be careful about correlated variables

Descriptor Subsets

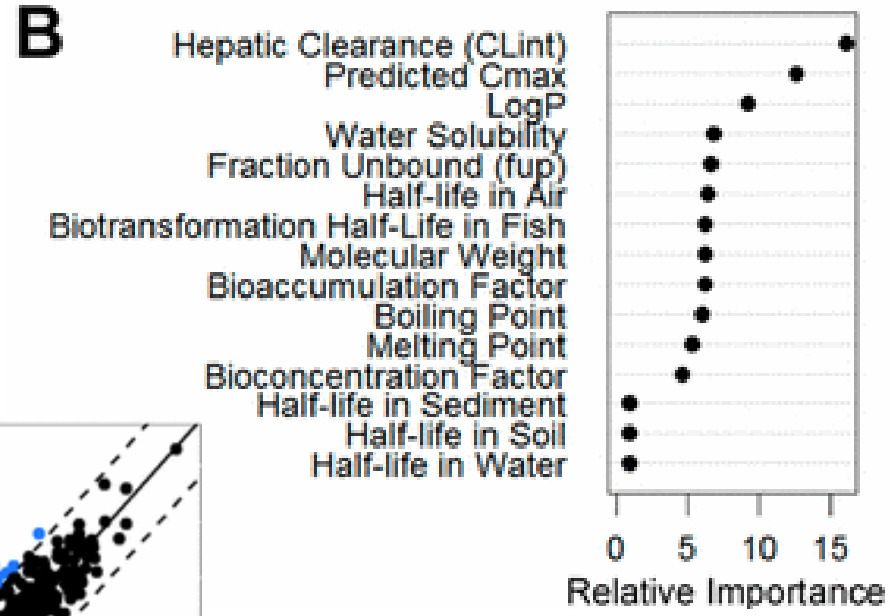
- In Random Forests, each decision point only considers a random subset of the available predictors (Ho, 1995)
- Allows for large predictors sets, such as chemical structure chemotypes (thousands of structure features)
- Because only a subset of predictors are evaluated for each branching point in each tree, can evaluate how well do trees that include predictor X perform relative to trees that don't include predictor X
- This is a measure of predictor importance (Archer et al., 2008)
- Can tell us which parameters drive the model
- Need to be careful about correlated variables

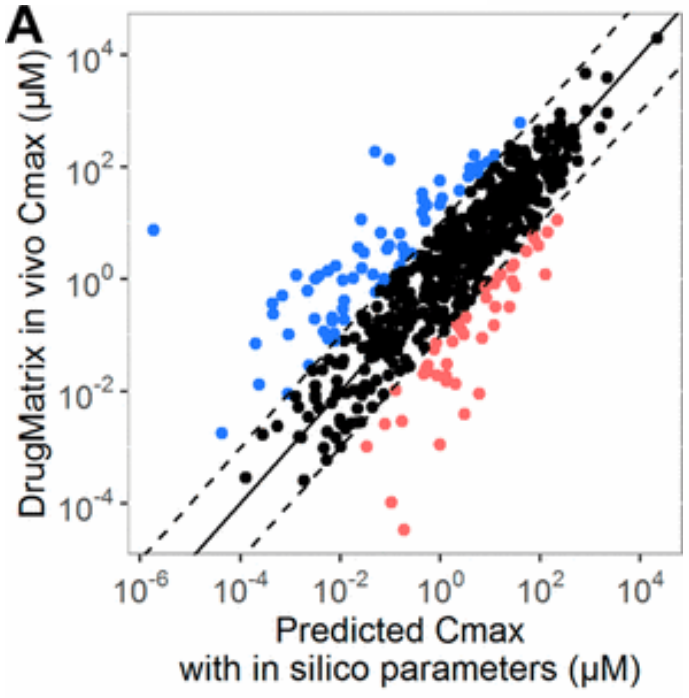


- In Sipes et al. (2017), chemical concentrations were predicted in order to compare in vitro bioactivity with exposure data

Descriptor Subsets

- In Random Forests, each decision point only considers a random subset of the available predictors (Ho, 1995)
- Allows for large predictors sets, such as chemical structure chemotypes (thousands of structure features)
- Because only a subset of predictors are evaluated for each branching point in each tree, can evaluate how well do trees that include predictor X perform relative to trees that don't include predictor X
- This is a measure of predictor importance (Archer et al., 2008)
- Can tell us which parameters drive the model
- Need to be careful about correlated variables





- In Sipes et al. (2017), chemical concentrations were predicted in order to compare in vitro bioactivity with exposure data

Model Performance

- In addition to OOB (out of bag) error rate, we can look at how well the average prediction of the models does in classifying the data: true positives (**TP**), false positives (**FP**), true negatives (**TN**), false negatives (**FN**)
- Sensitivity = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Want to do both, balanced accuracy = (Sensitivity + Specificity)/2
- Can use the option “sampsize” in randomForest R package to make sure that the training sets are balanced

	Actual Positives	Actual Negatives
Predicted Positives	TP	FP
Predicted Negatives	FN	TN

Confusion Matrix

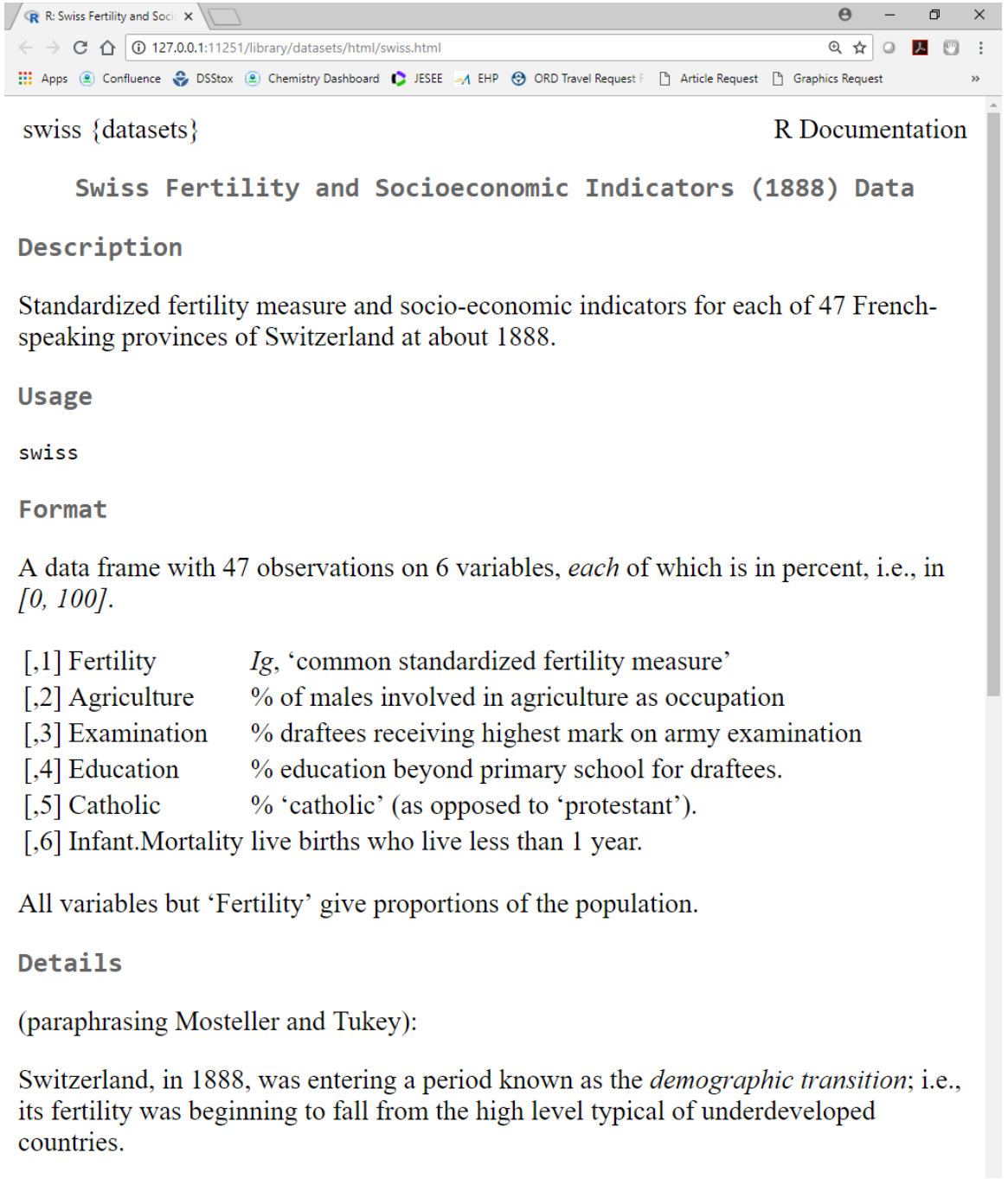
R Code

- Liaw and Wiener's R package "randomForest" ported the original Random Forests Fortran by Leo Breiman and Adele Cutler into R
- Can do both classification and regression (we have not discussed regression much here)

Actual R code:

```
#Built in 1888 Swiss Fertility Data

help(swiss) # Display data information
```



R: Swiss Fertility and Soci

127.0.0.1:11251/library/datasets/html/swiss.html

Apps Confluence DSStox Chemistry Dashboard JESEE EHP ORD Travel Request Article Request Graphics Request

swiss {datasets} R Documentation

Swiss Fertility and Socioeconomic Indicators (1888) Data

Description

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Usage

```
swiss
```

Format

A data frame with 47 observations on 6 variables, *each* of which is in percent, i.e., in *[0, 100]*.

[,1] Fertility	*Ig*, 'common standardized fertility measure'
[,2] Agriculture	% of males involved in agriculture as occupation
[,3] Examination	% draftees receiving highest mark on army examination
[,4] Education	% education beyond primary school for draftees.
[,5] Catholic	% 'catholic' (as opposed to 'protestant').
[,6] Infant.Mortality	live births who live less than 1 year.

All variables but 'Fertility' give proportions of the population.

Details

(paraphrasing Mosteller and Tukey):

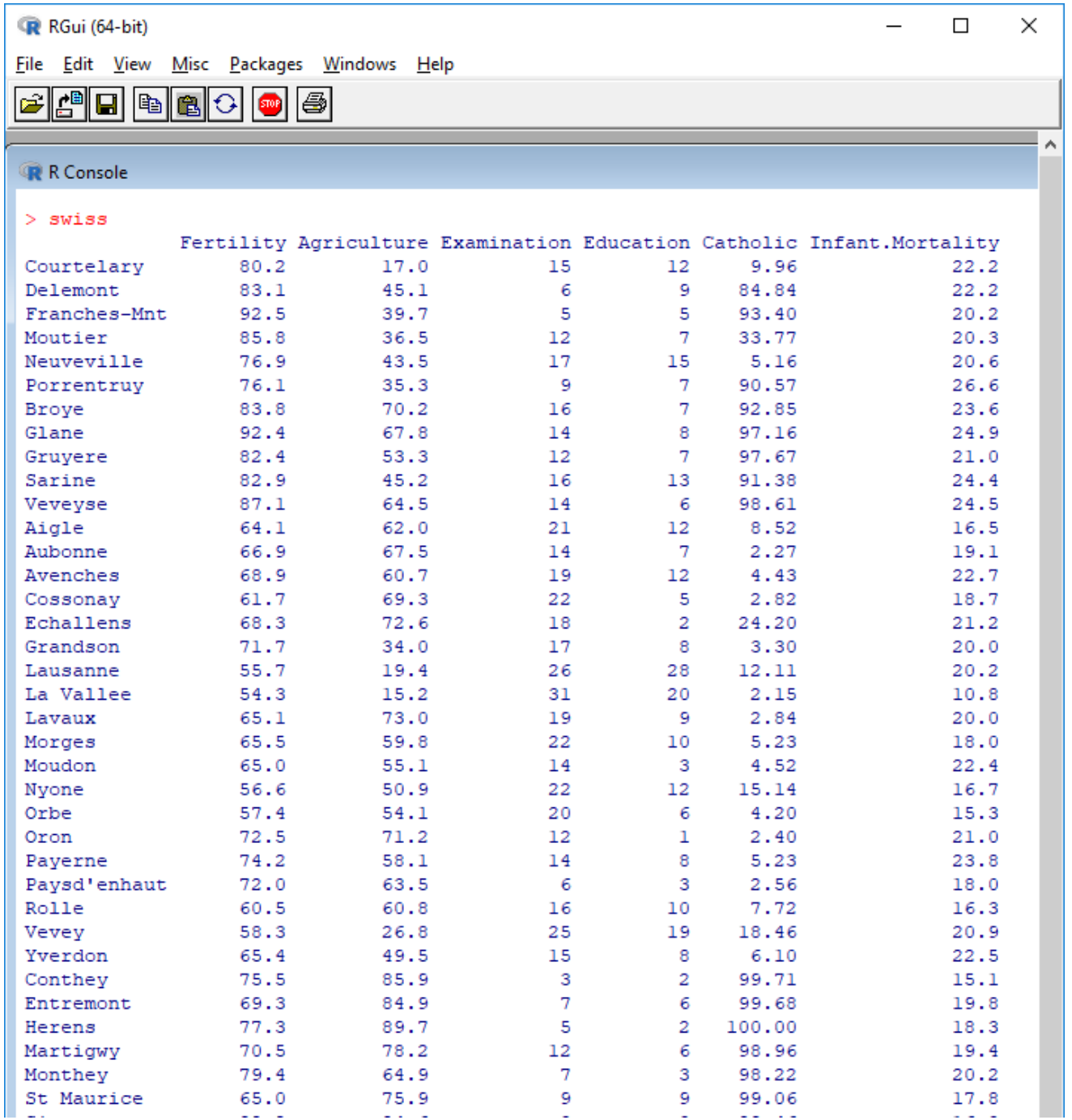
Switzerland, in 1888, was entering a period known as the *demographic transition*; i.e., its fertility was beginning to fall from the high level typical of underdeveloped countries.

R Code

- Liaw and Wiener's R package "randomForest" ported the original Random Forests Fortran by Leo Breiman and Adele Cutler into R
- Can do both classification and regression (we have not discussed regression much here)

Actual R code:

```
swiss #display the data
```



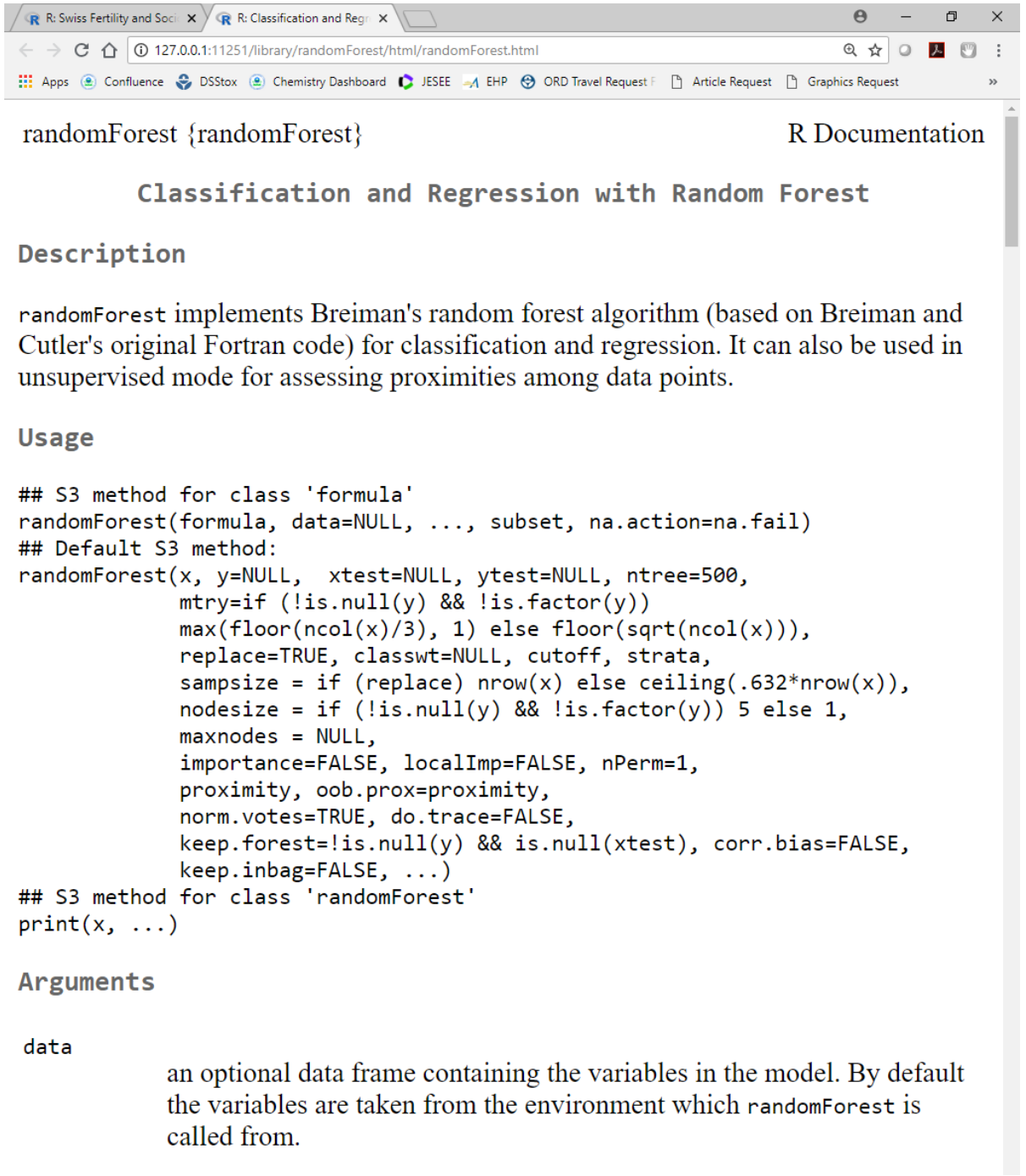
	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4
Nyone	56.6	50.9	22	12	15.14	16.7
Orbe	57.4	54.1	20	6	4.20	15.3
Oron	72.5	71.2	12	1	2.40	21.0
Payerne	74.2	58.1	14	8	5.23	23.8
Paysd'enhaut	72.0	63.5	6	3	2.56	18.0
Rolle	60.5	60.8	16	10	7.72	16.3
Vevey	58.3	26.8	25	19	18.46	20.9
Yverdon	65.4	49.5	15	8	6.10	22.5
Conthey	75.5	85.9	3	2	99.71	15.1
Entremont	69.3	84.9	7	6	99.68	19.8
Herens	77.3	89.7	5	2	100.00	18.3
Martigwy	70.5	78.2	12	6	98.96	19.4
Monthey	79.4	64.9	7	3	98.22	20.2
St Maurice	65.0	75.9	9	9	99.06	17.8

R Code

- Liaw and Wiener's R package "randomForest" ported the original Random Forests Fortran by Leo Breiman and Adele Cutler into R
- Can do both classification and regression (we have not discussed regression much here)

Actual R code:

```
library(randomForest) #load Random Forest
help(randomForest) # get help
```



R Code

- Liaw and Wiener's R package "randomForest" ported the original Random Forests Fortran by Leo Breiman and Adele Cutler into R
- Can do both classification and regression (we have not discussed regression much here)

Actual R code:

```
# Set regions with below median fertility to
false, above median to true:
swiss$Fertility <-
swiss$Fertility>median(swiss$Fertility)
# Turn it into a "factor" which is how R
describes a classifcation
swiss$Fertility <-
as.factor(swiss$Fertility)
# Build a random forests model:
mdl <- randomForest(Fertility~.,data=swiss)
```

```
Call:
 randomForest(formula = Fertility ~ ., data = swiss)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 25.53%
Confusion matrix:
      FALSE TRUE class.error
FALSE    18    6   0.2500000
TRUE      6   17   0.2608696
> |
```

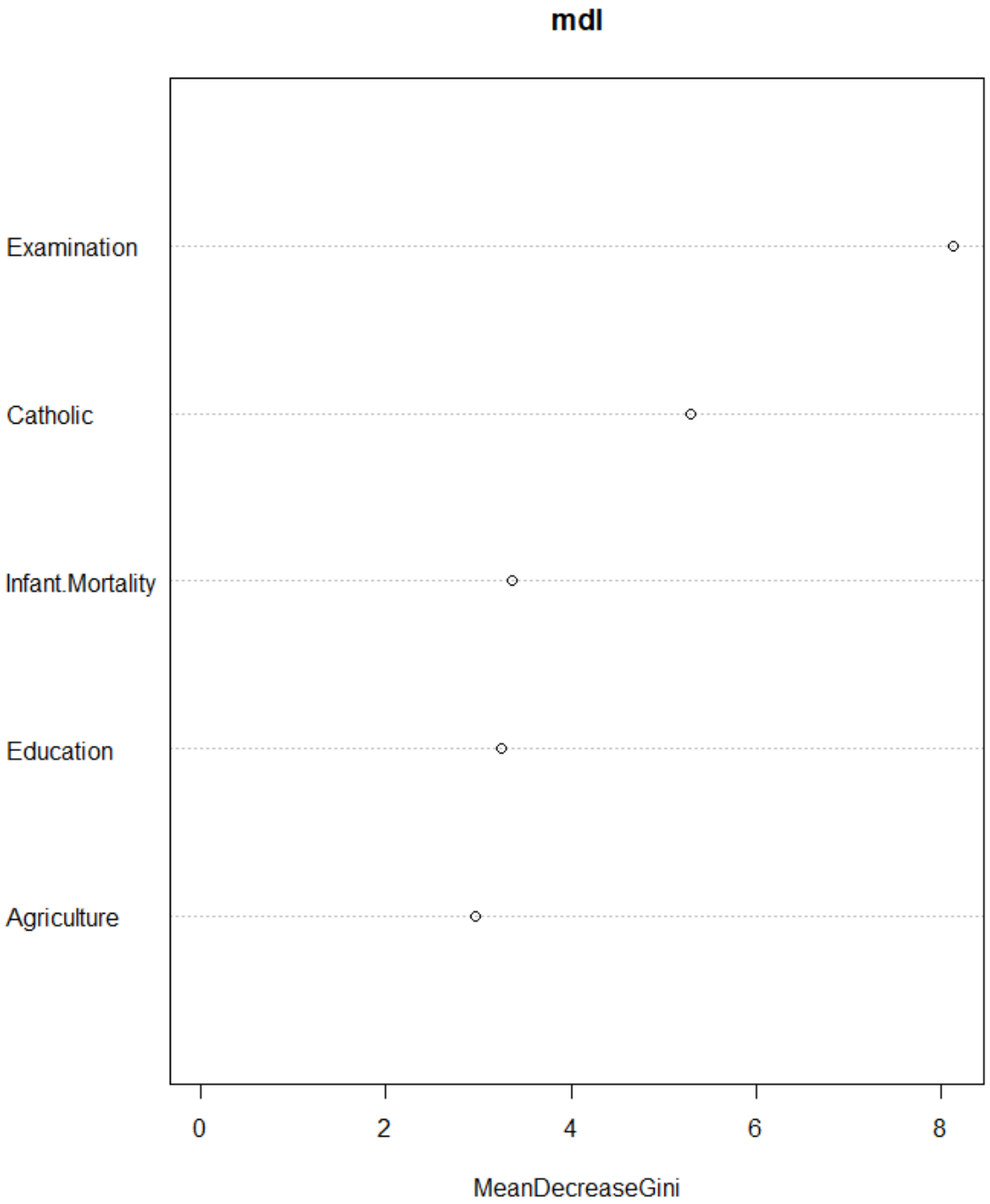
R Code

- Liaw and Wiener's R package "randomForest" ported the original Random Forests Fortran by Leo Breiman and Adele Cutler into R
- Can do both classification and regression (we have not discussed regression much here)

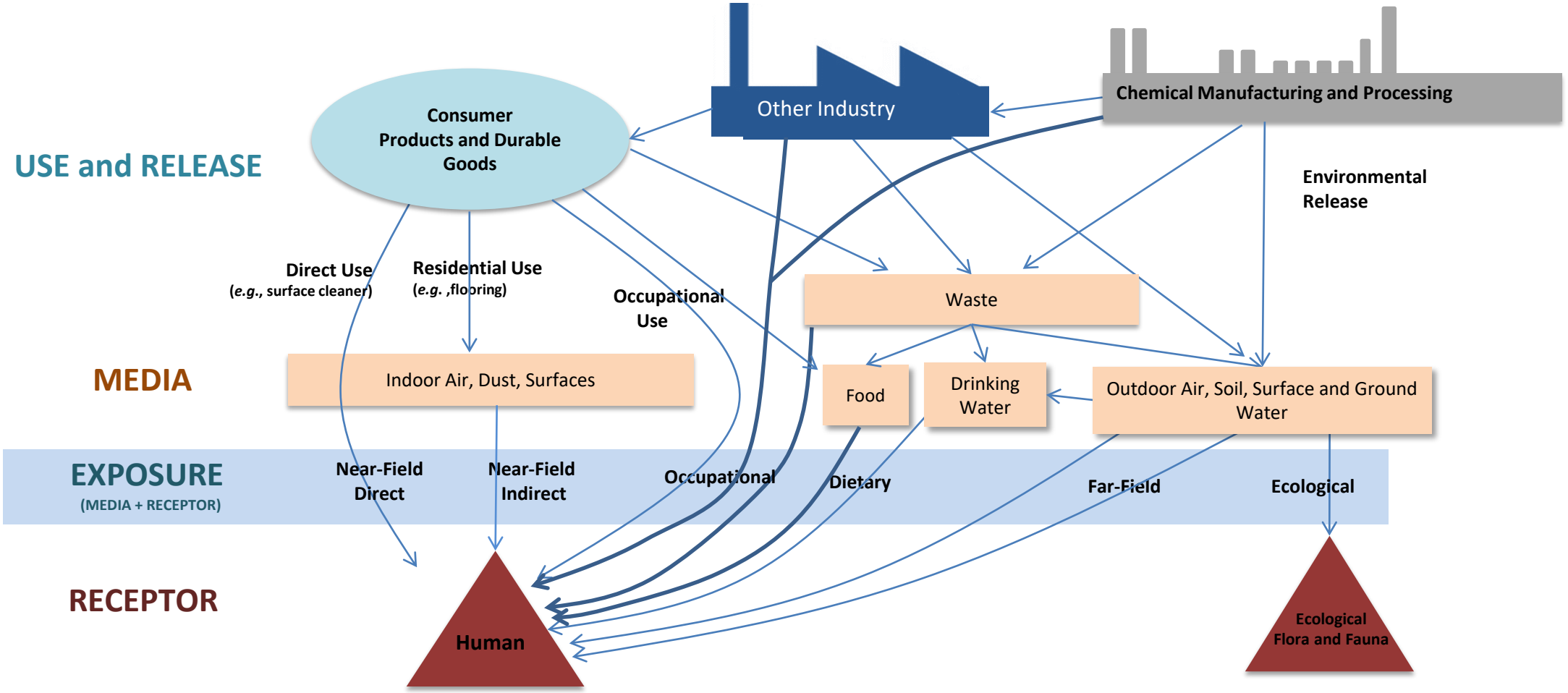
Actual R code:

```
varImpPlot(mdl) # Variable importance
```

Examination: percent of draftees receiving highest mark on army examination



Understanding Exposure is a Systems Problem



- **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)

Finding the Right Data

- Data curation is the rate limiting step for application of Random Forests
- Need a set of positive **and** negative examples with descriptors
- Thanks to various software tools and the speed of modern computers, once you have the data building the models is relatively easy

We used 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

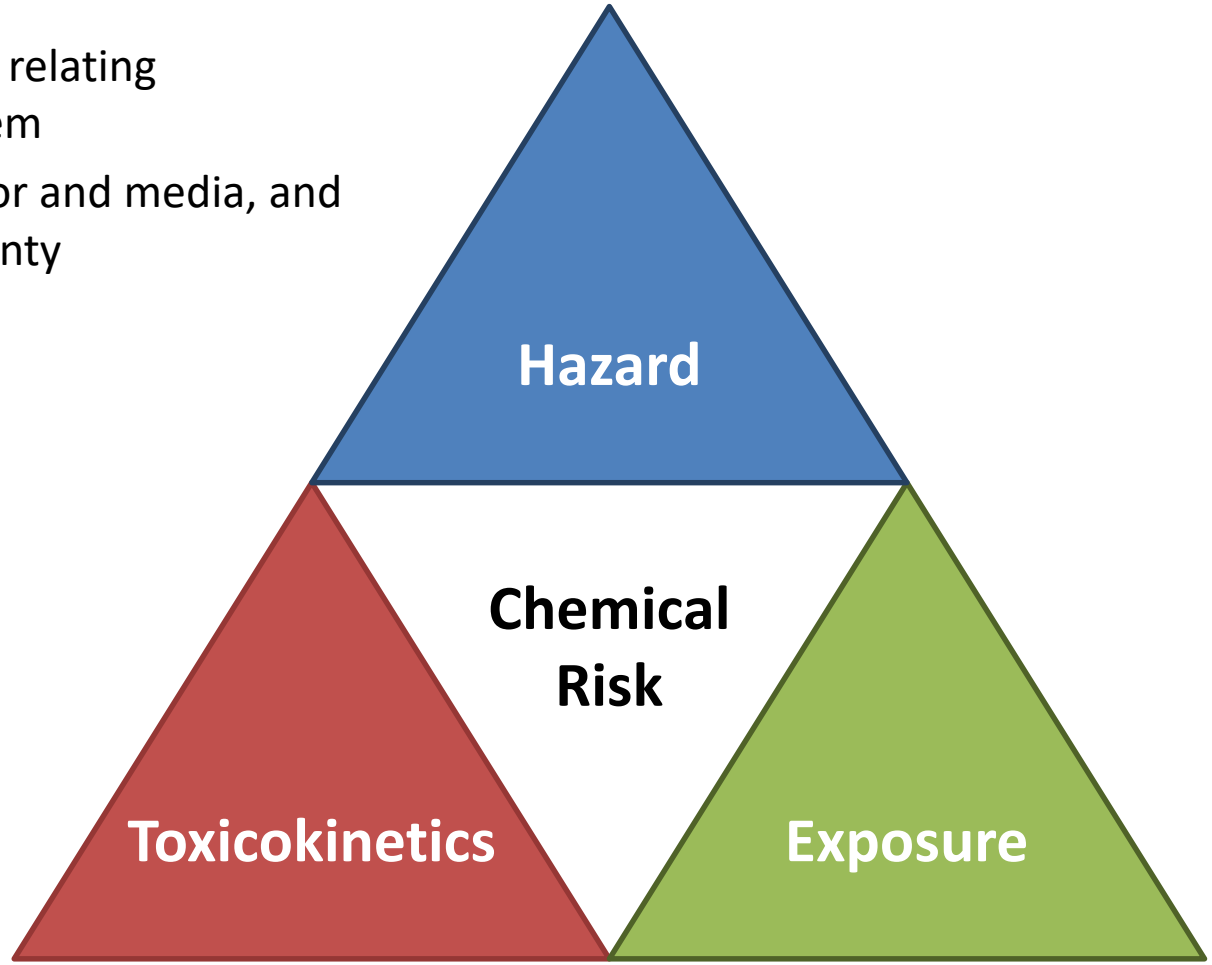
Ring et al., submitted

https://comptox.epa.gov/dashboard

- There are low levels of thousands of chemicals in commerce, relating exposures and health effects is an important unsolved problem
- The exposure pathway is the actual interaction of the receptor and media, and this event is often confounded by various sources of uncertainty

EPA's CompTox dashboard (Williams et al, 2017) can help you:

- Identify chemicals
- Find toxicity data
- Find lists of chemicals
- Find metabolites
- Identify products
- Find toxicokinetic information
- Get physicochemical properties
- Batch download data



Chemical Safety for Sustainability (CSS) Research Program

Rapid Exposure and Dosimetry (RED) Project

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

NRMRL
Xiaoyu Liu

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

NERL
Cody Addington*
Craig Barber
Namdi Brandon*
Peter Egeghy
Hongtai Huang*
Kristin Isaacs
Ashley Jackson*
Charles Lowe*
Dawn Mills*
Seth Newton
Katherine Phillips
Paul Price
Jeanette Reyes*
Randolph Singh*
Jon Sobus
John Streicher*
Mark Strynar
Mike Tornero-Velez
Elin Ulrich
Dan Vallero
Barbara Wetmore

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

***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
University of California, Davis
Deborah Bennett
University of Michigan
Olivier Jolliet
University of Texas, Arlington
Hyeong-Moo Shin

EPA United States Environmental Protection Agency

References

- Filer, Dayne L., et al. "tcpl: The ToxCast Pipeline for High-Throughput Screening Data." Bioinformatics (2016): btw680.
- Howgate, E., et al. “Prediction of in vivo drug clearance from in vitro data. I: impact of inter-individual variability” Xenobiotica 2006;36:473-497
- Israili and Dayton “Human Alpha-1-Glycoprotein and Its Interactions with Drugs” Drug metabolism reviews 2001;33:161-235
- Jamei, et al. “The Simcyp® population-based ADME simulator.” Expert opinion on drug metabolism & toxicology 2009b;5:211-223
- Johnson, et al. “Prediction of the clearance of eleven drugs and associated variability in neonates, infants and children.” Clinical pharmacokinetics (2006)
- Judson, R. S., et al., (2010) “In Vitro Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. Environmental Health Perspectives 118(4), 485-492.
- McNally, et al., “PopGen: a virtual human population generator.” Toxicology *2014)
- 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, et al. “httk: R Package for High-Throughput Toxicokinetics.” Journal of Statistical Software, (2017a)
- Pearce. Robert, et al. ““Evaluation and Calibration of High-Throughput Predictions of Chemical Distribution to Tissues,” Journal of Pharmacokinetics and Pharmacodynamics (2017b)
- Price et al., “Instructions for Use of Software Physiological Parameters for PBPK Modeling Version 1.3 (P3MTM 1.3).” 2003
- Ring , Caroline, et al., “Identifying populations sensitive to environmental chemicals by simulating toxicokinetic variability”, Environment International, *in press*
- Rotroff, Daniel, et al., (2010) “Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening.” Tox. Sciences 117(2), 348-58
- Routledge, P., “The plasma protein binding of basic drugs. British journal of clinical pharmacology 1986;22:499-506
- Sipes, Nisha, et al. “An Intuitive Approach for Predicting Potential Human Health Risk with the Tox21 10k Library”, Environmental Science and Technology (2017)
- Wambaugh, John F., et al. "High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals." *Env. science & technology* (2014).
- Wambaugh, John F., et al. "Toxicokinetic triage for environmental chemicals." Toxicological Sciences (2015): kfv118.
- Wang, Y.-H. (2010). “Confidence Assessment of the Simcyp Time-Based Approach and a Static Mathematical Model in Predicting Clinical Drug-Drug Interactions for Mechanism-Based CYP3A Inhibitors.” Drug Metabolism and Disposition 38(7), 1094-1104
- Wetmore, Barbara A., et al. "Integration of dosimetry, exposure and high-throughput screening data in chemical toxicity assessment." *Tox. Sciences* (2012)
- 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.
- Yasuda, et al., “The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies.” Clinical Pharmacology & Therapeutics 2008;84:417-423
- Yoon, M., et al. (2014). “Evaluation of simple in vitro to in vivo extrapolation approaches for environmental compounds.” Toxicology in Vitro 28(2), 164-170,