http://www.orcid.org/0000-0002-2668-4821

Progress in Using Big Data in Chemical Toxicity Research at the National Center for Computational Toxicology

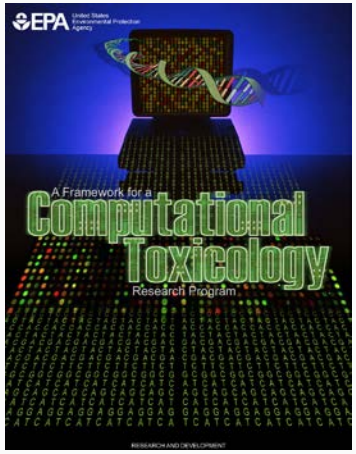
Antony Williams

U.S. Environmental Protection Agency, RTP, NC

June 1st, 2017

Proctor and Gamble, Cincinnati, OH

National Center for Computational Toxicology

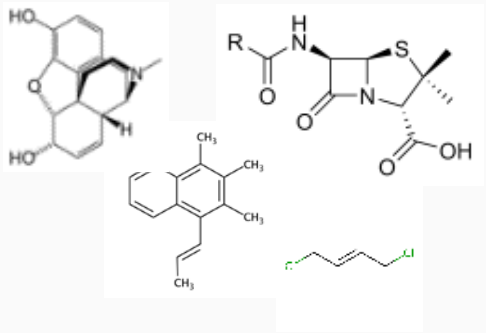


- National Center for Computational Toxicology (2005) established to integrate:
  - High-throughput and high-content technologies
  - Modern molecular biology
  - Data mining and statistical modeling
  - Computational biology and chemistry

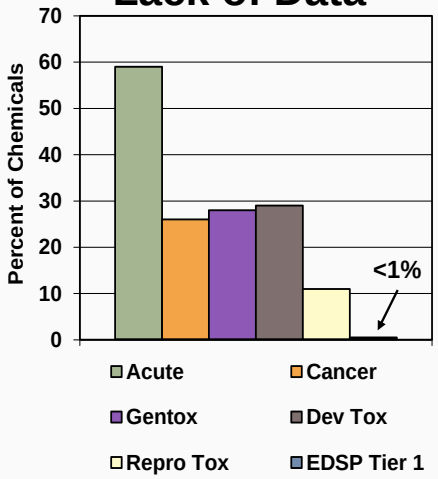
- Currently staffed by ~60 employees as part of EPA's Office of Research and Development
- Home of ToxCast & ExpoCast research efforts
- Key partner in U.S. Tox21 federal consortium

Why we must innovate…

Number of Chemicals /Combinations

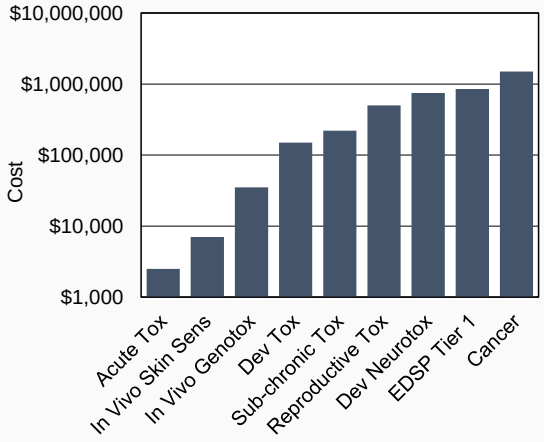


Lack of Data



Ethical Concerns

Economics



Toxic Substances Control List

- Inventory was initially published in 1979
- Second version, containing about 62k chemical substances, was published in 1982
- Continues to grow and now lists ~85k chemicals, about 15k are confidential business information

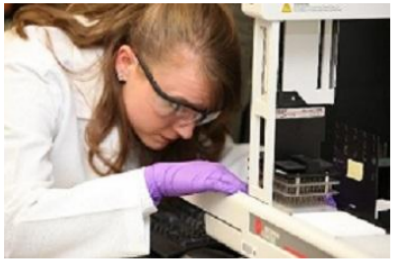
https://www.epa.gov/tsca-inventory/how-access-tsca-inventory

The need for data and derivative models and algorithms

- NCCT outputs: include a lot of data, models, algorithms and software applications
- We produce Open Data – we want people to interrogate it, learn from it, develop understanding

Toxicity Forecasting

Advancing the Next Generation of Chemical Evaluation



EPA needs rapid and efficient methods to prioritize, screen and evaluate thousands of chemicals. EPA's Toxicity Forecaster (ToxCast) generates data and predictive models on thousands of chemicals of interest to the EPA. ToxCast uses high-throughput screening methods and computational toxicology approaches to rank and prioritize chemicals. In fact, EPA's Endocrine Disruption Screening Program (EDSP) is working to use ToxCast to rank and prioritize chemicals.

- ToxCast has data on over 1,800 chemicals from a broad range of sources including industrial and consumer products, food additives, and potentially "green" chemicals that could be safer alternatives to existing chemicals.
- ToxCast screens chemicals in over 700 high-throughput assays that cover a range of high-

Downloadable Computational Toxicology Data

EPA's computational toxicology research efforts evaluate the potential health effects of thousands of chemicals. The process of evaluating potential health effects involves generating data that investigates the potential harm, or hazard of a chemical, the degree of exposure to chemicals as well as the unique chemical characteristics.

As part of EPA's commitment to share data, all of the computational toxicology data is publicly available for anyone to access and use.

High-throughput Screening Data

EPA researchers use rapid chemical screening (called high-throughput screening assays) to limit the number of laboratory animal tests while quickly and efficiently testing thousands of chemicals for potential health effects.

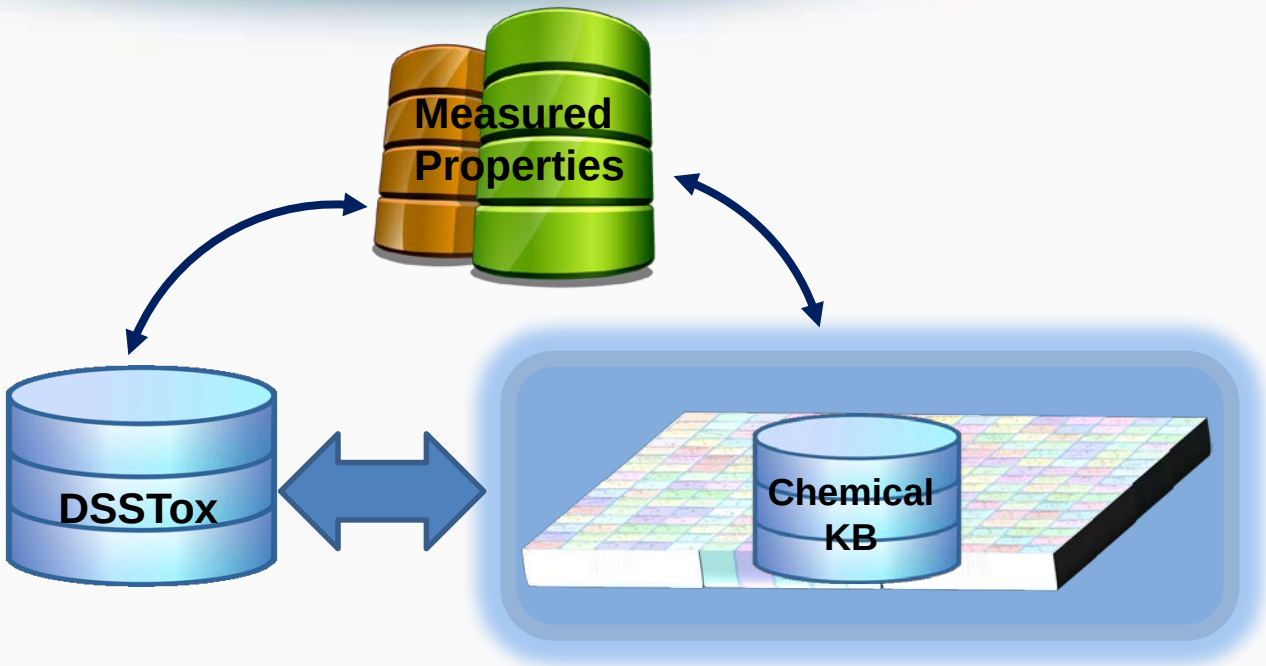
- ToxCast Data: High-throughput screening data on thousands of chemicals.

Rapid Exposure and Dose Data

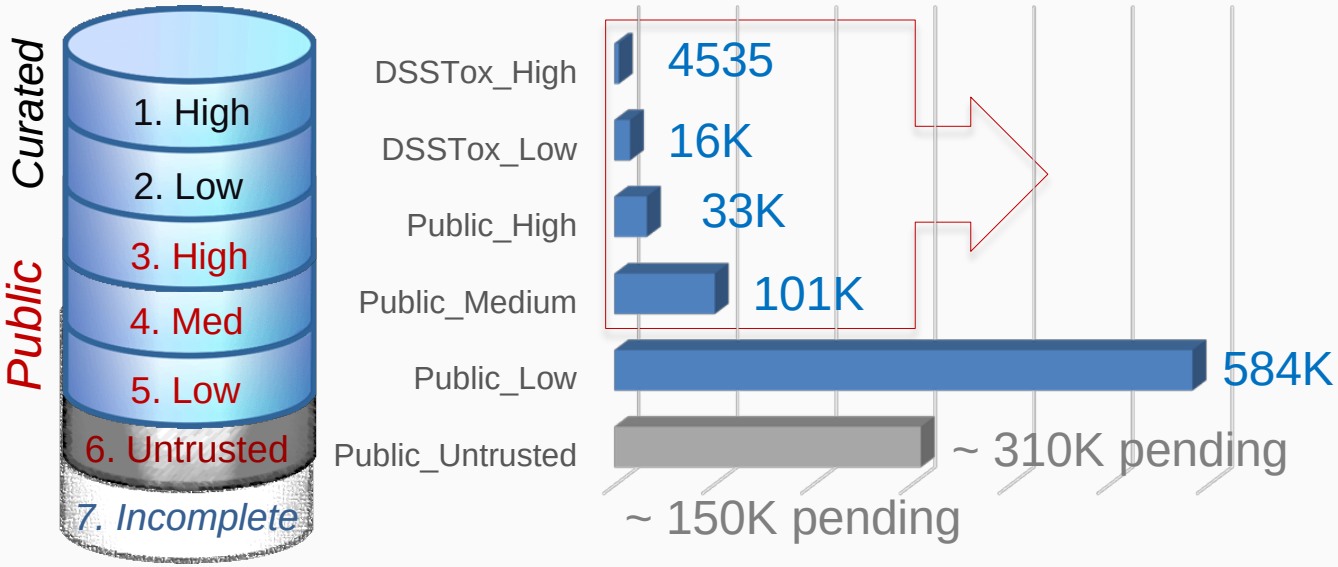
EPA researchers develop and use rapid exposure estimates to predict potential exposure for thousands of chemicals.

- High-throughput toxicokinetics data: It is important to link the external dose of a chemical to an internal blood or tissue concentration, this process is called toxicokinetics. EPA researchers measure the critical factors that determine the distribution

Underpinning with chemicals



Approximately 15 Years of Data… Growing with daily curation



QC Levels

DSSTox_High:	Hand curated and validated
DSSTox_Low:	Hand curated and confirmed using multiple public sources
Public_High:	Extracted from EPA SRS and confirmed to have no conflicts in ChemID and PubChem
Public_Medium:	Extracted from ChemID and confirmed to have no conflicts in PubChem
Public_Low:	Extracted from ACToR or PubChem
Public_Untrusted:	Postulated, but found to have conflicts in public sources

Is 750,000 substances BIG DATA?

- # of chemical substances in databases

PubChem

Compounds:	91,490,492
Substances:	230,805,802
BioAssays:	1,252,791
Tested Compounds:	2,395,818
Tested Substances:	3,818,645
RNAi BioAssays:	112
BioActivities:	233,481,319
Protein Targets:	10,342
Gene Targets:	22,079

ChemSpider
Search and share chemistry

59 Million
chemical structures

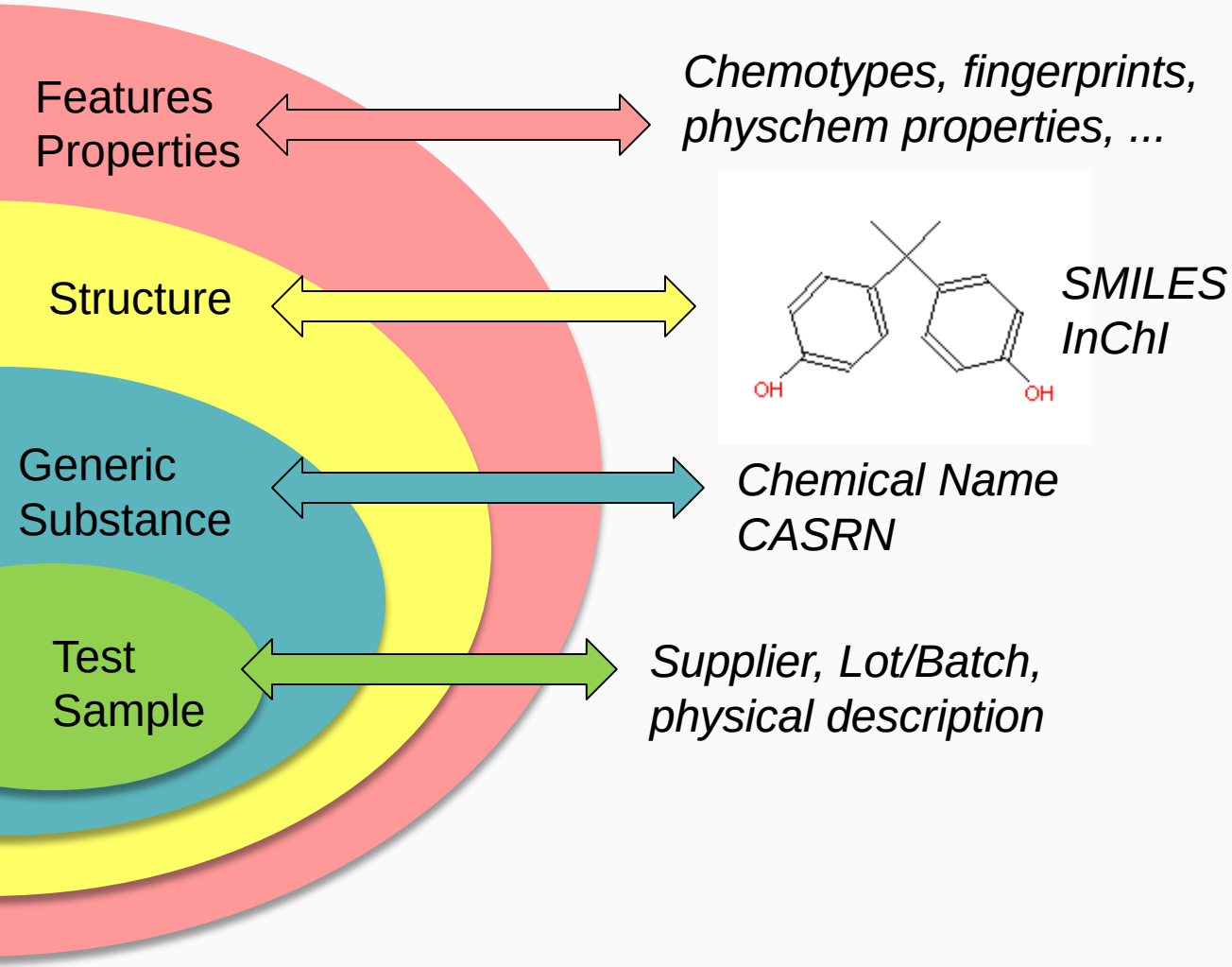
EMBL-EBI
ChEMBL

ChEMBL Statistics

- DB: ChEMBL_23
- Targets: 11,538
- Compound records: 2,101,843
- Distinct compounds: 1,735,442
- Activities: 14,675,320
- Publications: 67,722
- Release Notes

- **Data quality** is a challenge!!! Careful curation

Chemical representation levels supporting data integration



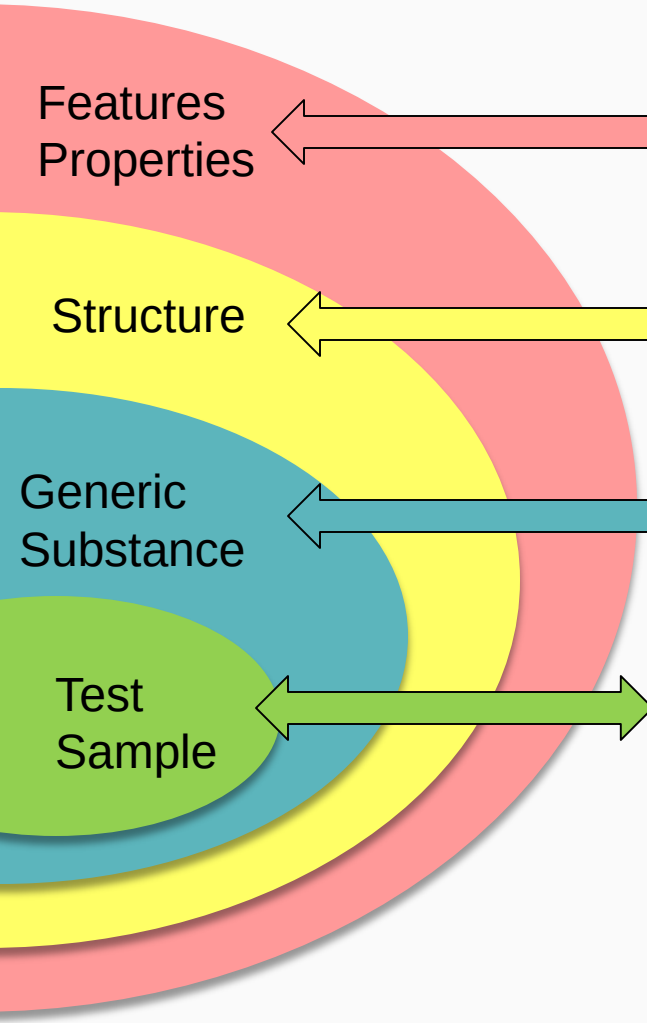
Chemotypes, fingerprints, physchem properties, ...

SMILES
InChI

Chemical Name
CASRN

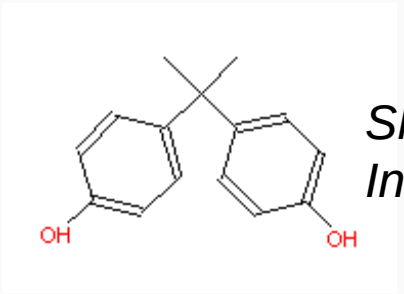
Supplier, Lot/Batch, physical description

Chemical representation levels supporting data integration



Chemotypes, fingerprints, physchem properties, ...

Chemical analogs, Read-across, SAR modeling



SMILES
InChI

Structure searching & modeling

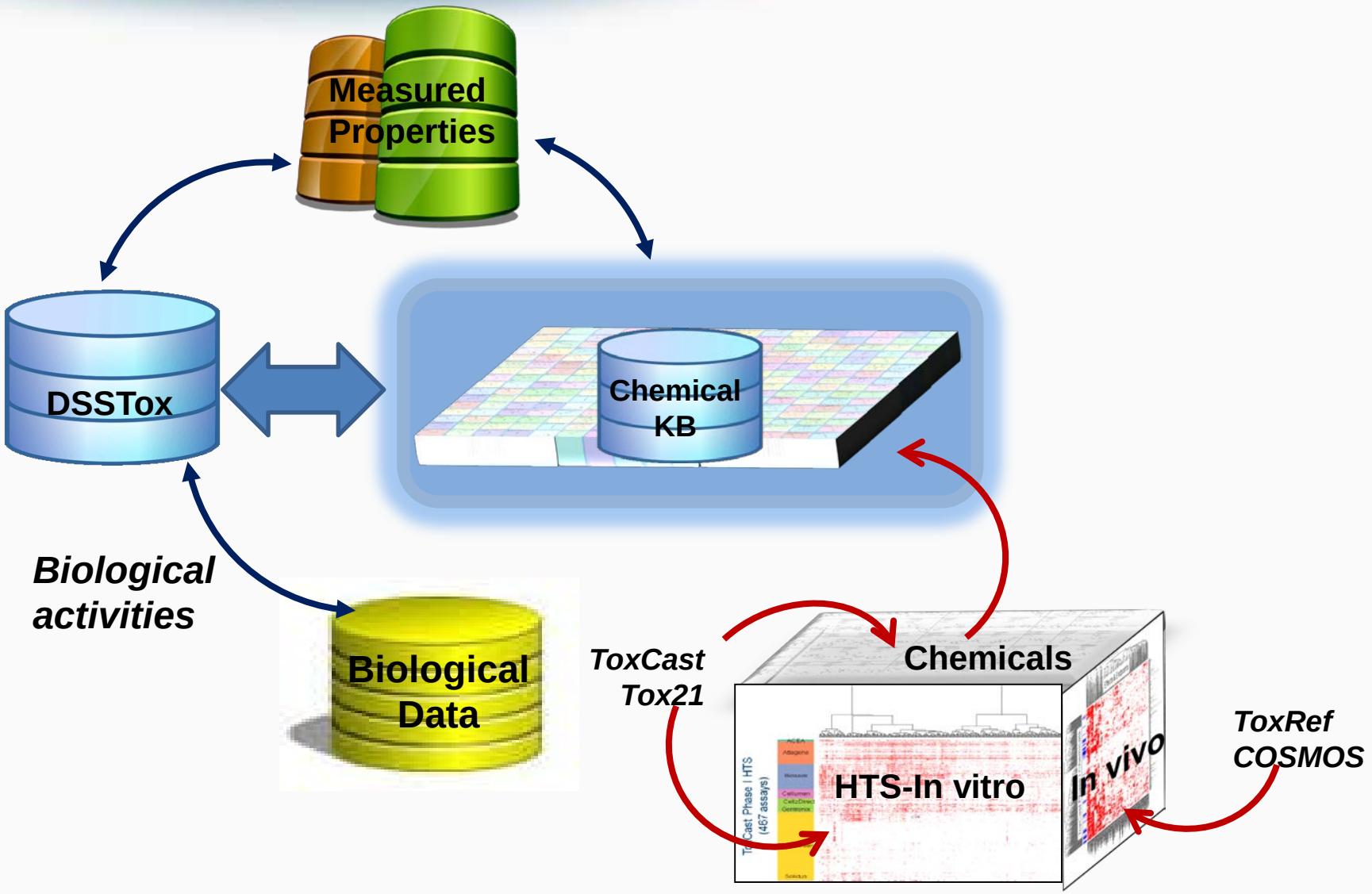
Chemical Name
CASRN

Public toxicity datasets

Supplier, Lot/Batch, physical description

Experimental Endpoint Data

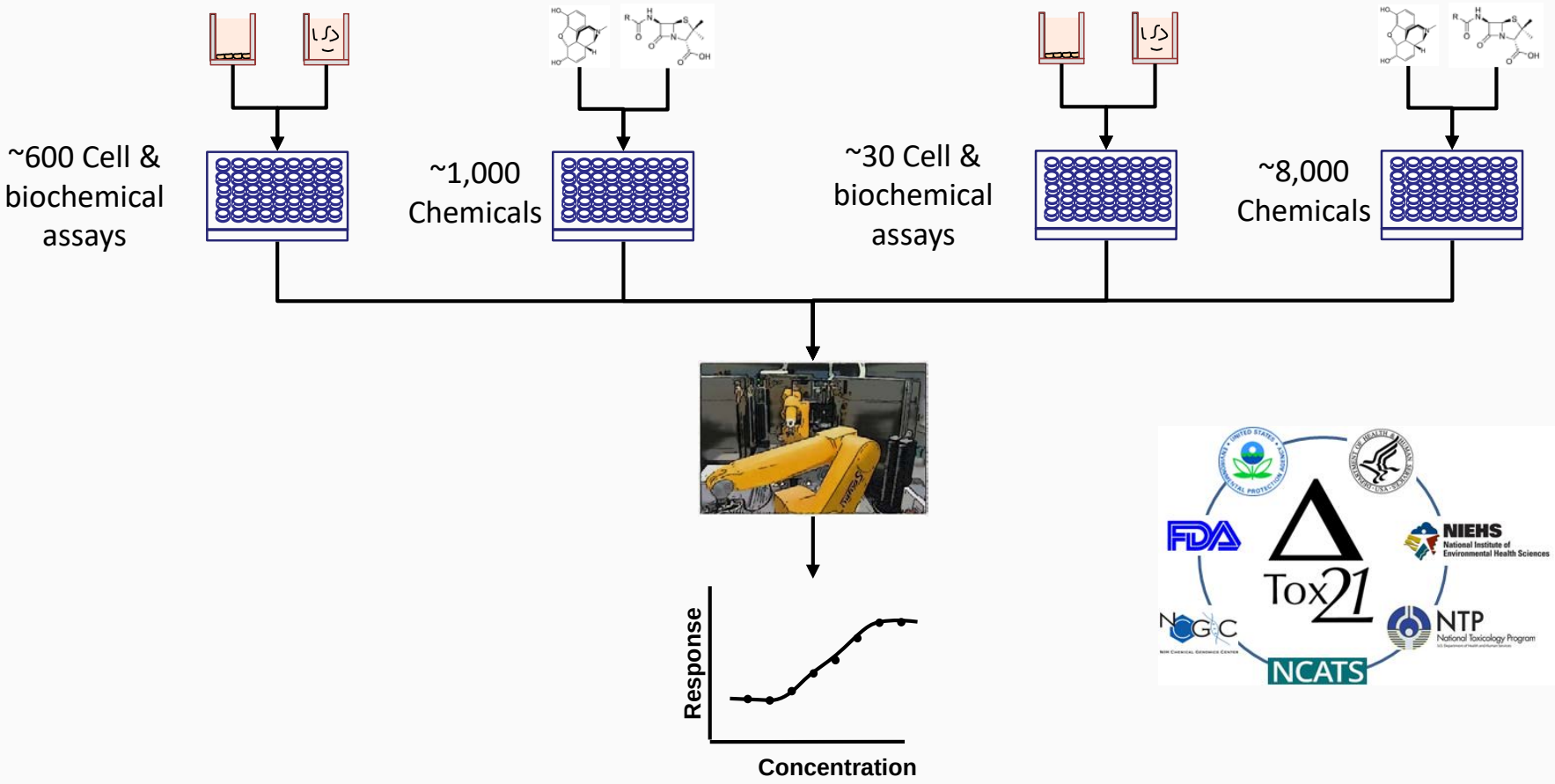
Integrating *in vitro* and *in vivo* data



Adding High-Throughput Hazard Screening Data

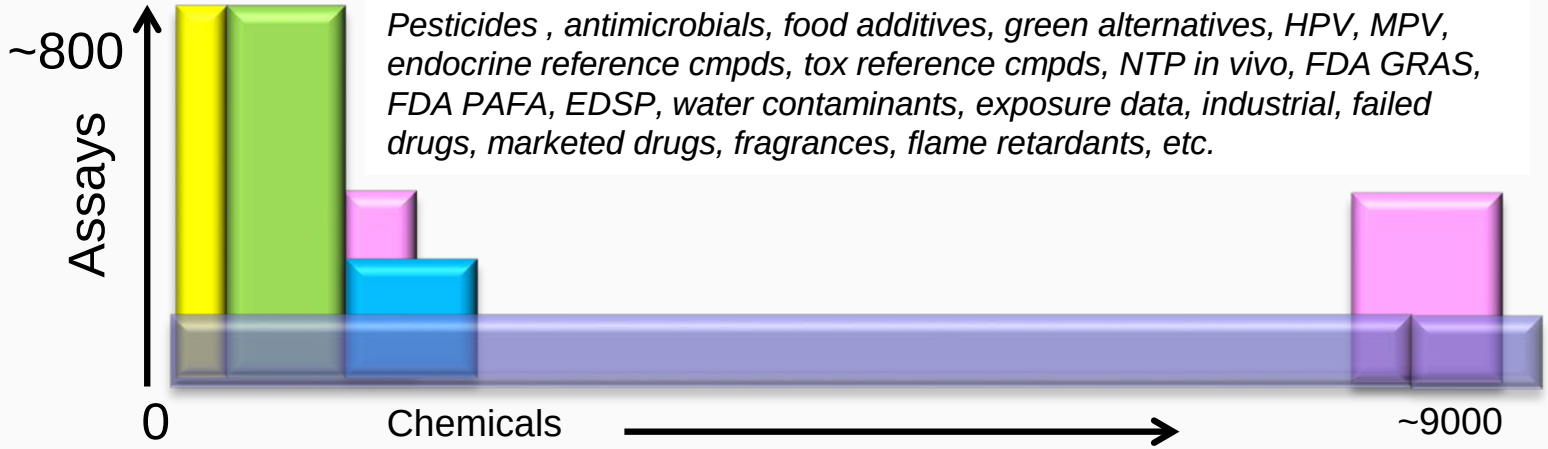
EPA United States Environmental Protection Agency

ToxCast | **Tox21**

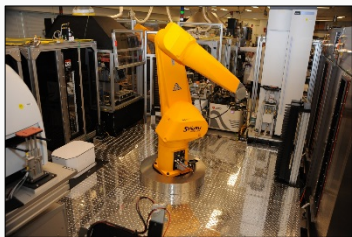


TEN YEARS of Assay Measurements: ToxCast & Tox21

Set	Chemicals	Assays	Endpoints	Completion
ToxCast Phase I	293	~600	~700	2011
ToxCast Phase II	767	~600	~700	03/2013
ToxCast E1K	800	~50	~120	03/2013
ToxCast Phase III	~900	~300	~300	In progress
Tox21	~9000	~80	~150	In progress

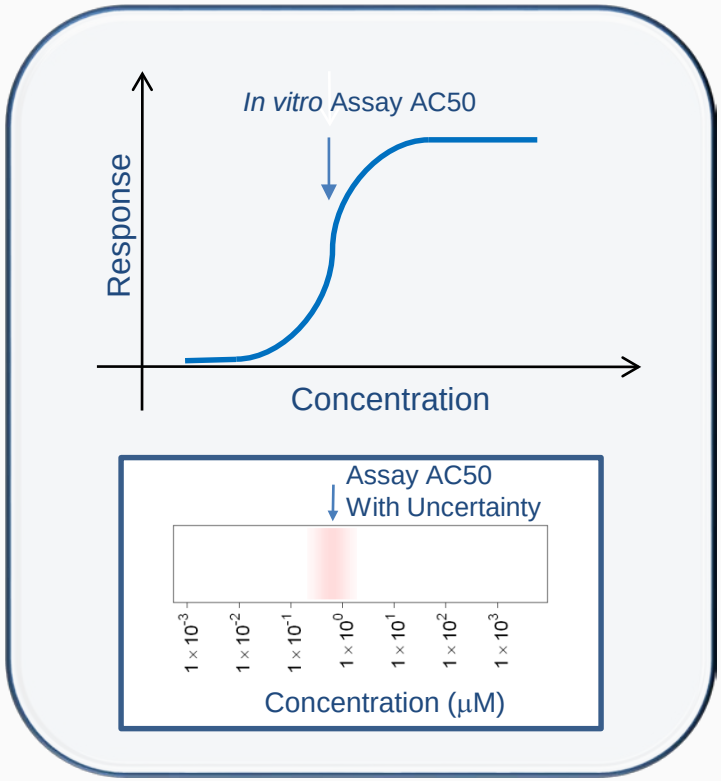


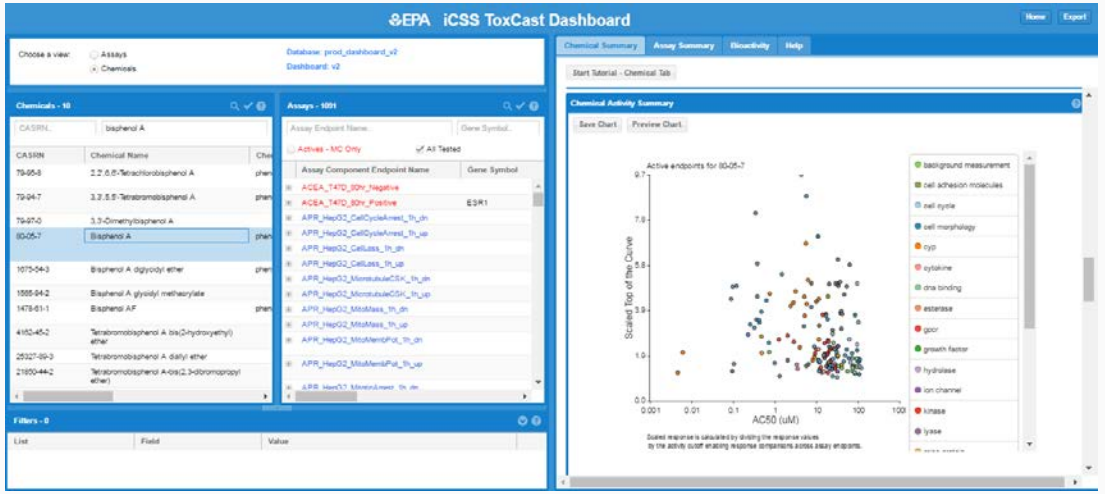
Pesticides , antimicrobials, food additives, green alternatives, HPV, MPV, endocrine reference cmpds, tox reference cmpds, NTP in vivo, FDA GRAS, FDA PAFA, EDSP, water contaminants, exposure data, industrial, failed drugs, marketed drugs, fragrances, flame retardants, etc.



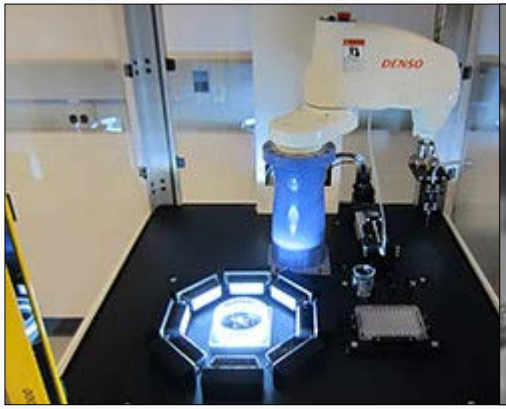
High-Throughput Bioactivity to Identify **Potential Hazards**

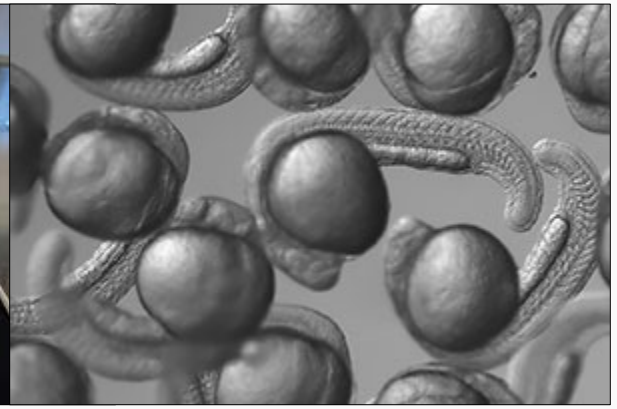
- Assays in dose-response format (AC50 – and efficacy if data described by a Hill function)



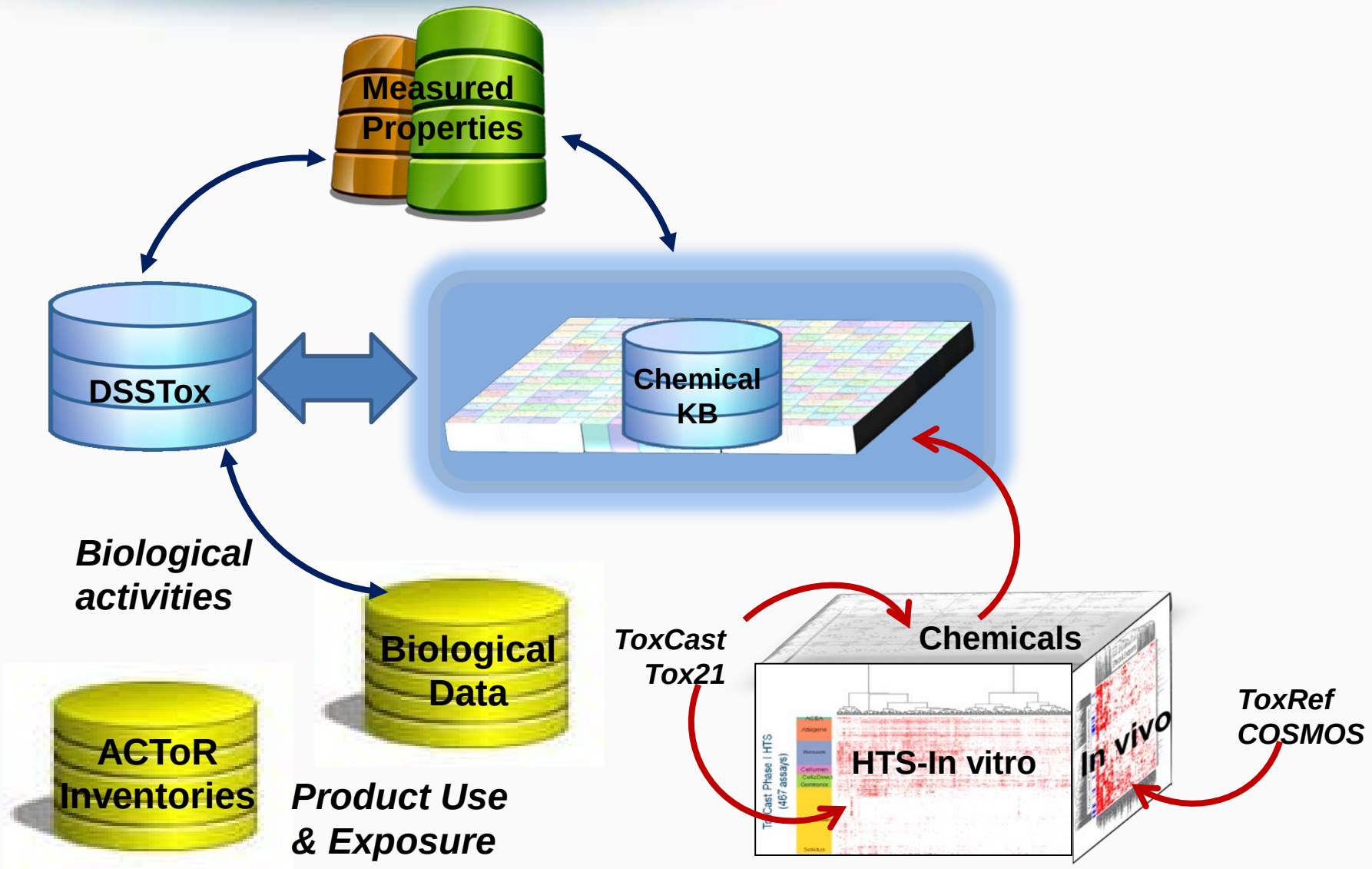


All data is made public: http://actor.epa.gov/dashboard/





Adding Product Use and Exposure



High Throughput Measurement to Identify **Exposure**

EPA United States Environmental Protection Agency

EPA US Environmental Protection Agency

Español | 中文: 繁體版 | 中文: 简体版 | Tiếng Việt | 한국어

Learn the Issues | Science & Technology | Laws & Regulations | About EPA

Search EPA.gov

Related Topics: Safer Chemicals Research

Contact Us Share

Rapid Chemical Exposure and Dose Research

EPA is responsible for ensuring the safety of thousands of chemicals. Quantitative exposure data are available for only a small fraction of registered chemicals. This type of exposure data is needed to thoroughly evaluate chemicals for potential risks to humans, wildlife and ecosystems. EPA is developing innovative methods to develop exposure estimates for thousands of chemicals to better protect human health and the environment. These innovative methods are called rapid exposure and dose assessments.

Rapid Exposure Predictions

Rapid, also called high-throughput, exposure predictions or ExpoCast provide rapid exposure estimates for thousands of chemicals. ExpoCast quickly and efficiently looks at multiple routes of exposure to provide exposure estimates. ExpoCast uses and enhances two well-known exposure models to estimate chemical exposure.

Pictured Above: Farfield Exposure Examples

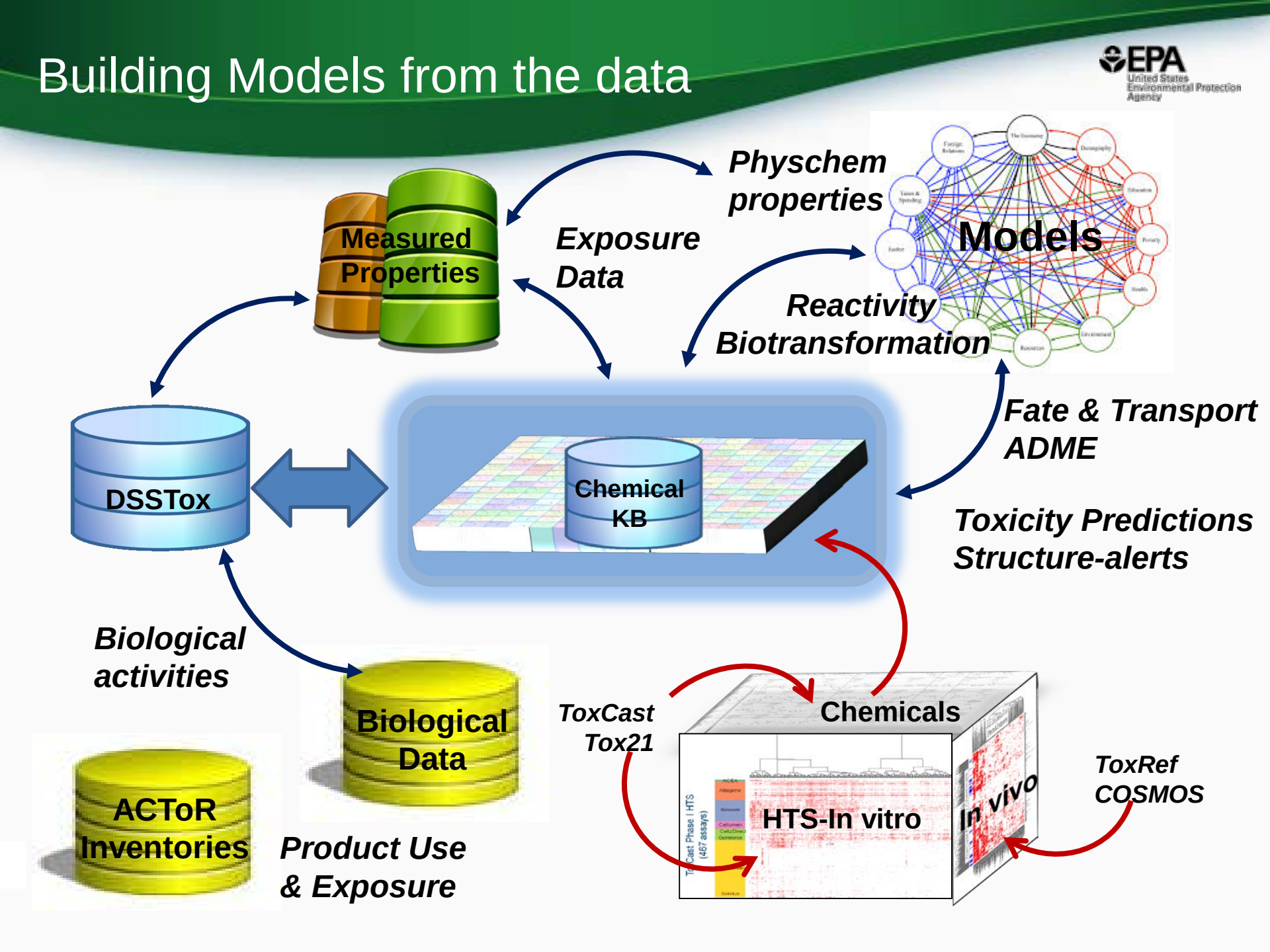
› Farfield Exposure Models

› Nearfield Exposure Models

Evaluating High-throughput Exposure Predictions

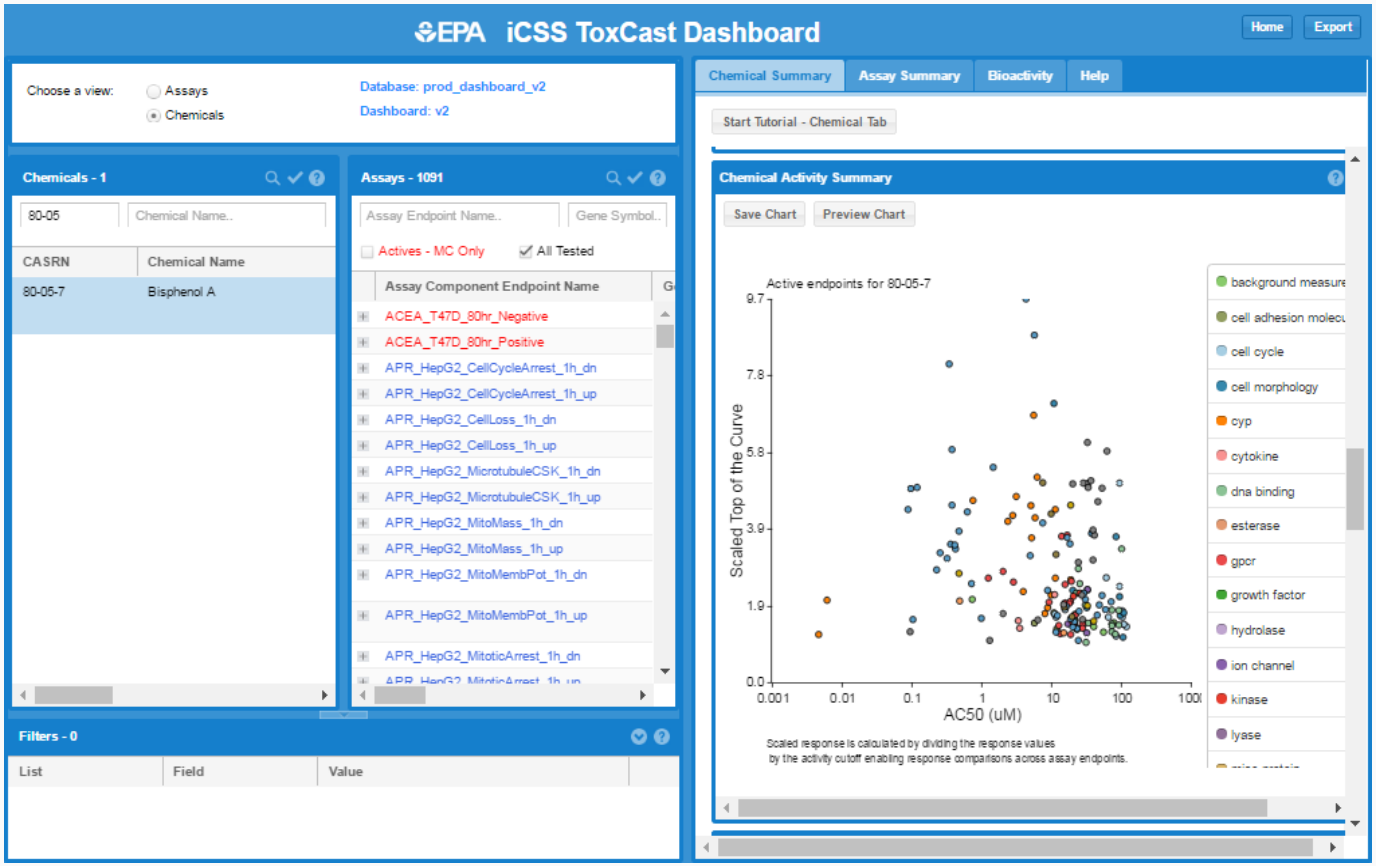
EPA is currently evaluating the effectiveness of high-throughput exposure models

Building Models from the data

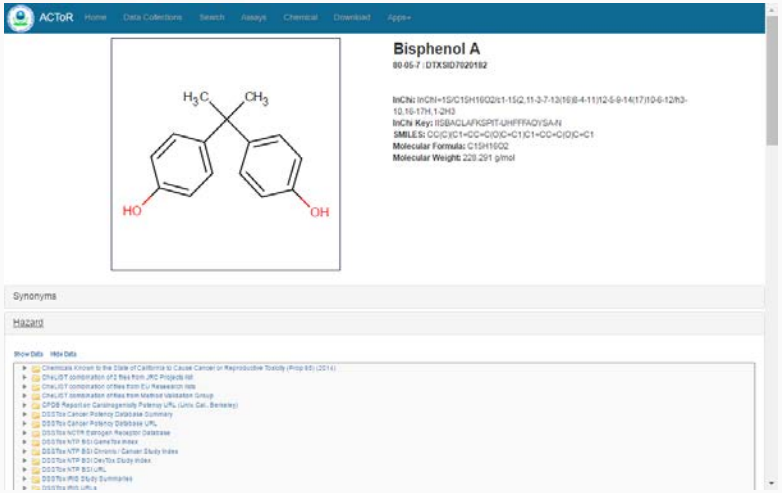


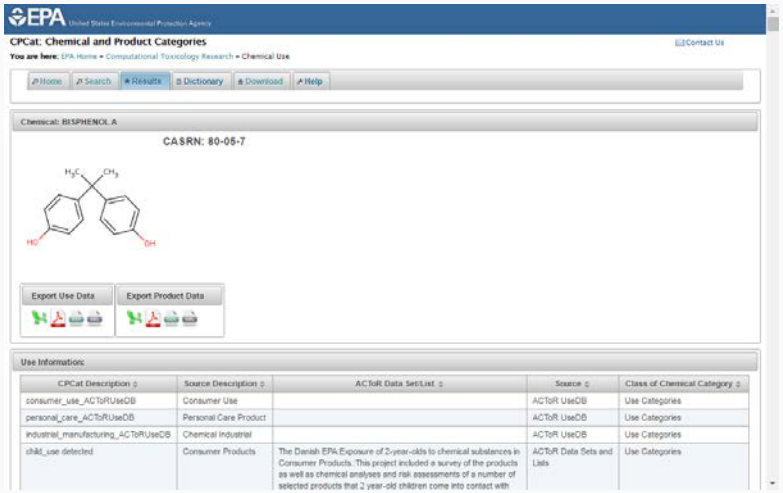
Toxcast Dashboard
https://actor.epa.gov/dashboard/

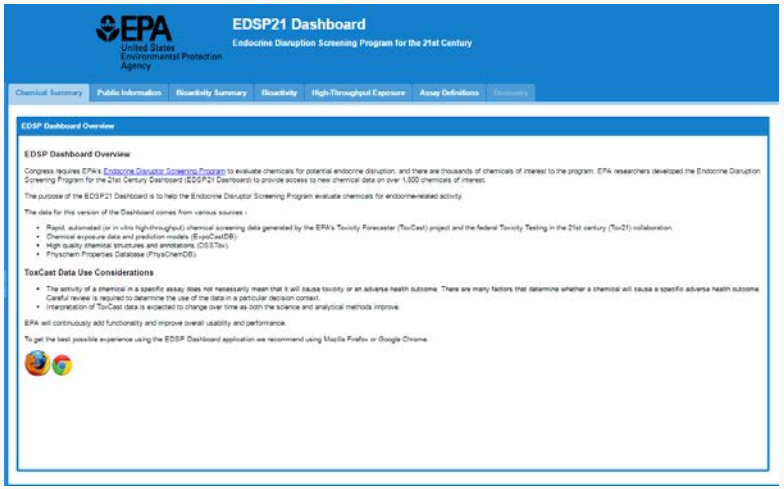
- Access and Interrogate chemical screening data from ToxCast and the Tox21 collaboration



Multiple Dashboards Delivered



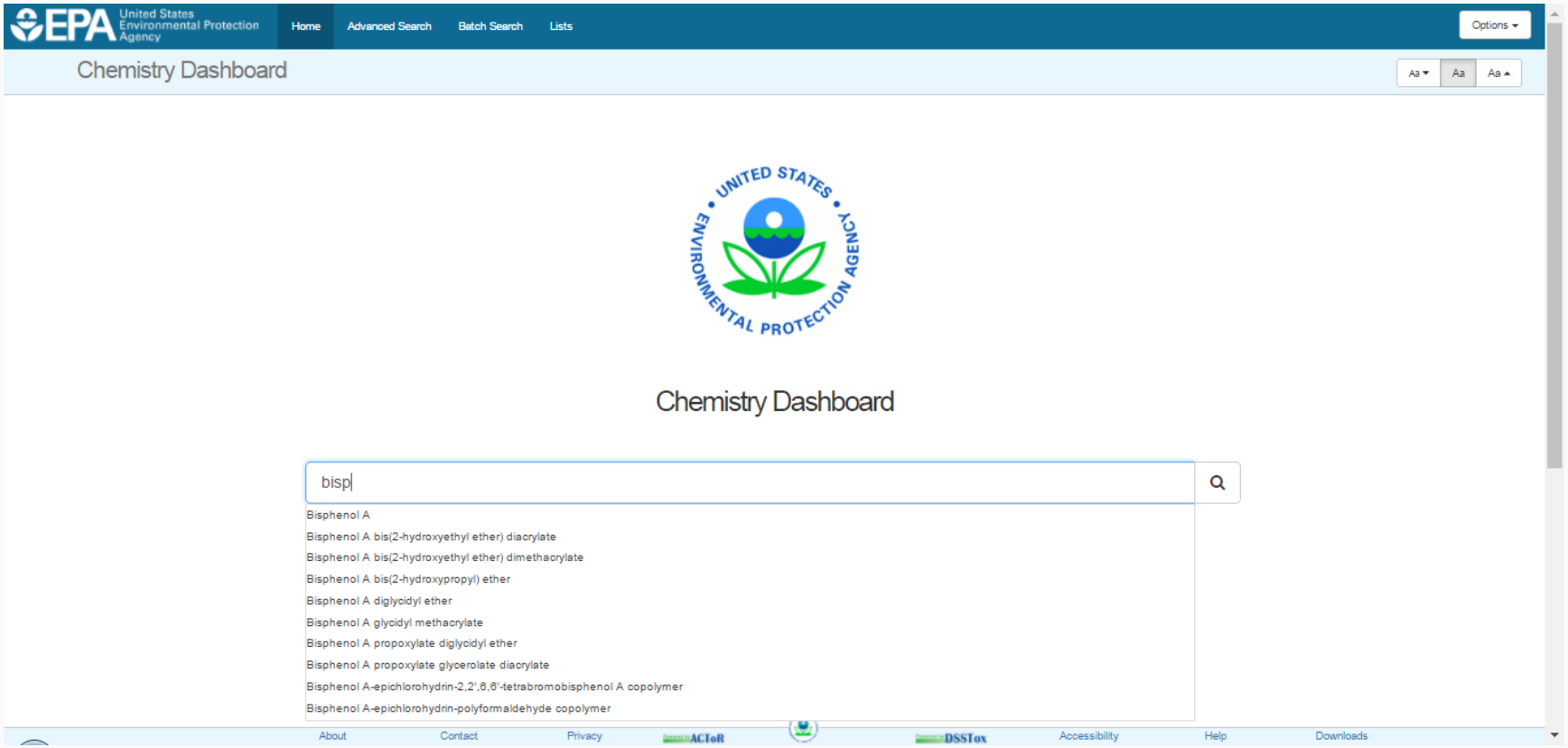




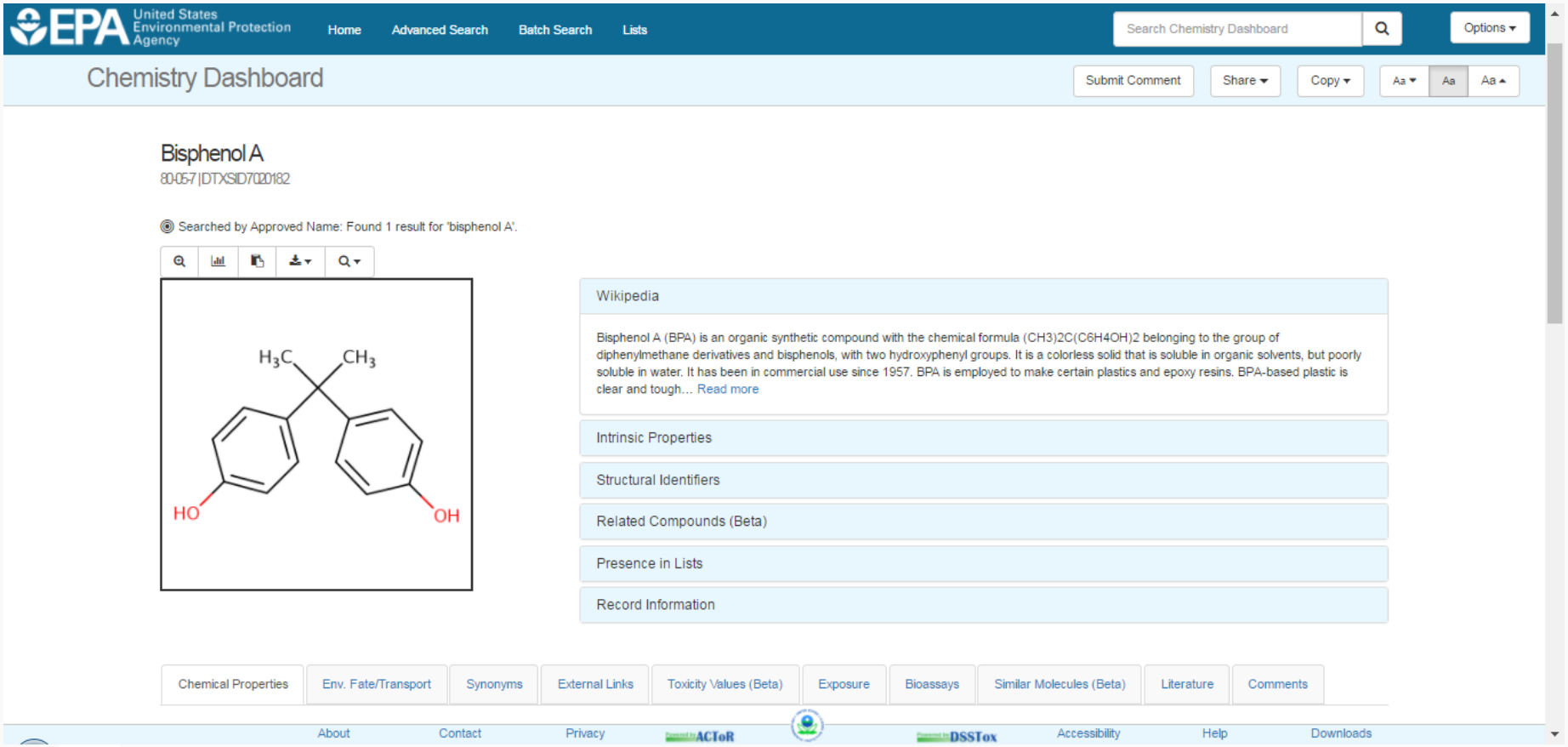
What we have learned…

- Data curation, standardization and versioning is essential
- Prototype application development suffices for research projects but production development requires managed processes
- ODOSOS (Open Data, Open Source and Open Standards) endows many benefits
- With this in mind…

Our **Latest** Dashboard
https://comptox.epa.gov

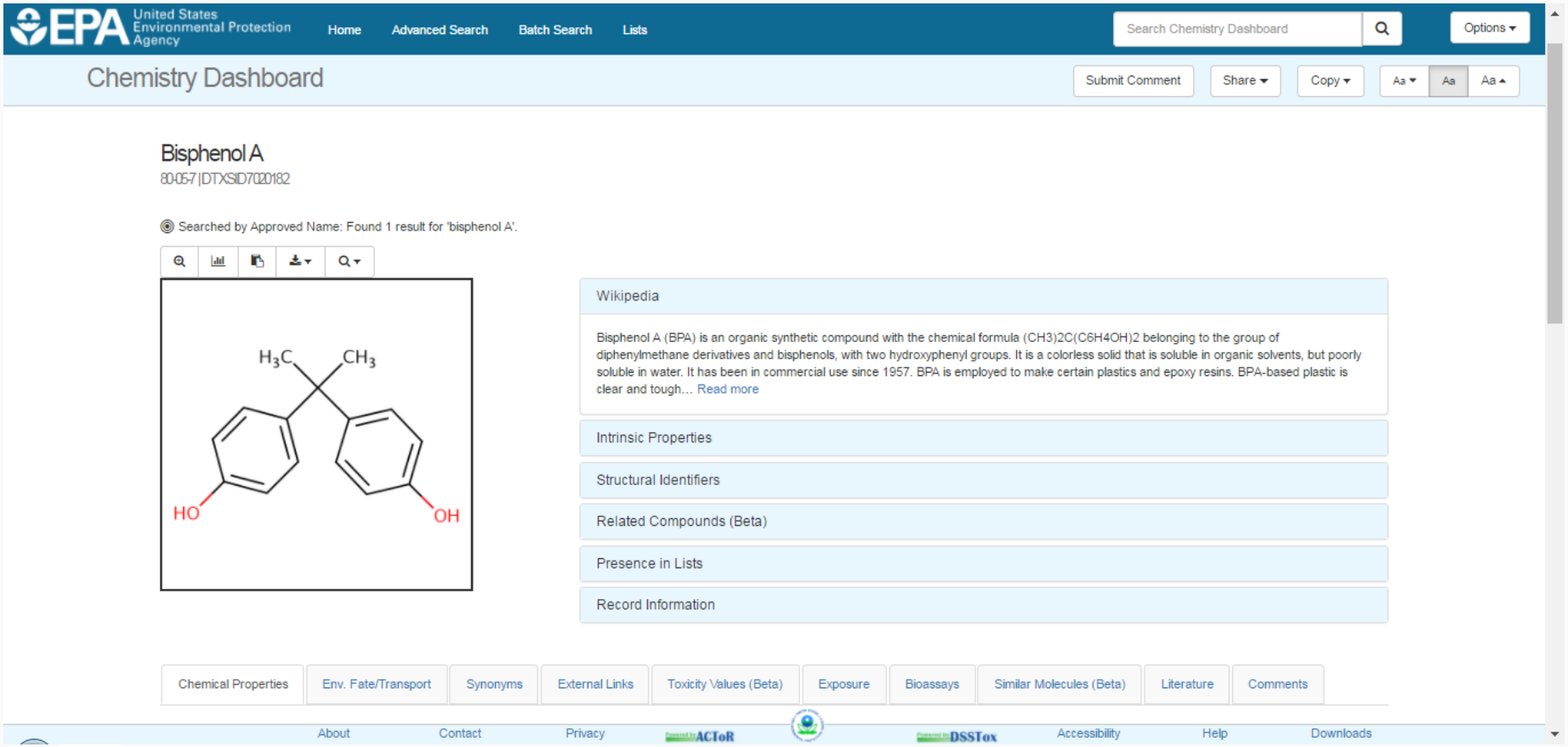


Bisphenol A



Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Chemical Properties



Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Chemical Properties

- Consuming and producing open data

Summary

- LogP: Octanol-Water
- Water Solubility
- Density
- Melting Point
- Boiling Point
- Surface Tension
- Vapor Pressure
- LogKoa: Octanol-Air
- Henry's Law
- Index of Refraction
- Molar Refractivity
- pKa Acidic Apparent

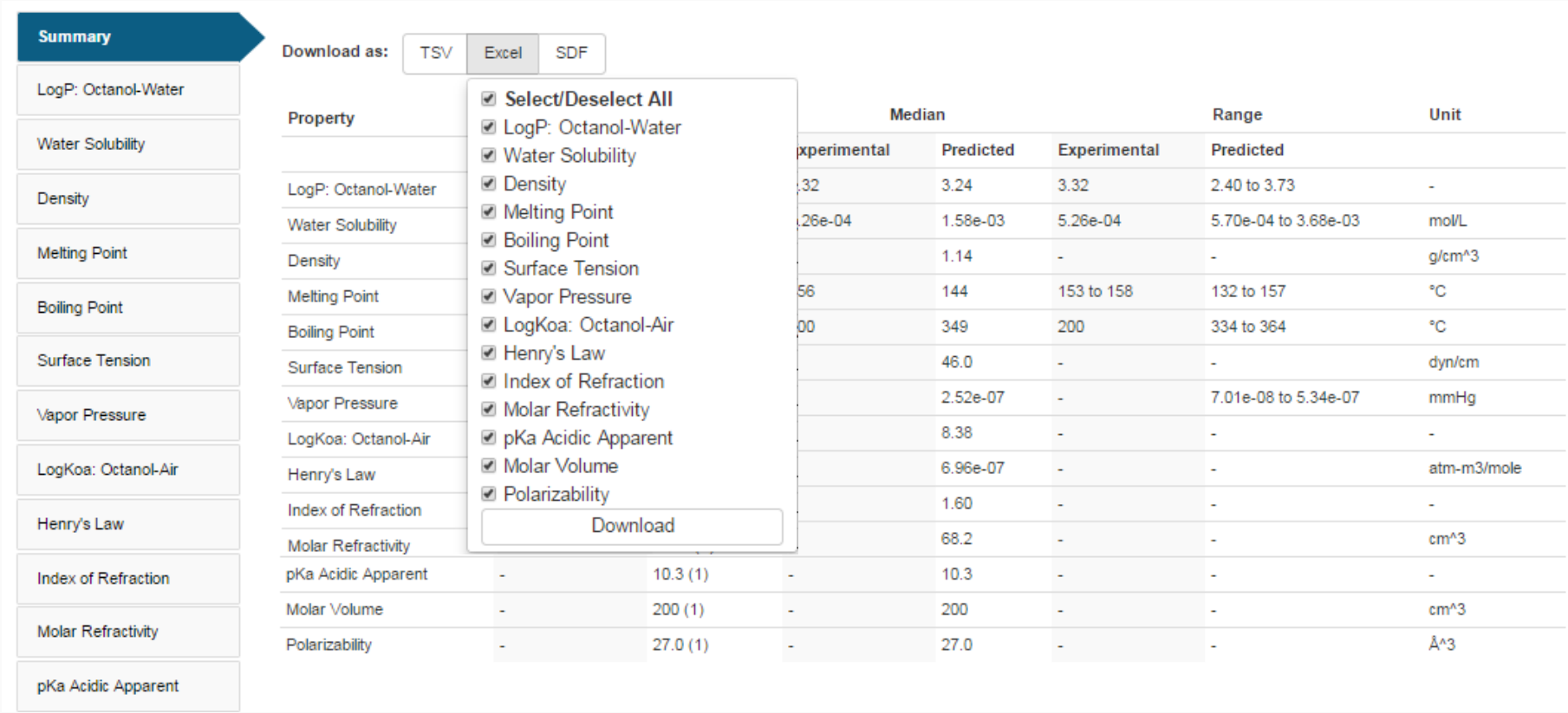
Download as: TSV | Excel | SDF

Property	Average		Median		Range		Unit
	Experimental	Predicted	Experimental	Predicted	Experimental	Predicted	
LogP: Octanol-Water	3.32 (1)	3.24 (4)	3.32	3.24	3.32	2.40 to 3.73	-
Water Solubility	5.26e-04 (1)	1.58e-03 (4)	5.26e-04	1.58e-03	5.26e-04	5.70e-04 to 3.68e-03	mol/L
Density	-	1.14 (1)	-	1.14	-	-	g/cm^3
Melting Point	155 (7)	144 (3)	156	144	153 to 158	132 to 157	°C
Boiling Point	200 (1)	349 (3)	200	349	200	334 to 364	°C
Surface Tension	-	46.0 (1)	-	46.0	-	-	dyn/cm
Vapor Pressure	-	2.52e-07 (3)	-	2.52e-07	-	7.01e-08 to 5.34e-07	mmHg
LogKoa: Octanol-Air	-	8.38 (1)	-	8.38	-	-	-
Henry's Law	-	6.96e-07 (1)	-	6.96e-07	-	-	atm-m3/mole
Index of Refraction	-	1.60 (1)	-	1.60	-	-	-
Molar Refractivity	-	68.2 (1)	-	68.2	-	-	cm^3
pKa Acidic Apparent	-	10.3 (1)	-	10.3	-	-	-
Molar Volume	-	200 (1)	-	200	-	-	cm^3
Polarizability	-	27.0 (1)	-	27.0	-	-	Å^3

Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Data Distribution

- Consuming and producing open data



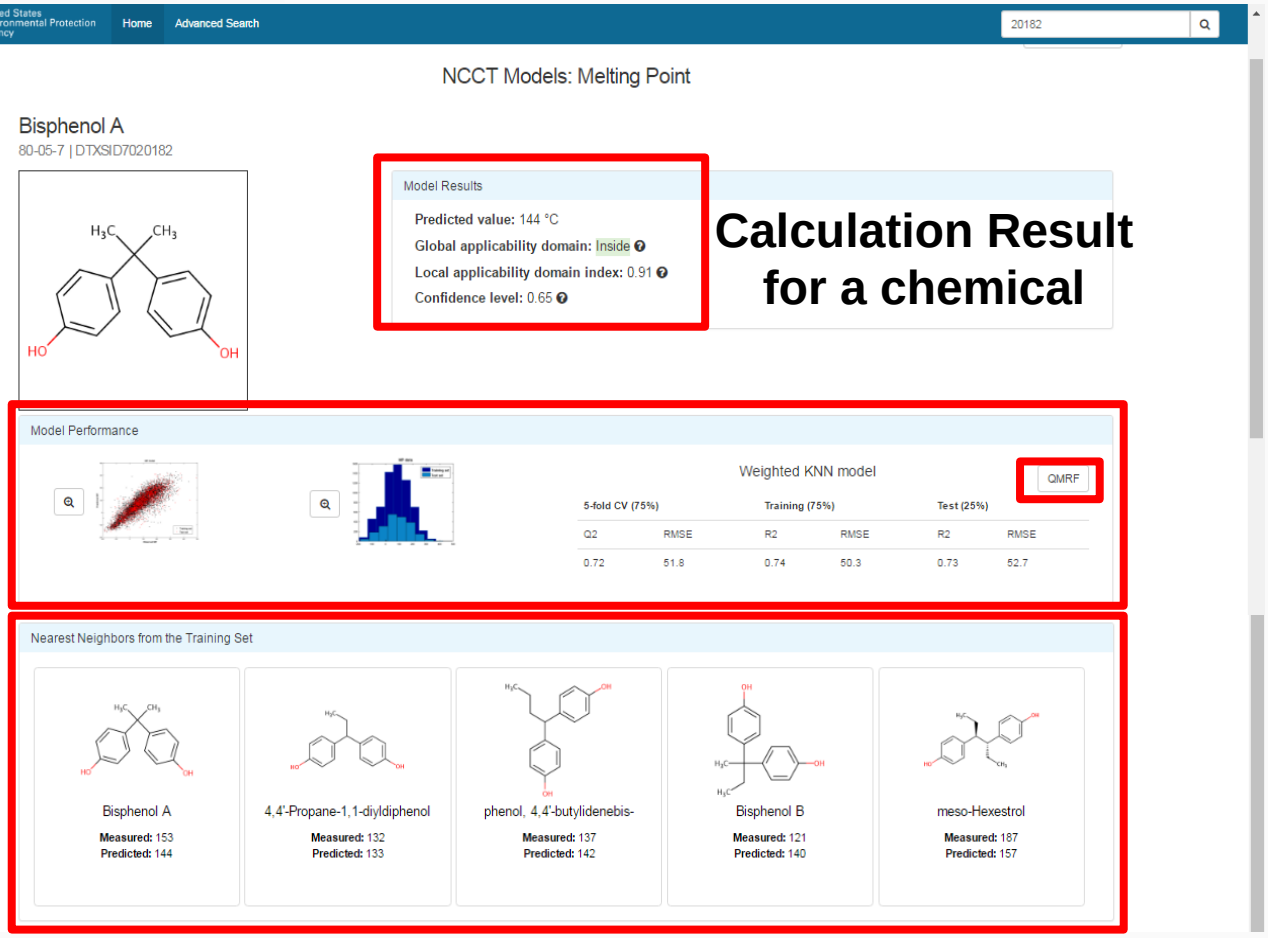
Data Distribution

	A	B	C	D	E	F	G	H
1	**Property**	**Average**		**Median**		**Range**		**Unit**
2		**Experimental**	**Predicted**	**Experimental**	**Predicted**	**Experimental**	**Predicted**	
3	LogP: Octanol-Water	3.32 (1)	3.24 (4)	3.32	3.24	3.32	2.40 to 3.73	-
4	Water Solubility	5.26e-04 (1)	1.58e-03 (4)	0.000526	0.00158	0.000526	5.70e-04 to 3.68e-03	mol/L
5	Density	-	1.14 (1)	-	1.14	-	-	g/cm^3
6	Melting Point	155 (7)	144 (3)	156	144	153 to 158	132 to 157	°C
7	Boiling Point	200 (1)	349 (3)	200	349	200	334 to 364	°C
8	Surface Tension	-	46.0 (1)	-	46	-	-	dyn/cm
9	Vapor Pressure	-	2.52e-07 (3)	-	0.000000252	-	7.01e-08 to 5.34e-07	mmHg
10	LogKoa: Octanol-Air	-	8.38 (1)	-	8.38	-	-	-
11	Henry's Law	-	6.96e-07 (1)	-	0.000000696	-	-	atm-m3/mole
12	Index of Refraction	-	1.60 (1)	-	1.6	-	-	-
13	Molar Refractivity	-	68.2 (1)	-	68.2	-	-	cm^3
14	pKa Acidic Apparent	-	10.3 (1)	-	10.3	-	-	-
15	Molar Volume	-	200 (1)	-	200	-	-	cm^3
16	Polarizability	-	27.0 (1)	-	27	-	-	Å^3

Modeling Details

Model Performance with full QMRF

Nearest Neighbors from Training Set



Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Prediction Details and QMRF Report

Model Results

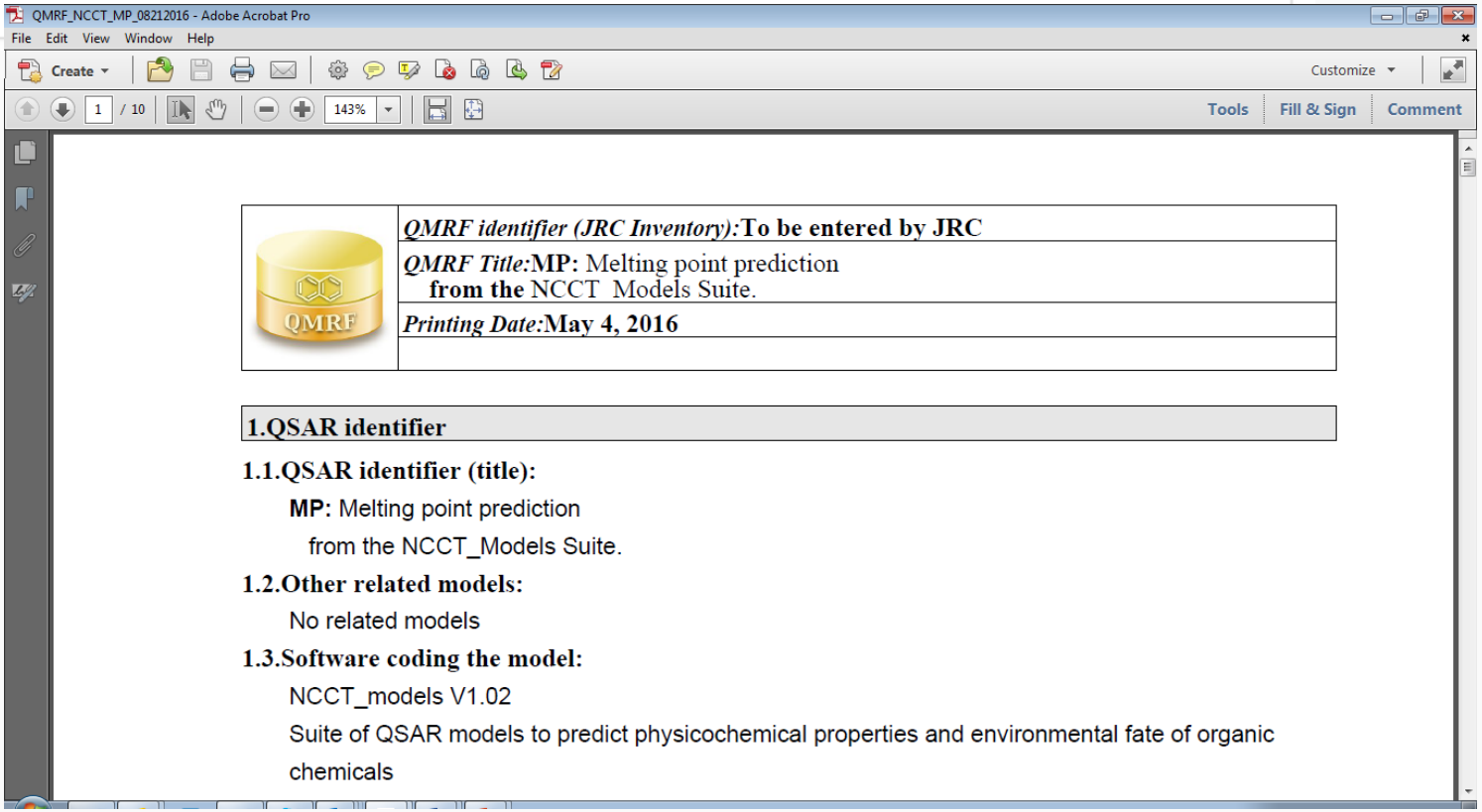
Predicted value: 144 °C

Global applicability domain: Inside

Applicability domain using the leverage approach. All training set space considered. More details in QMRF.

Local applicability domain index: 0.91

Confidence level: 0.65



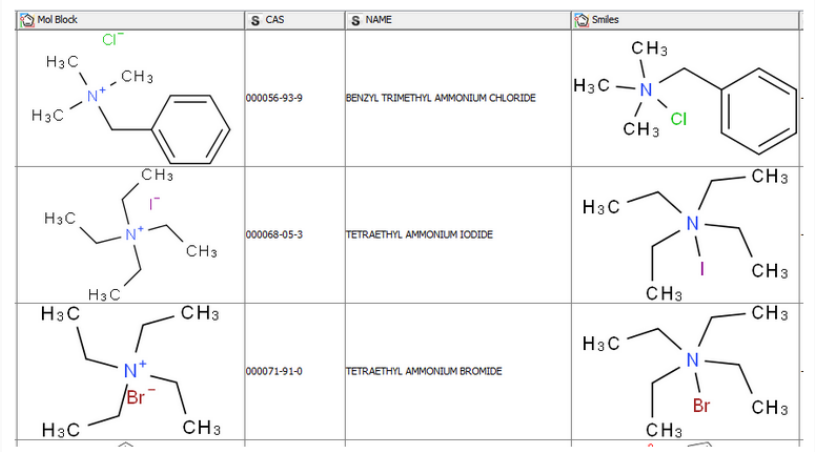
Developing “NCCT Models”

- Our approach to modeling:
  - Obtain **high quality** training sets
  - Apply appropriate modeling approaches
  - **Validate** performance of models
  - Define the applicability domain and model limitations
  - Use models to predict properties across our full datasets
  - Release as **Open Data and Open Models**

Consuming and Curating Public Data

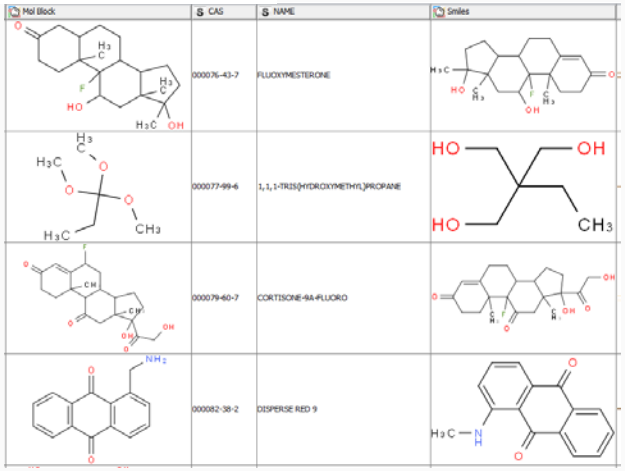
EPA United States Environmental Protection Agency

Public data should be curated prior to modeling



Covalent Halogens

Identical Chemicals



Mismatches

LogP dataset: 15,809 structures

- CAS Checksum: 12163 valid, 3646 **invalid** (>23%)
- **Invalid** names: 555
- **Invalid** SMILES 133
- Valence **errors**: 322 Molfile, 3782 SMILES (>24%)
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry (~8%)
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds (~3%)

Workflow Details and Data

Journal
SAR and QSAR in Environmental Research ›
Volume 27, 2016 - Issue 11: 17th International Conference on QSAR in Environmental and Health Sciences (QSAR 2016) - Part II. Guest Editors: C.G. Barber and G.J. Myatt

Enter keywords, authors, DOI etc.

258
Views
4
CrossRef citations
16
Altmetric

Articles

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling $

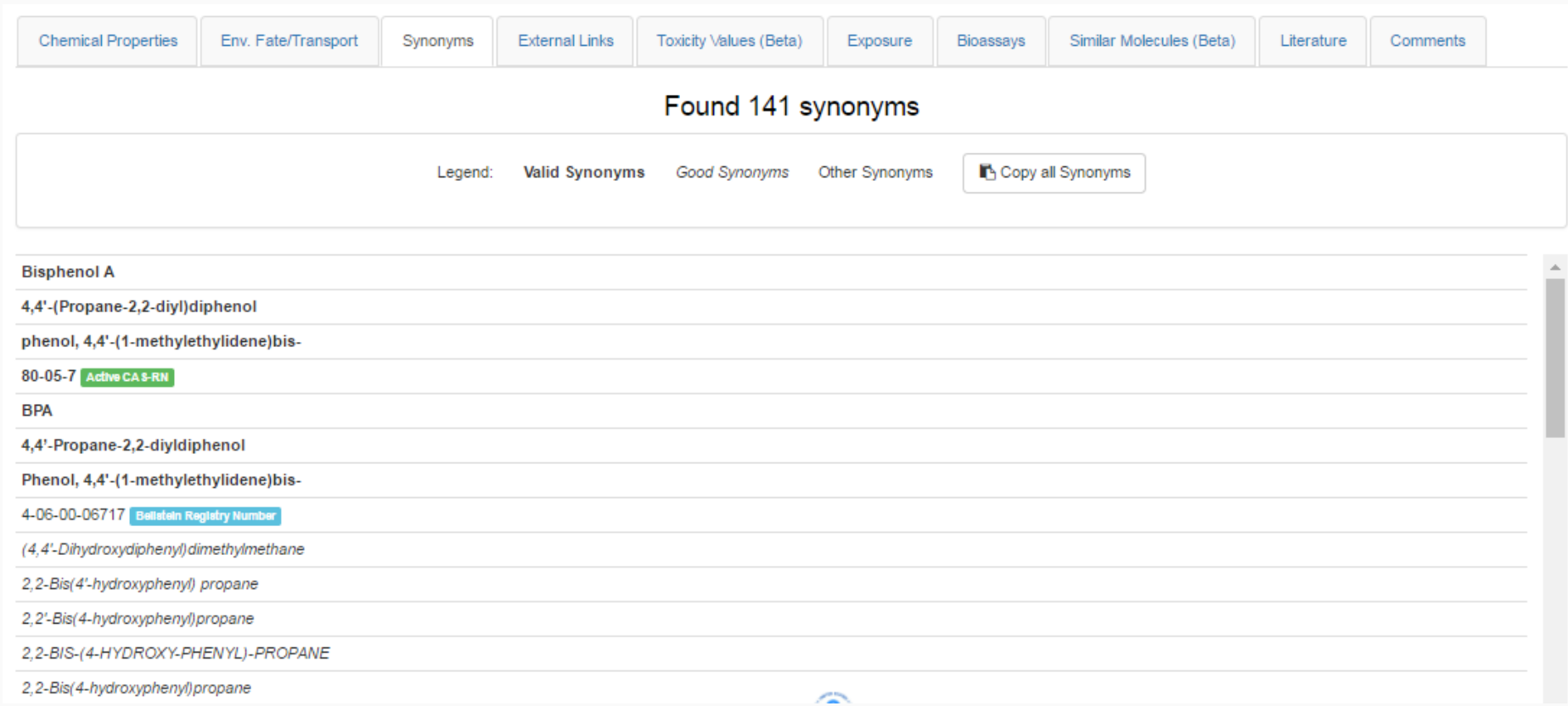
K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

Pages 911-937 | Received 03 Sep 2016, Accepted 24 Oct 2016, Published online: 25 Nov 2016

Download citation http://dx.doi.org/10.1080/1062936X.2016.1253611 Check for updates

OPERA Models: https://github.com/kmansouri/OPERA

Synonyms – curated and Open



Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Taking advantage of online data

Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

General

- EPA Substance Regist...
- HMDB
- Wikipedia
- MSDS Lookup
- ToxPlanet
- ChemHat: Hazards an...
- ChEMBL
- Consumer Product Inf...
- ECHA Brief Profile
- ECHA Infocard
- Sigma-Aldrich Chemic...

Toxicology

- ACToR
- eChemPortal
- EDSP Dashboard
- Gene-Tox
- HSDB
- ToxCast Dashboard 2
- LactMed
- International Toxicity E...
- ACToR PDF Report

Publications

- Toxline
- Google Patents
- PubMed
- BioCaddie DataMed
- Federal Register
- Regulations.gov
- RSC Publications
- Springer Materials
- IRIS Assessments
- CORE Literature Sear...
- Bielefeld Academic Se...

Analytical

- RSC Analytical Abstra...
- FOR-IDENT
- MONA: MassBank No...
- NEMI: National Enviro...
- Tox21 Analytical Data

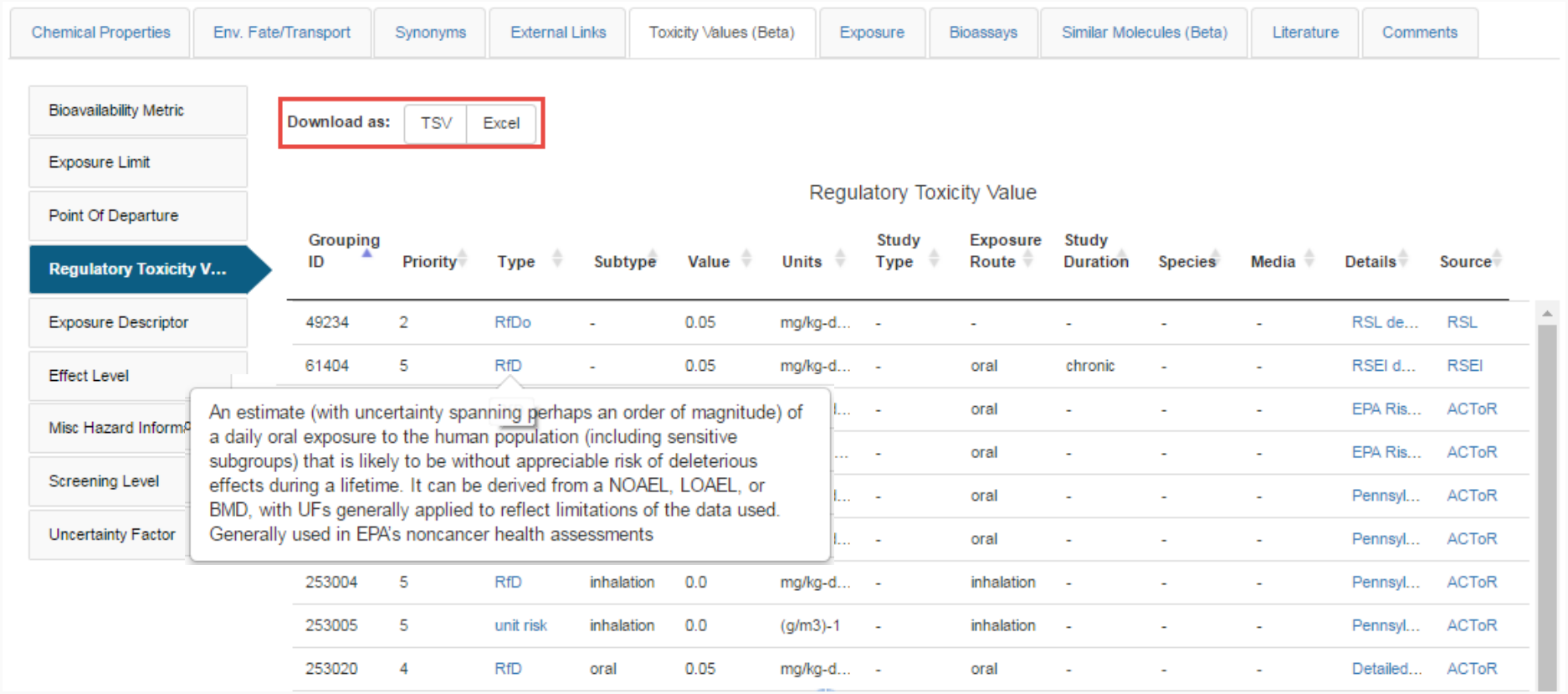
Prediction

- Chemicalize
- Proton NMR Prediction
- Carbon-13 NMR Pred...
- 2D NMR HSQC/HMB...
- ChemRTP Predictor

eChemPortal provides free public access to information on properties of chemicals. Direct links to collections of chemical hazard and risk information prepared for government chemical review programmes at national, regional and international levels are obtained.

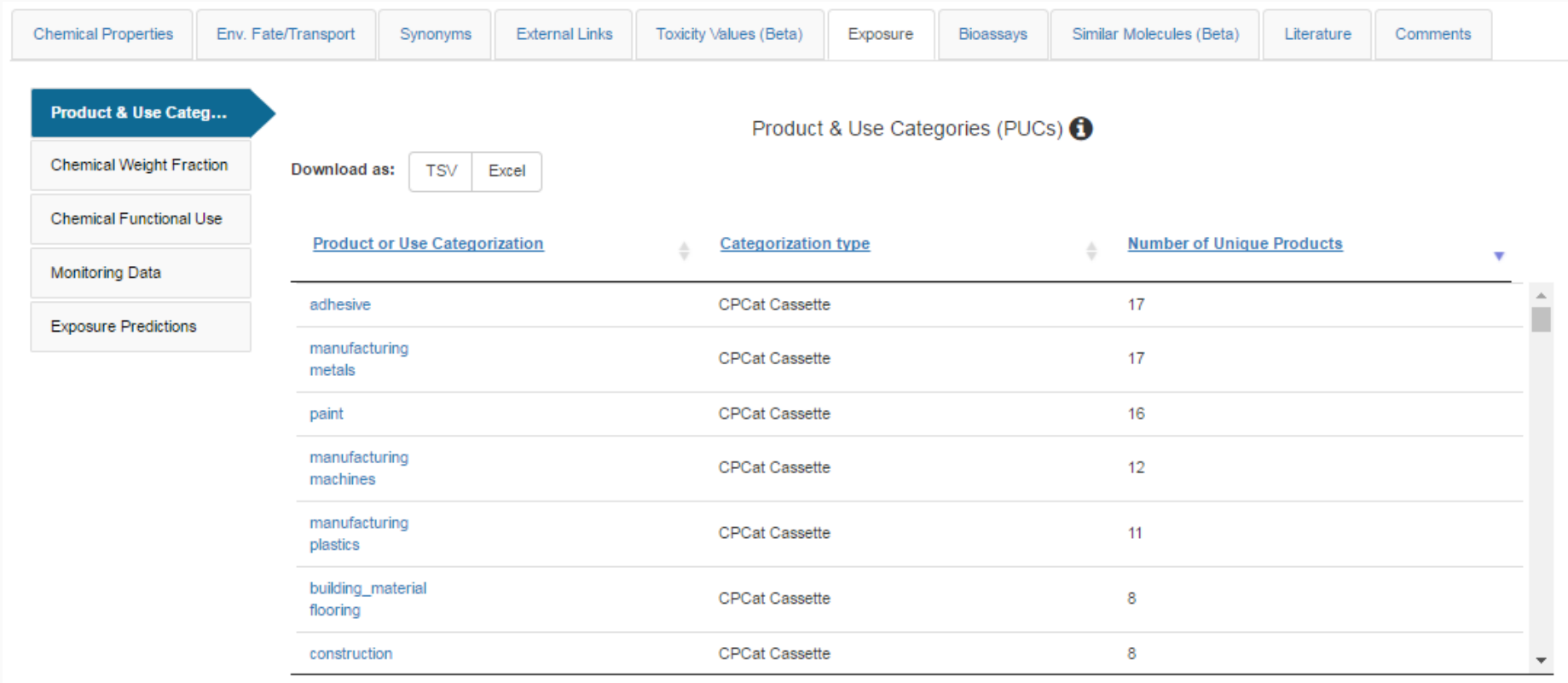
Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Consuming Available Toxicity Value Data



Chemical Properties
Env. Fate/Transport
Synonyms
External Links
Toxicity Values (Beta)
Exposure
Bioassays
Similar Molecules (Beta)
Literature
Comments

Exposure



Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Consumer Product data

- Data gathered from multiple online data sources – text mining and data downloads

- CASRN and name mappings to produce curated structure set with functional uses
- Produce Functional Use (FUSEDB) and build predicted functional use models

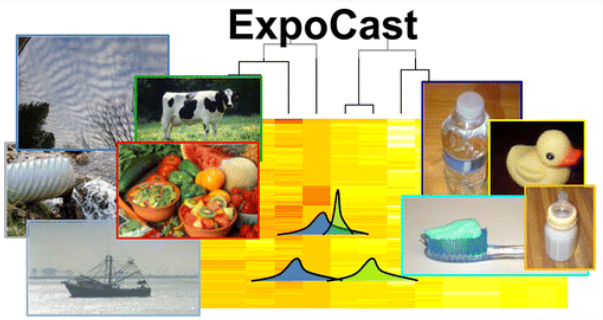
Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

Exposure

National Health and Nutrition Examination Survey

High-Throughput Models for Exposure-Based Chemical Prioritization in the ExpoCast Project

Environ. Sci. Technol., **2013**, *47* (15), pp 8479–8488



- Product & Use Categori...
- Chemical Weight Fraction
- Chemical Functional Use
- **Monitoring Data**
- Exposure Predictions

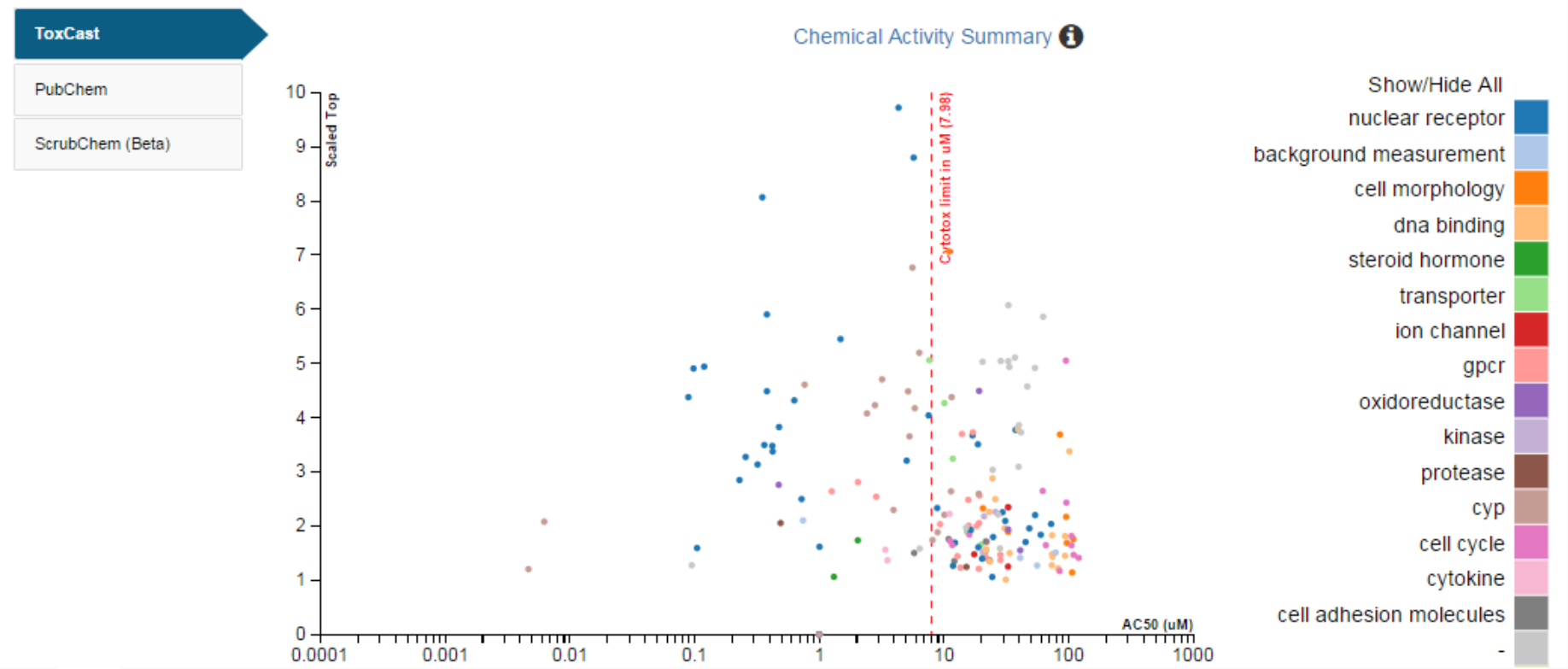
Download as: TSV | Excel

National Health and Nutrition Examination Survey (NHANES) Inferences (mg/kg-bw/day)

	Lower 95th Limit	Upper 95th Limit	Median
Ages 6-11	3.80e-05	4.92e-05	4.33e-05
Ages 12-19	2.55e-05	3.38e-05	2.93e-05
Ages 20-65	2.79e-05	3.27e-05	3.02e-05
Ages 65+	1.91e-05	2.31e-05	2.10e-05
BMI > 30	2.38e-05	2.74e-05	2.55e-05
BMI < 30	3.02e-05	3.30e-05	3.16e-05
Repro. Age Females	2.83e-05	3.31e-05	3.06e-05
Females	2.58e-05	3.03e-05	2.80e-05
Males	2.94e-05	3.37e-05	3.15e-05
Total	2.86e-05	3.08e-05	2.97e-05

Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

ToxCast and Tox21 Bioassays



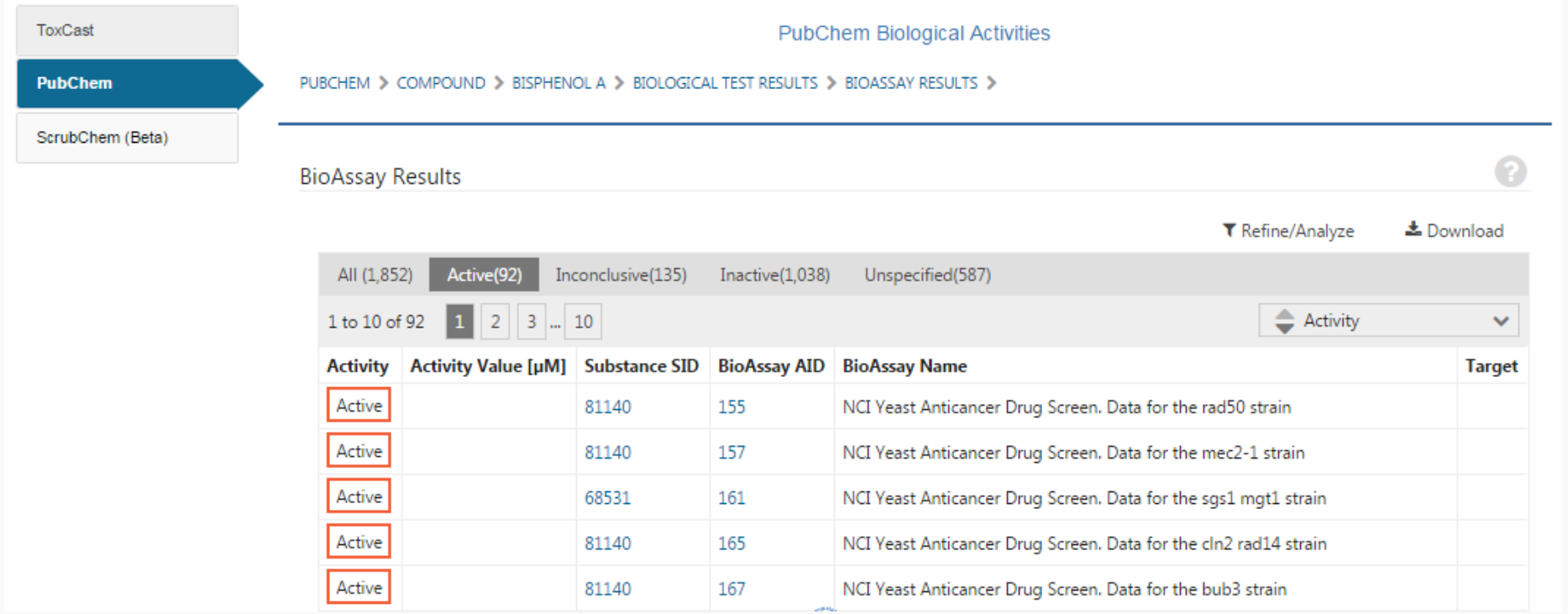
Chemical Properties | Env. Fate/Transport | Synonyms | External Links | Toxicity Values (Beta) | Exposure | Bioassays | Similar Molecules (Beta) | Literature | Comments

ToxCast and Tox21 Bioassays

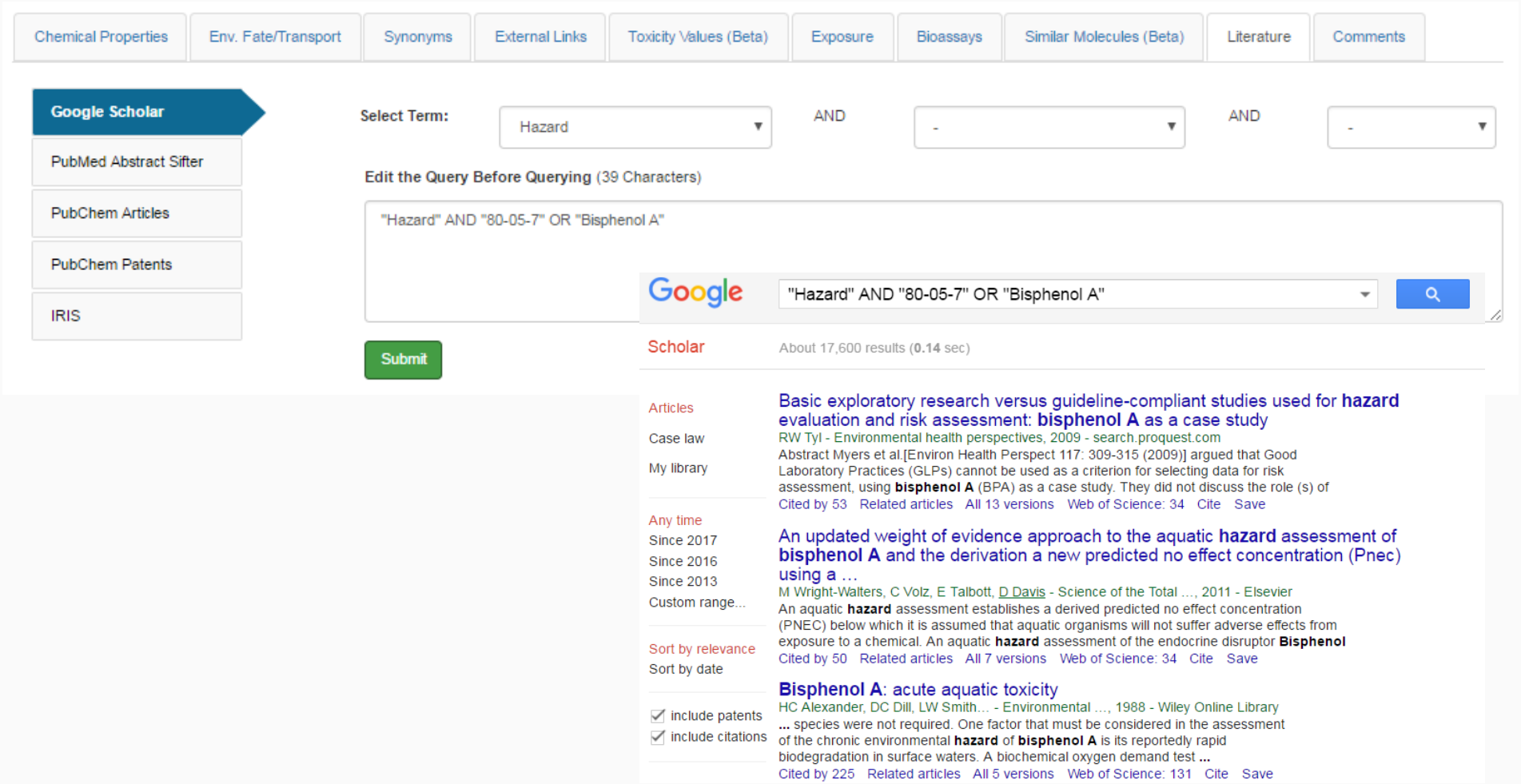
Download as: TSV | Excel Show: Inactive | Background

Assay Name	Hit Call	Top	Scaled Top	AC50	log AC50 ↓	Intended Target Family
APR_Hepat_CellLoss_48hr_dn	ACTIVE	1.41	1.41	120	2.08	cell cycle
APR_HepG2_MitoMass_24h_dn	ACTIVE	0.874	1.76	109	2.04	cell morphology
APR_HepG2_OxidativeStress_24h_up	ACTIVE	1.20	1.47	110	2.04	cell cycle
APR_Hepat_DNADamage_48hr_up	ACTIVE	1.84	1.14	107	2.03	cell morphology
APR_HepG2_CellLoss_24h_dn	ACTIVE	1.20	1.81	106	2.02	cell cycle
APR_HepG2_OxidativeStress_72h_up	ACTIVE	1.80	1.65	106	2.02	cell cycle
ATG_HSE_CIS_up	ACTIVE	1.59	3.38	102	2.01	dna binding

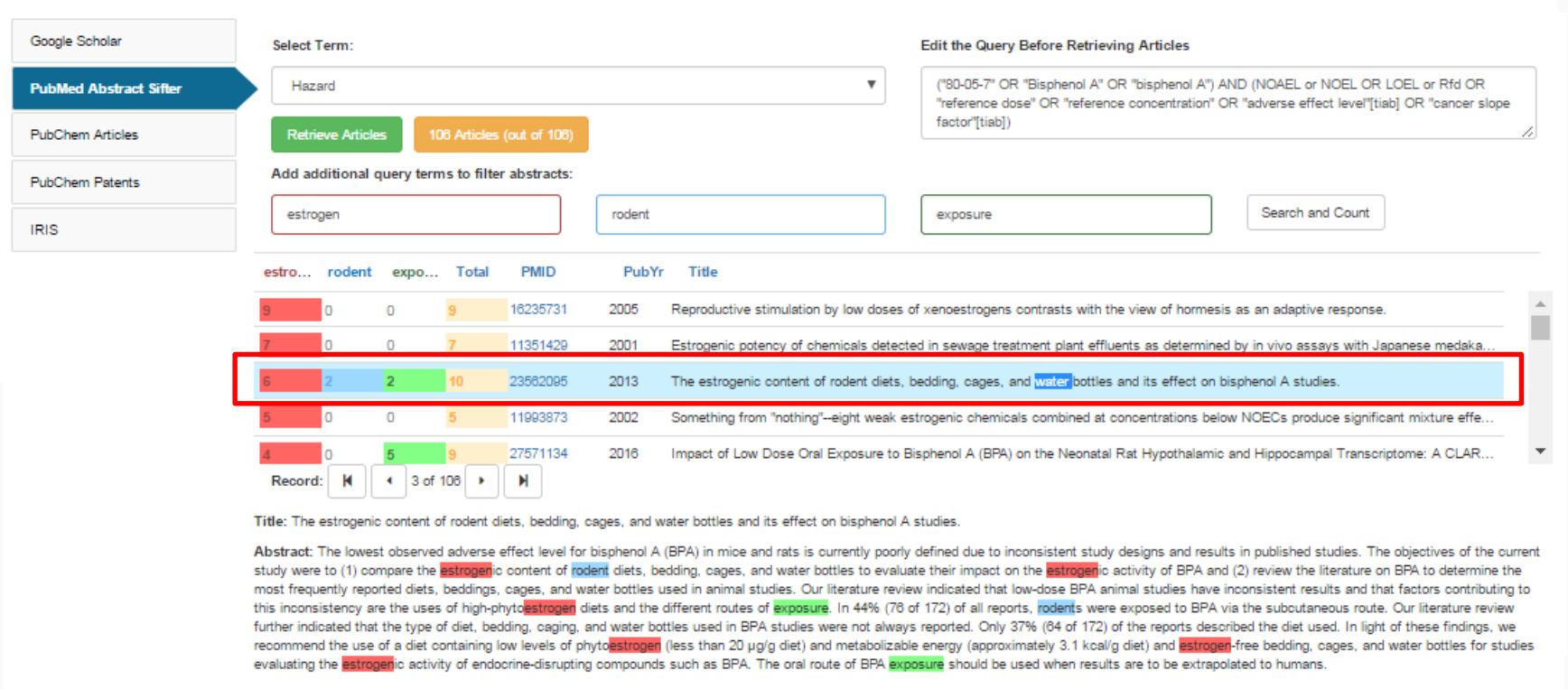
PubChem Bioassay Data Integration



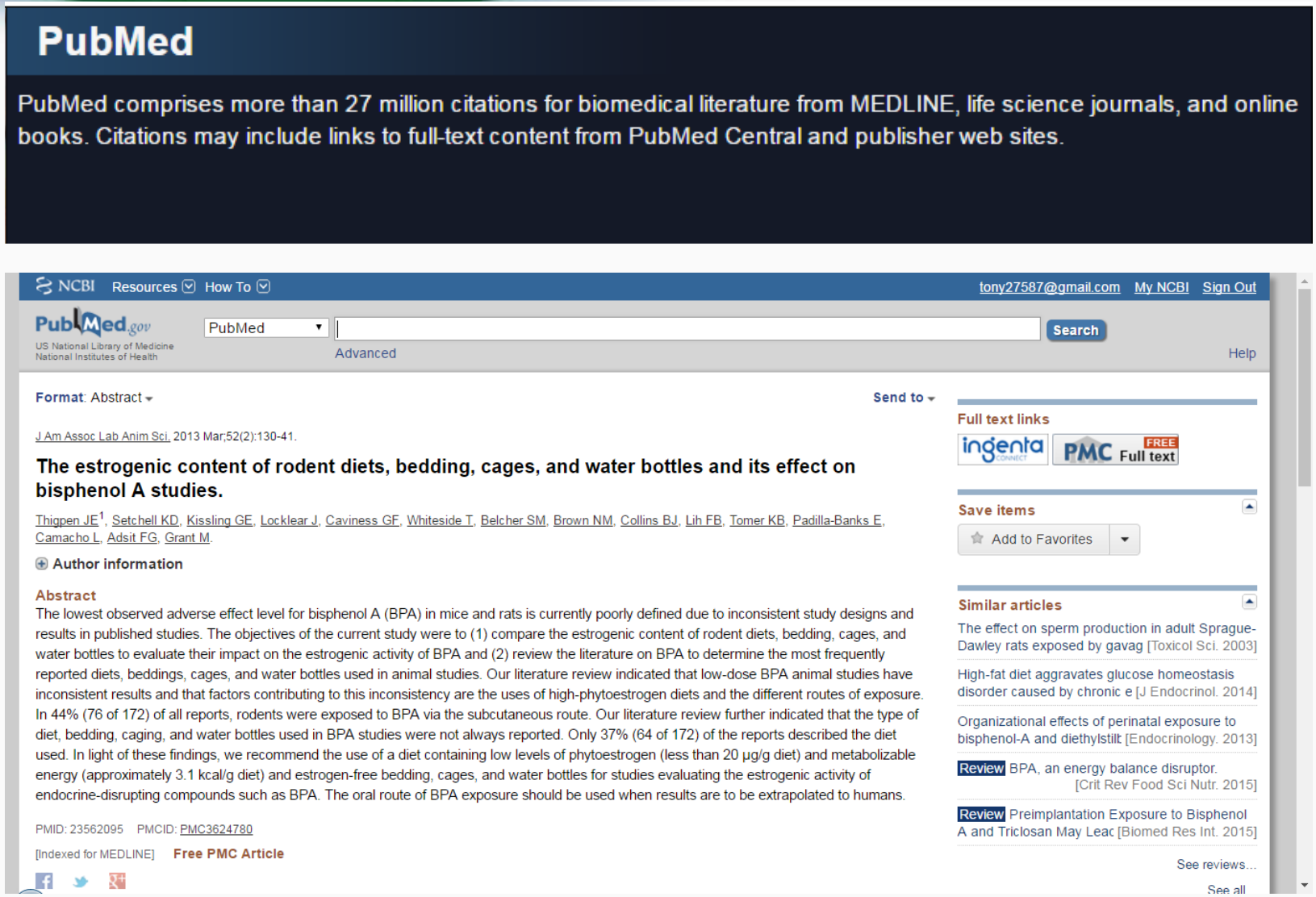
Literature Search Integration
Google Scholar



PubMed: BIG DATA Literature Search Integration



Linked Directly to PubMed



Answering Questions

- I have a 1000 CAS Numbers (or Names) – is there data available?
  - Has any Toxcast data been run?
  - Are there Toxicity Data values available?
  - Are there predicted exposure data (via Expocast)?
  - Can I get predicted physchem data for my model?
- Identifying chemicals by analytical chemistry

Mass Spectrometry Analyses

- Targeted Analysis:
  - We know exactly what we're looking for
  - 10s – 100s of chemicals
- Suspect Screening Analysis (SSA):
  - We have chemicals of interest
  - 100s – 1,000s of chemicals
- Non-Targeted Analysis (NTA):
  - We have no preconceived lists
  - 1,000s – 10,000s of chemicals
  - In dust, soil, food, air, water, products, etc.

NTA Applications

- Exposure surveillance
  - What chemicals are in food, products, dust, blood, etc.?

- Chemical prioritization
  - What are relevant chemicals & mixtures?

- Exposure forensics
  - What are chemical signatures of exposure sources?

- Effect-directed analysis
  - What are the biologically active chemicals in complex mixtures?

- Biomarker discovery
  - What chemicals are predictive of bioactivity/health impairment?

Batch Searching - Formulae

Select Input Type(s)

- [ ] Chemical Name
- [ ] CAS-RN
- [ ] InChIKey
- [ ] DSSTox Substance ID
- [x] Exact Molecular Formula

Include top 3 hits

Enter Identifiers to Search

C12H25NO2
C11H21NO3
C10H9N5O
C11H21NOS
C10H14ClNO2
C12H9NO3
C12H13N3O
C9H18N3OP
C8H14ClN5

Select Output Format

Excel

Customize Results

Chemical Identifiers

- [x] Chemical Name
- [x] DTXSID
- [x] CAS-RN
- [ ] InChIKey
- [x] IUPAC Name

Structures

- [ ] Mol File
- [x] SMILES
- [ ] InChI String

Intrinsic Properties

- [ ] Molecular Formula
- [ ] Average Mass
- [ ] Monoisotopic Mass

Metadata

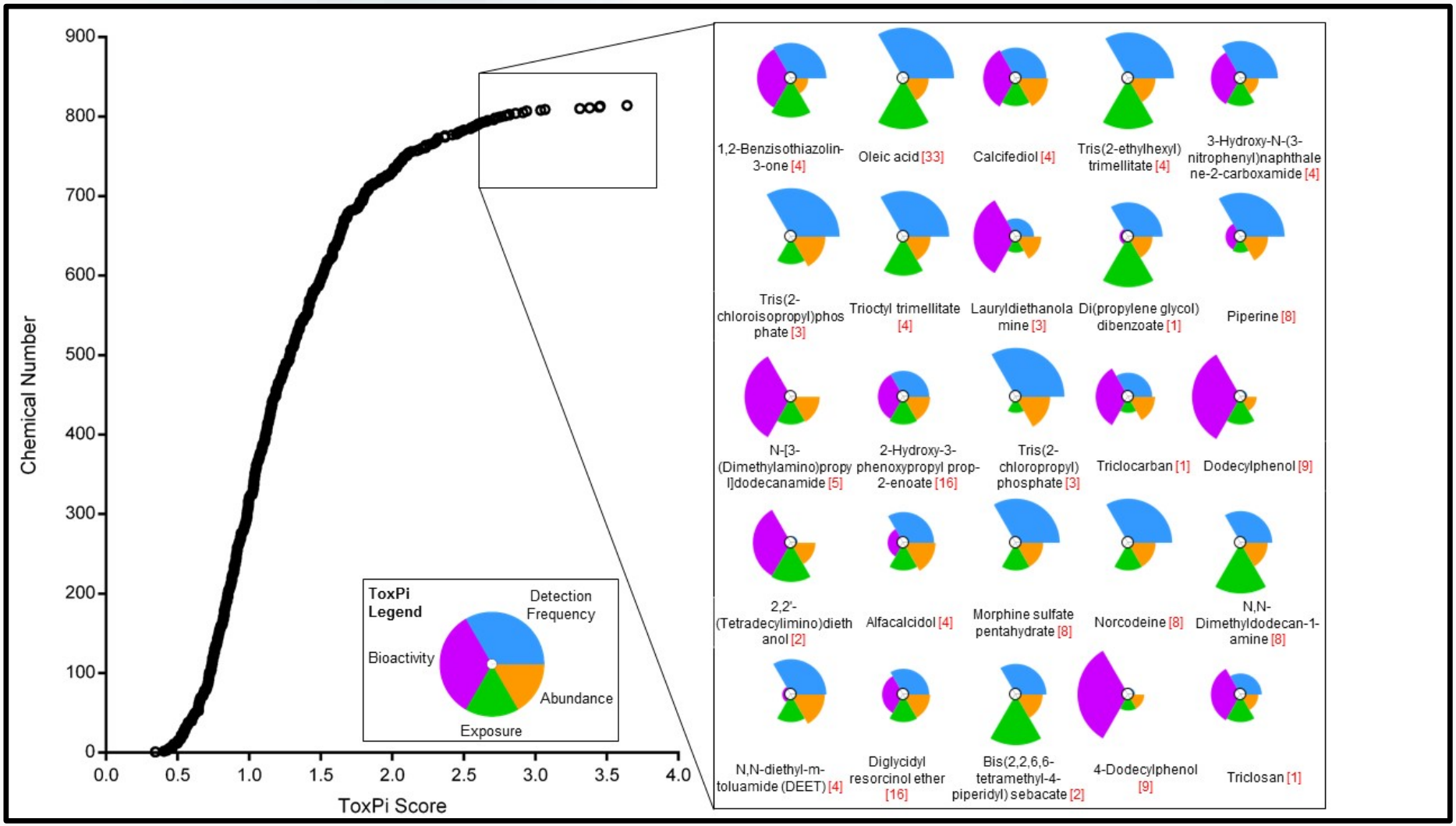
- [ ] Curation Level Details
- [x] Data Sources

Batch Searching - Formulae

- Uses data source ranking
- Production volume ranking in development

	A	B	C	D	E	F	G
1	INPUT	DTXSID	CASRN	PREFERRED NAME	MOL FORMULA	MONOISOTOPIC MASS	DATA SOURCES
2	C14H22N2O3	DTXSID2022628	29122-68-7	Atenolol	C14H22N2O3	266.163042576	46
3	C14H22N2O3	DTXSID0021179	6673-35-4	Practolol	C14H22N2O3	266.163042576	32
4	C14H22N2O3	DTXSID4048854	841-73-6	Bucolome	C14H22N2O3	266.163042576	20
5	C14H22N2O3	DTXSID1045407	13171-25-0	Trimetazidine dihydrochloride	C14H24Cl2N2O3	338.116398	19
6	C14H22N2O3	DTXSID0045753	56715-13-0	R-(+)-Atenolol	C14H22N2O3	266.163042576	19
7	C14H22N2O3	DTXSID2048531	5011-34-7	Trimetazidine	C14H22N2O3	266.163042576	14
8	C14H22N2O3	DTXSID10239405	93379-54-5	Esatenolol	C14H22N2O3	266.163042576	12
9	C14H22N2O3	DTXSID50200634	52662-27-8	N-(2-Diethylaminoethyl)-2-(4-hydroxyphenoxy)acetamide	C14H22N2O3	266.163042576	7
10	C14H22N2O3	DTXSID4020111	51706-40-2	dl-Atenolol hydrochloride	C14H23ClN2O3	302.1397203	6
11	C14H22N2O3	DTXSID1068693	51963-82-7	Benzenamine, 2,5-diethoxy-4-(4-morpholinyl)-	C14H22N2O3	266.163042576	5
12	C18H34N2O6S	DTXSID3023215	154-21-2	Lincomycin	C18H34N2O6S	406.213757997	35
13	C18H34N2O6S	DTXSID7047803	859-18-7	Lincomycin hydrochloride	C18H35ClN2O6S	442.1904357	22
14	C18H34N2O6S	DTXSID20849438	1398534-62-7	PUBCHEM_71432748	C18H35ClN2O6S	442.1904357	1
15	C10H12N2O	DTXSID1047576	486-56-6	Cotinine	C10H12N2O	176.094963014	40
16	C10H12N2O	DTXSID8075330	50-67-9	Serotonin	C10H12N2O	176.094963014	22
17	C10H12N2O	DTXSID8044412	2654-57-1	4-Methyl-1-phenylpyrazolidin-3-one	C10H12N2O	176.094963014	18
18	C10H12N2O	DTXSID80165186	153-98-0	Serotonin hydrochloride	C10H13ClN2O	212.0716407	11
19	C10H12N2O	DTXSID2048870	29493-77-4	(4R,5S)-4-methyl-5-phenyl-4,5-dihydro-1,3-oxazol-2-amine	C10H12N2O	176.094963014	10
20	C10H12N2O	DTXSID10196105	443-31-2	6-Hydroxytryptamine	C10H12N2O	176.094963014	9
21	C10H12N2O	DTXSID90185693	31822-84-1	1,4,5,6-Tetrahydro-5-phenoxypyrimidine	C10H12N2O	176.094963014	7
22	C10H12N2O	DTXSID40178777	2403-66-9	2-Benzimidazolepropanol	C10H12N2O	176.094963014	7
23	C10H12N2O	DTXSID80157026	13140-86-8	N-Cyclopropyl-N'-phenylurea	C10H12N2O	176.094963014	6
24	C10H12N2O	DTXSID30205607	570-14-9	4-Hydroxytryptamine	C10H12N2O	176.094963014	6
25	C14H18N4O3	DTXSID5023900	17804-35-2	Benomyl	C14H18N4O3	290.137890456	68
26	C14H18N4O3	DTXSID3023712	738-70-5	Trimethoprim	C14H18N4O3	290.137890456	51
27	C14H18N4O3	DTXSID40209671	60834-30-2	Trimethoprim hydrochloride	C14H19ClN4O3	326.1145682	8
28	C14H18N4O3	DTXSID70204210	55687-49-5	Benzenemethanol, 4-((2,4-diamino-5-pyrimidinyl)methyl)-2,	C14H18N4O3	290.137890456	5
29	C14H18N4O3	DTXSID20152671	120075-57-2	6-Methoxy-4-(3-(N,N-dimethylamino)propylamino)-5,8-quin	C14H18N4O3	290.137890456	4
30	C14H18N4O3	DTXSID30213742	63931-79-3	1H-1,2,4-Benzotriazepine-3-carboxylic acid, 4,5-dihydro-4-	C14H18N4O3	290.137890456	3
31	C14H18N4O3	DTXSID30219608	69449-07-6	2,4-Pyrimidinediamine, 5-((3,4,5-trimethoxyphenyl)methyl)-	C14H20N4O4	308.14845514	3
32	C14H18N4O3	DTXSID20241155	94232-27-6	L-Aspartic acid, compound with 5-((3,4,5-trimethoxyphenyl	C18H25N5O7	423.175398165	3
33	C14H18N4O3	DTXSID80241156	94232-28-7	L-Glutamic acid, compound with 5-((3,4,5-trimethoxypheny	C19H27N5O7	437.191048229	3

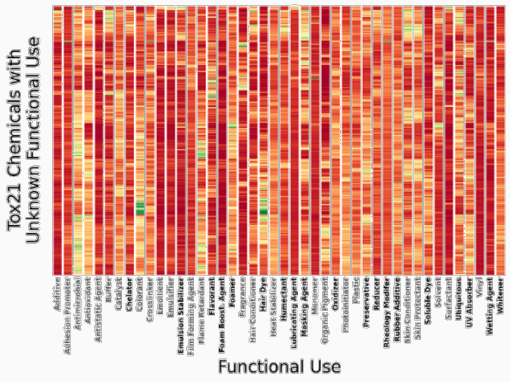
Chemical Prioritization: Dust Case Study

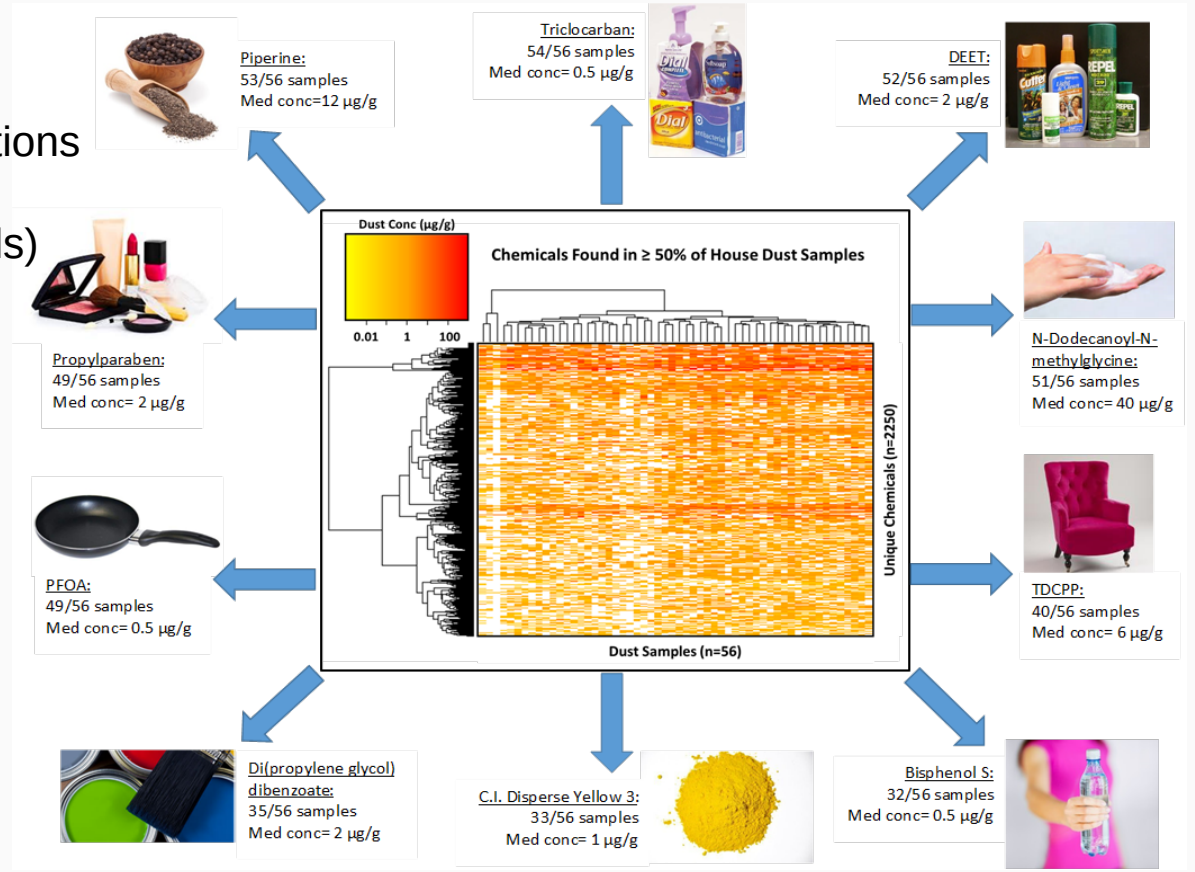


Rager et al., *Environ. Int.,* v88, March 2016, pp 269–280. DOI: 10.1016/j.envint.2015.12.008

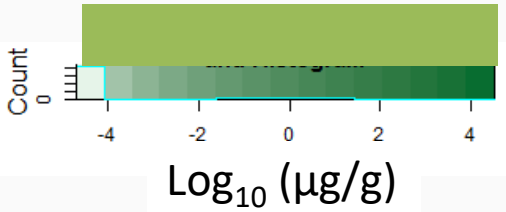
Suspect Screening Results: Dust Case Study

- Chemical confirmation from ToxCast
- Detection frequency
- External calibration for concentrations
- Cluster analysis (homes/chemicals)
- Chemical functional use





Improving Exposure Estimates and Characterization



- Air freshener
- Baby soap
- Carpet
- Carpet padding
- Cereals
- Cotton clothing
- Deodorant
- Fabric upholstery
- Glass cleaners
- Hand soap
- Indoor house paint
- Lipstick
- Plastic children's toys
- Shampoo
- Shaving cream
- Shower curtain
- Skin lotion
- Sunscreen
- Toothpaste
- Vinyl upholstery

- **GCXGC-MS with DCM Extraction**
- **1606 tentative and confirmed chemical identifications**

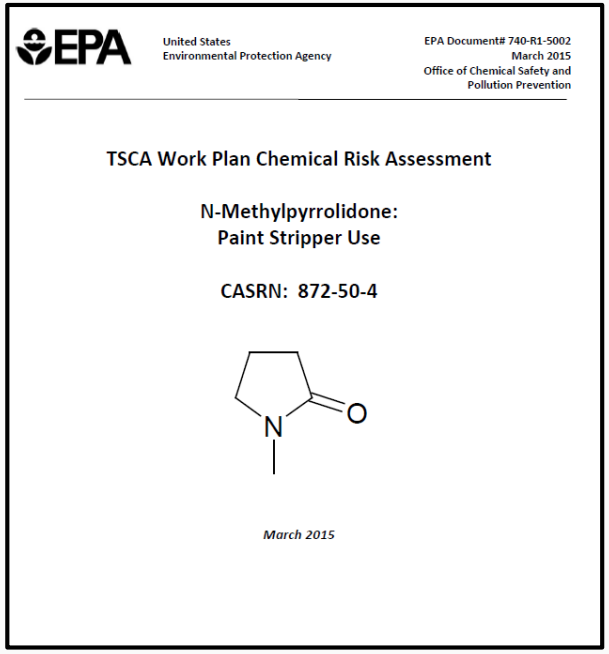
- Common Chemical (n>19)
- ToxCast
- Flame Retardant
- Potent ER

Wambaugh et al. Unpublished

“RapidTox” in Development

- Semi-automated **decision support tool** for high-throughput risk assessments
- Use the Dashboard “architecture” and data streams and add appropriate functionality
- Combine diverse data streams into **quantitative toxicity values** with uncertainty estimates
- Test out new prototype capabilities for internal users – e.g. GenRA (Generalized Read Across)

Risk Assessments Generally Contain a Standard Set of Components



EPA United States Environmental Protection Agency

EPA Document# 740-R1-5002
March 2015
Office of Chemical Safety and Pollution Prevention

TSCA Work Plan Chemical Risk Assessment

N-Methylpyrrolidone: Paint Stripper Use

CASRN: 872-50-4

March 2015

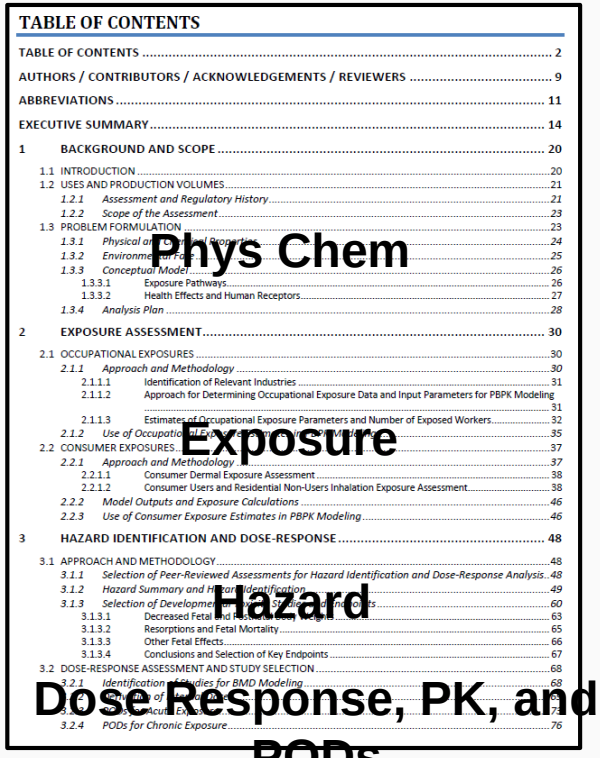
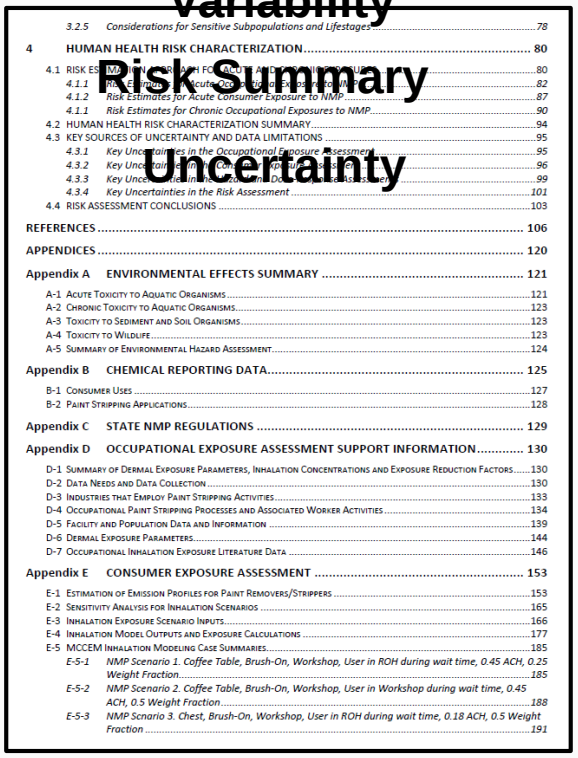


TABLE OF CONTENTS



Phys Chem

Exposure

Hazard

Dose Response, PK, and PODs

Variability

Risk Summary

Uncertainty

We are assembling these components to deliver RapidTox

Problem #1 – SuperFund site

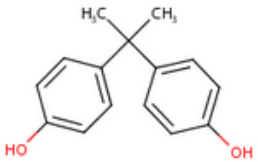
- At a SuperFund site, EPA has identified 600 unique chemicals
- 300 of these have “good” reference doses/concentration (RfD/RfC), but 300 do not
- Can we determine “good enough” RfD/RfC to aid in cleanup planning?
- Can we do it “quickly”?

Problem #2 – Prioritize inerts

- Office of Pesticide Programs (OPP) petitioned to perform **risk assessments** on 100s of pesticidal "inert ingredients"
- Companies have **not** typically been required to submit ***in vivo* data** on individual inerts
- Can we **prioritize** which of these chemicals should be the focus of detailed risk assessments?

RapidTox Dashboard in development
Risk Assessment Tool

RapidTox Report: Bisphenol A



Property	Value
CASRN	80-05-7
Preferred Name	Bisphenol A
IUPAC Name	4,4'-propane-2,2-diyldiphenol
SMILES	CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1
InCHI Key	IISBACLAFKSPIT-UHFFFAOYSA-N
Formula	C15H16O2

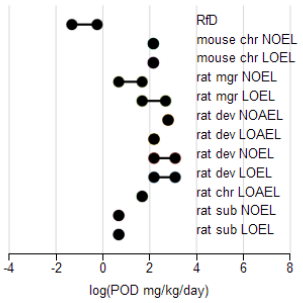
External Links

- ACToR
- Chemistry Dashboard
- RapidTox Dashboard
- EDSP21 Dashboard
- ToxCast Dashboard
- CPCat

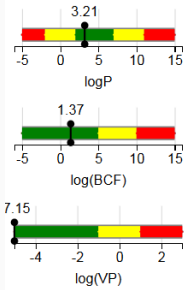
Executive Summary

- **Quantitative Risk Assessment Value**
  - IRIS values available
  - EPA RSL values available
  - Minimum RfD: 0.050 mg/kg-day (IRIS, oral, -, -)
  - Minimum oral POD: 5.0 mg/kg-day (ToxRefDB, LOEL, sub, rat)
  - No RfC calculated
  - No inhlation POD values
  - HTTK Css(95%)= 1.66, Css(median)=0.569 um/(mg/kg-day)
  - IVIVE POD = 0.0536 to 0.156 mg/kg-day, based on NVS_NR_hCAR_Antagonist AC50=0.089 uM
  - Genes with activity below 1 uM: ESR1, ESR2, Esr1, PPARA, NR1I3, ESR1, NR1I2, MMP3, Cyp2c11, Tpo

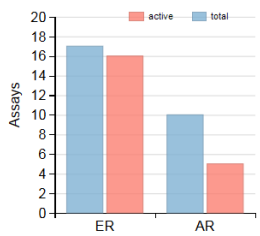
Oral PODs



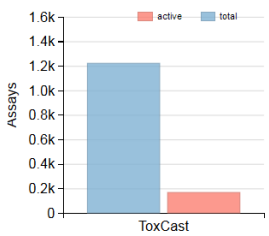
Summary Plots



EDSP



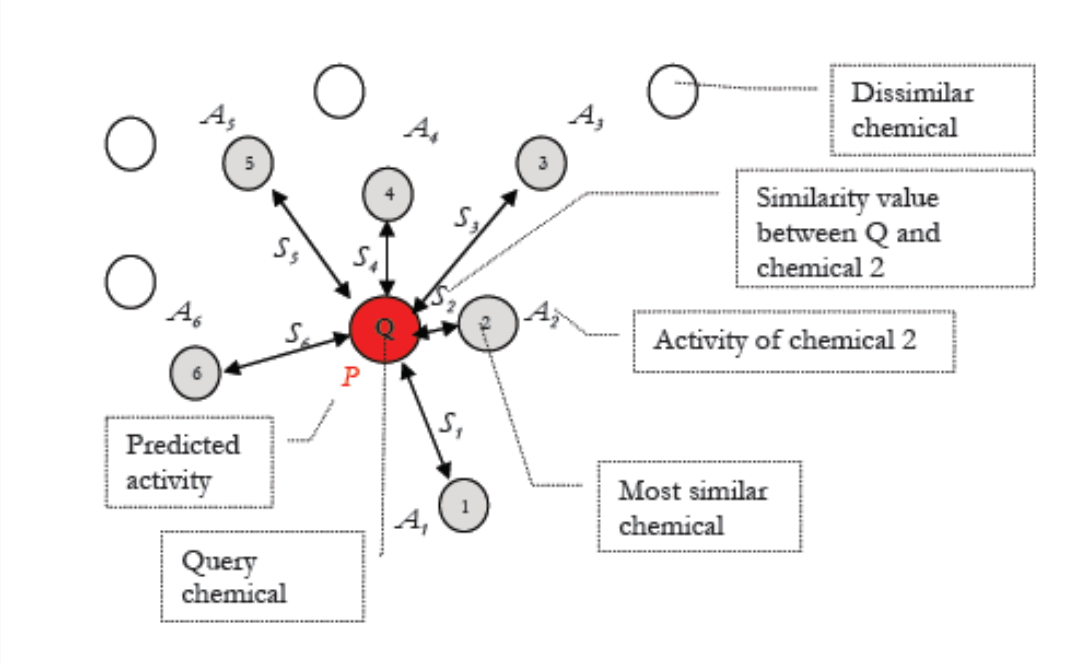
ToxCast



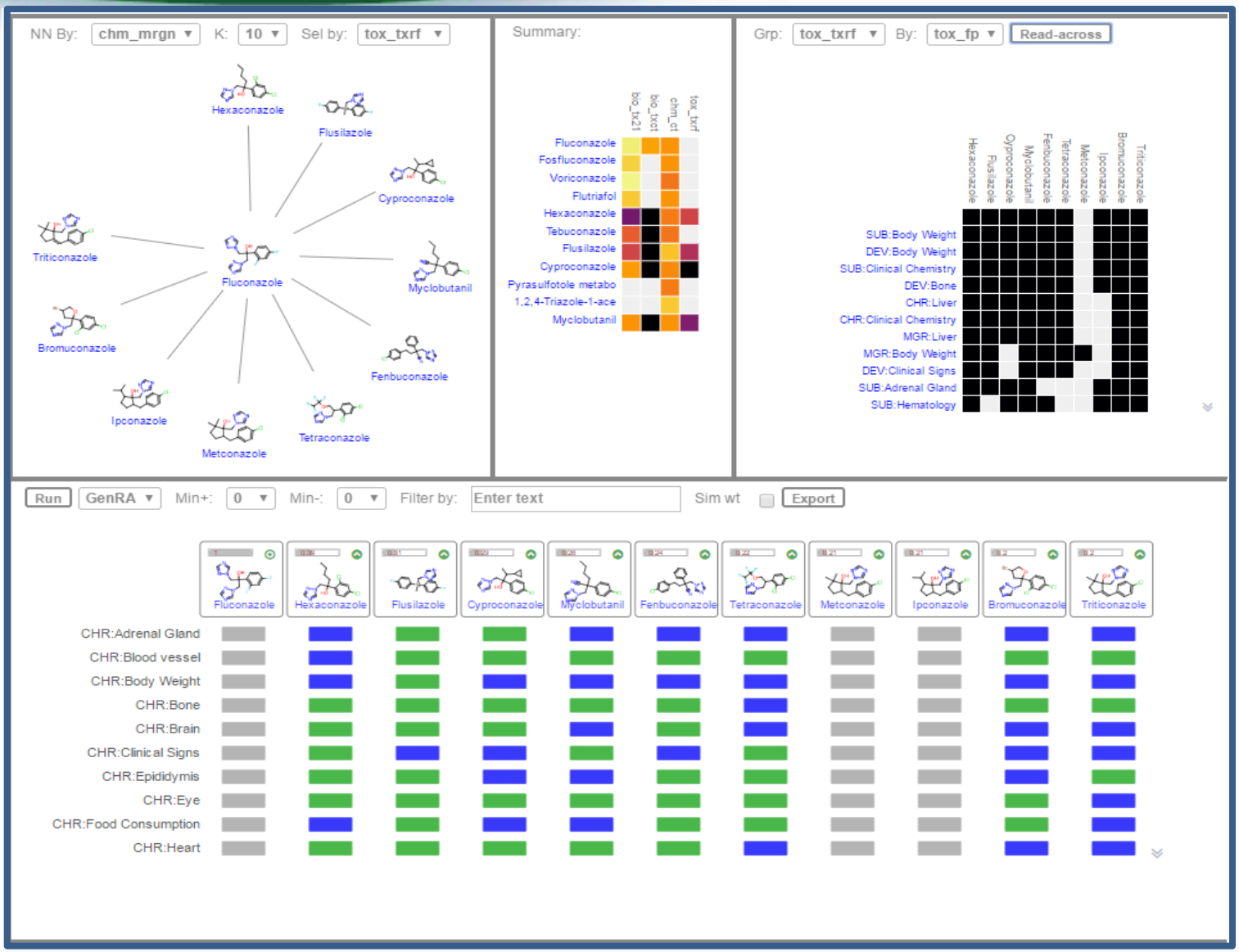
Quantifying Uncertainty & Assessing Performance of Read-Across - GenRA

- "**Chemical-Biological Read-Across**" - predict toxicity as a similarity-weighted activity of nearest neighbors

- Evaluates read-across performance and uncertainty using available data



Work-in-progress



Deliver Data and Models for Reuse

Toxicity ForeCaster (ToxCast™) Data

EPA's most updated, publicly available high-throughput toxicity data on thousands of chemicals. This data is generated through the EPA's ToxCast resesarch effort. ToxCast is part of the Toxicology in the 21st Century (Tox21) federal collaboration. All data is available for download and includes the following data sets. The release date and version names for the data sets are provided in the table below.

As part of EPA's commitment to share data, all of the computational toxicology data is publicly available for anyone to access and use. EPA's computational toxicology data is considered "open data", and thus all of the data below are free of all copyright restrictions, and fully and freely available for both non-commercial and commercial use.

- **ToxCast & Tox21 Chemicals**: A list of all chemicals screened, along with descriptions.
- **ToxCast & Tox21 High-throughput Assays**: Information about the hundreds of assays used to screen the chemicals.
- **ToxCast & Tox21 Summary Data**: Summary hit calls for all chemicals and concentration response curves for all assays.
- **MySQL Database**: Downloadable database that provides access to EPA's analysis of the ToxCast and Tox21 high-throughput screening data.
- **R Package**: The computer programming package used to process and model all EPA ToxCast and Tox21 chemical screening data.
- **Data Spreadsheet:** Spreadsheet that provides access to EPA's analysis of the ToxCast and Tox21 high-throughput screening data.
- **Concentration Response Curves:** Concentration response curves for all ToxCast & Tox21 assays.
- **Collaborative Estrogen Receptor Activity Prediction Project Data**: Data and supplemental files from CERAPP, a large-scale modeling project.

https://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data

Open Data means Reuse

Chemical Research in Toxicology

Big Data in Chemical Toxicity Research: The Use of High-Throughput Screening Assays To Identify Potential Toxicants

Hao Zhu,*[,†,‡] Jun Zhang,[†,‡] Marlene T. Kim,[†,‡] Abena Boison,[†] Alexander Sedykh,[‡] and Kimberlee Moran[§]

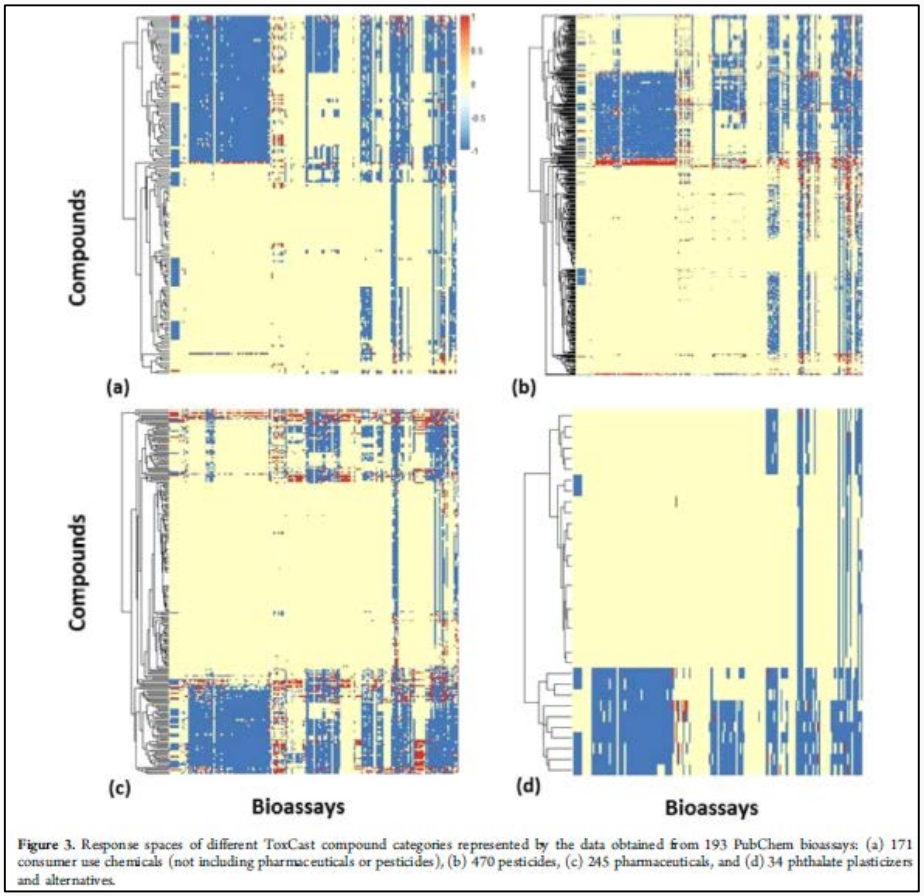
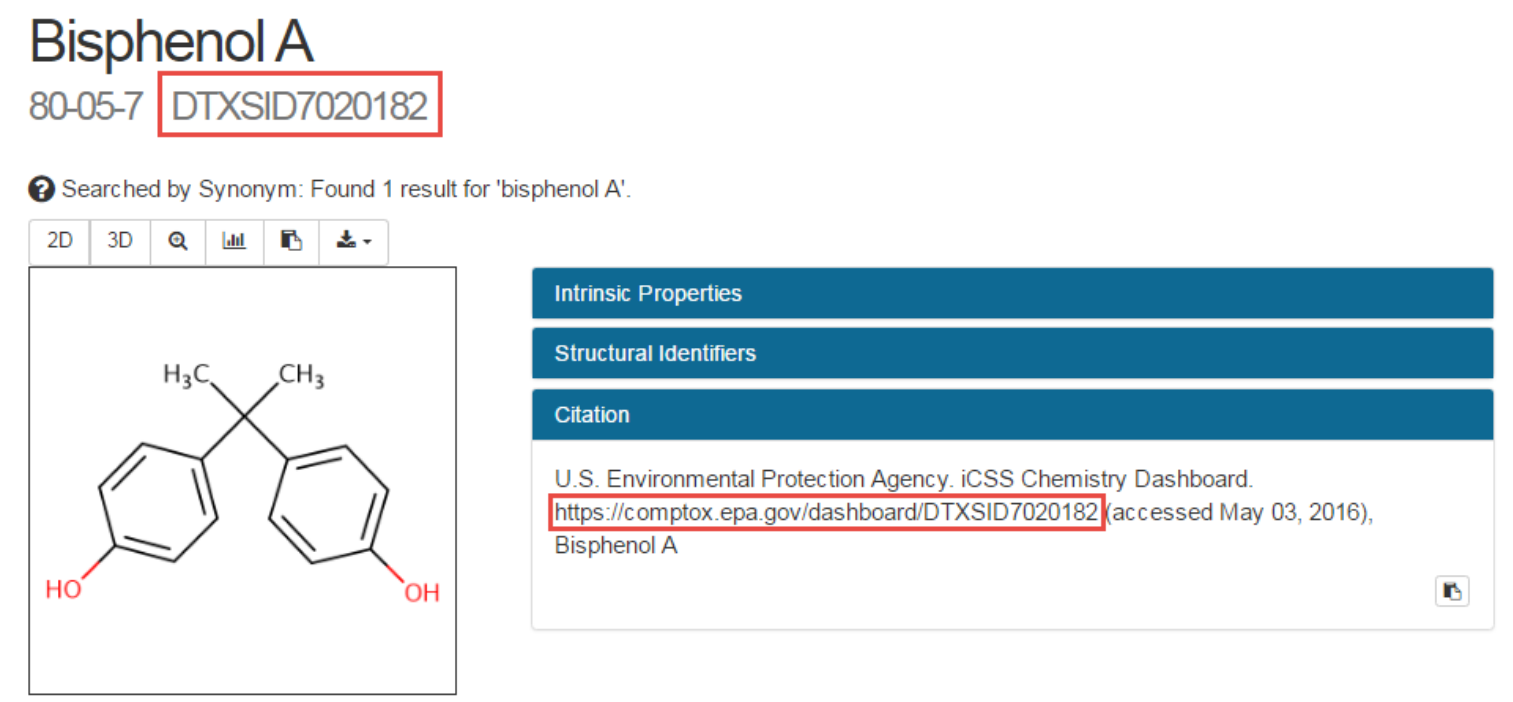


Figure 3. Response spaces of different ToxCast compound categories represented by the data obtained from 193 PubChem bioassays: (a) 171 consumer use chemicals (not including pharmaceuticals or pesticides), (b) 470 pesticides, (c) 245 pharmaceuticals, and (d) 34 phthalate plasticizers and alternatives.

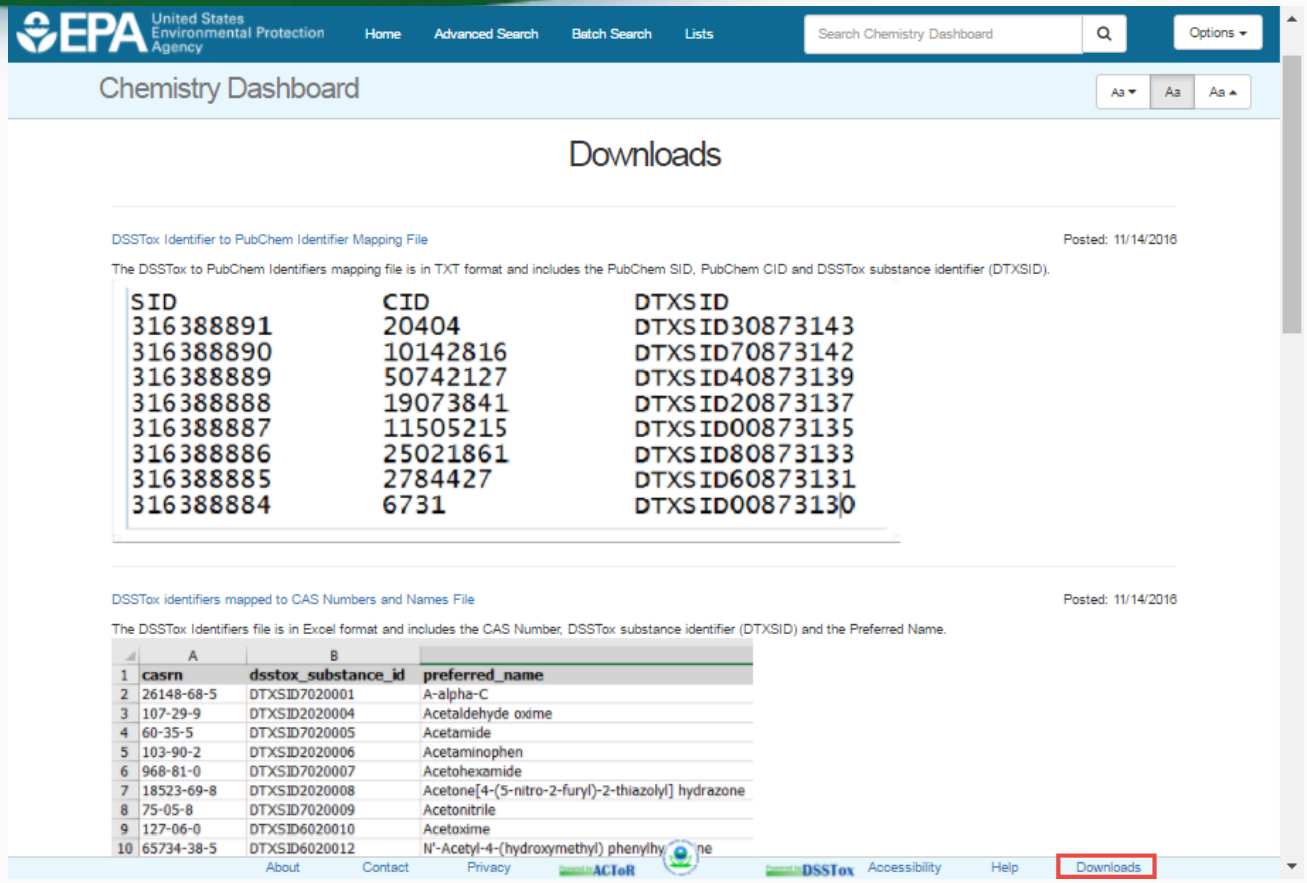
Delivering Data to the Community

- Linkages are simple: using the associated identifiers



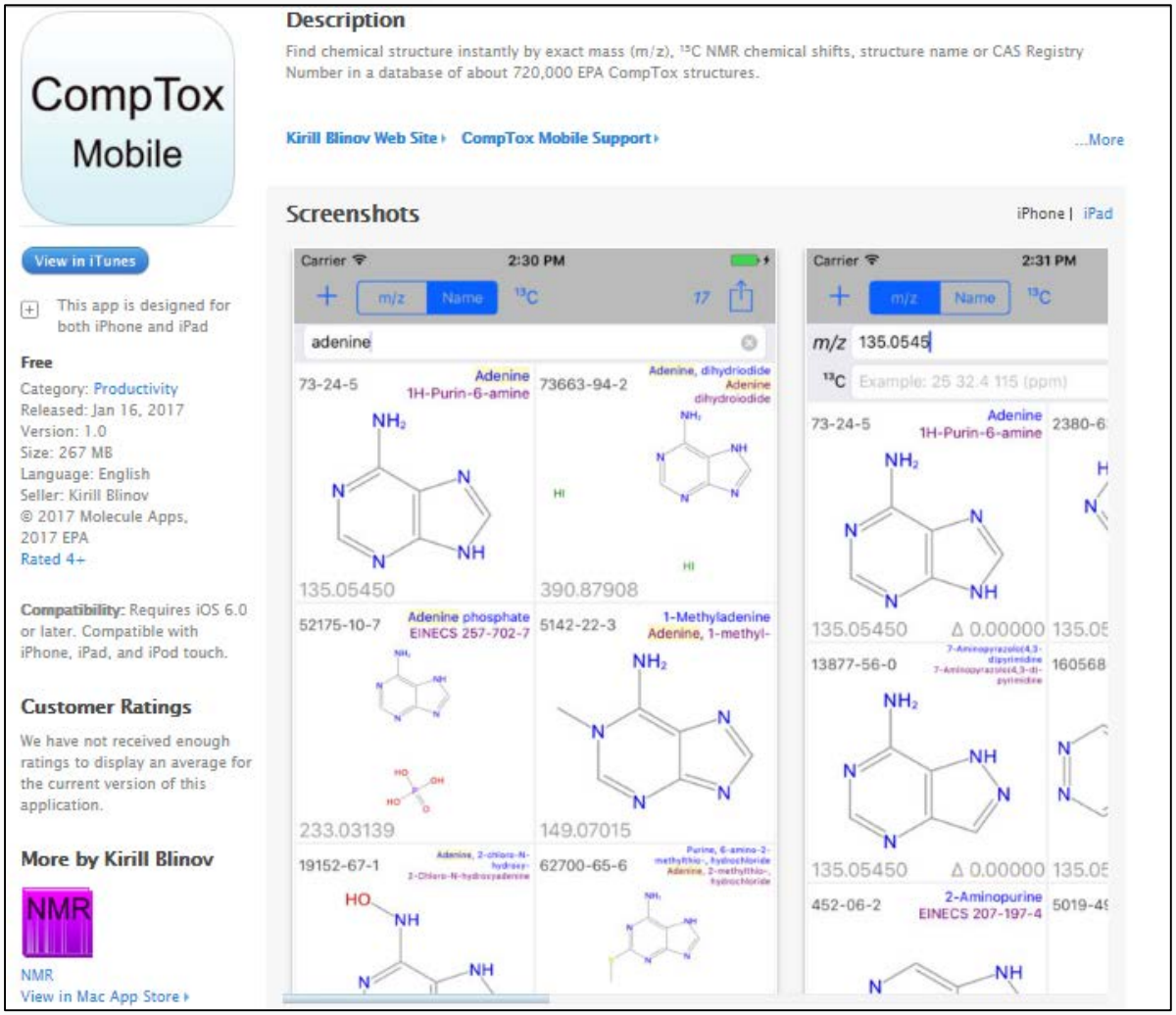
- Files of structures/identifiers mapped to DTXSIDs
- Integrated already - PubChem, EBI's UNICHEM, ChemSpider, ToxPlanet. Whoever wants the files…

Delivering our Chemistry Data

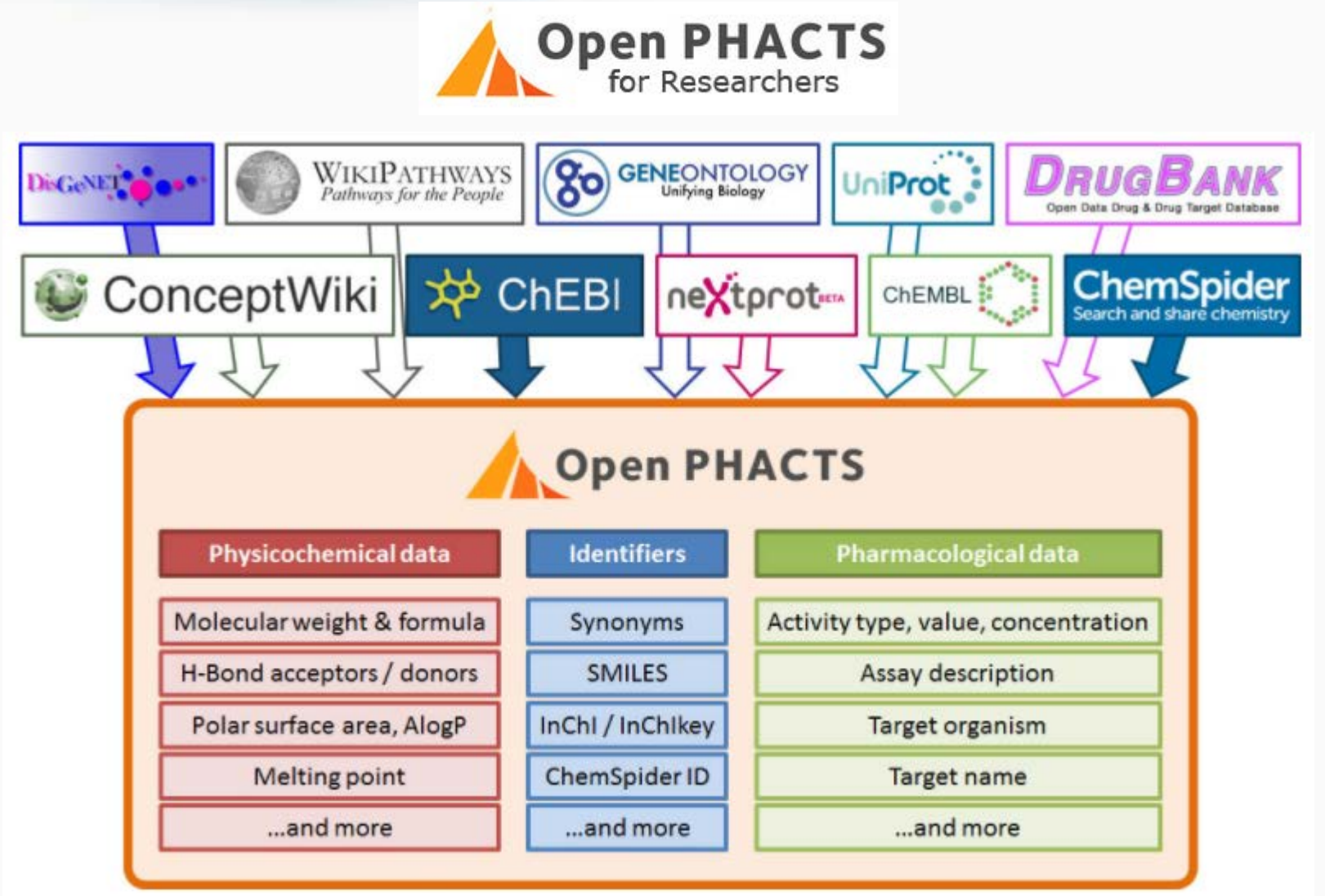


- Various types of data at FTP download site: ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard

Open Data Reuse on iOS



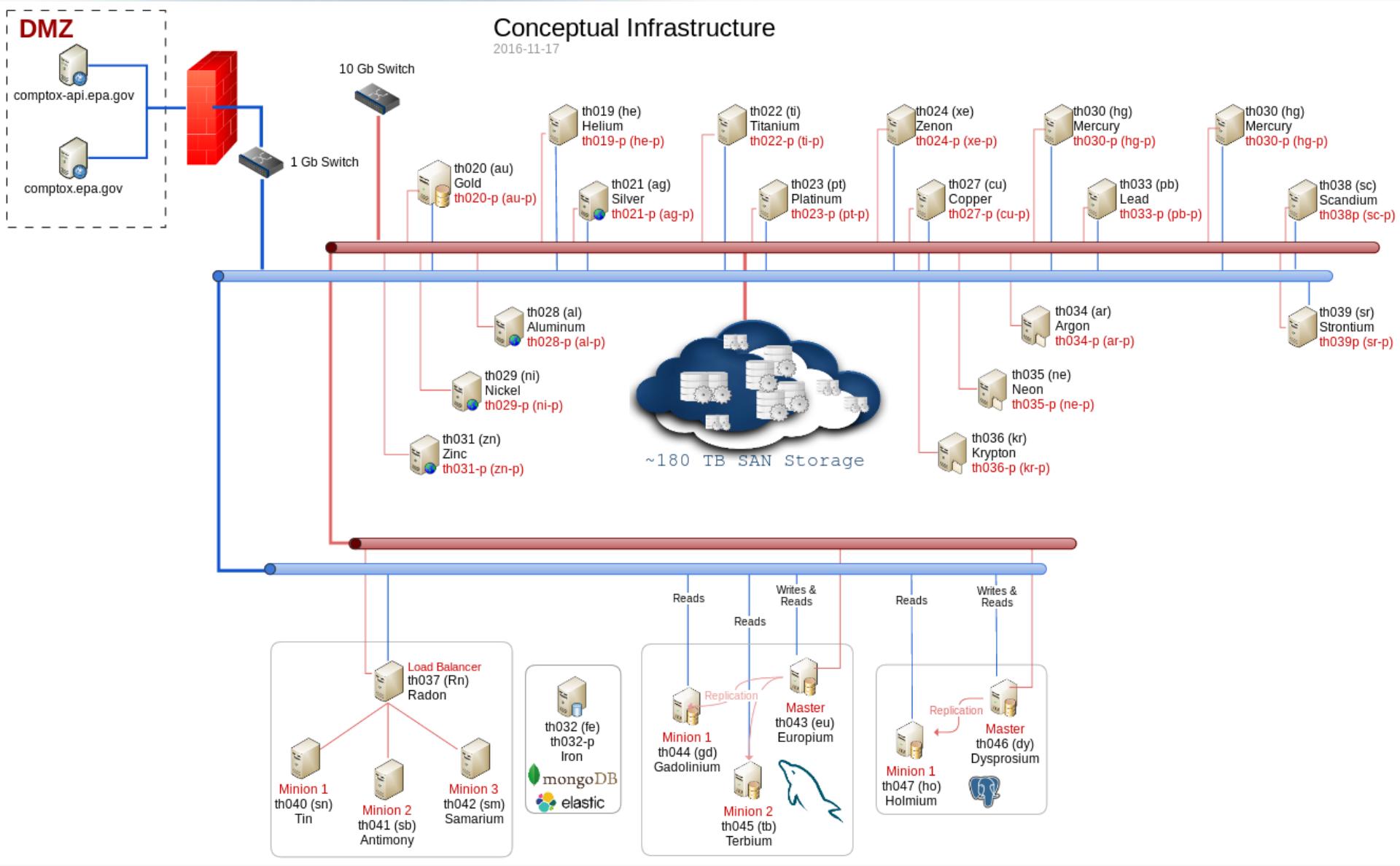
Delivering More Value from Our Data – TRIPLE STORE?



Computational Infrastructure and Processes

- 10 years of development – heterogeneous
  - MySQL, Java, R-code, Matlab, multiple Javascript environments, 100s of web services
  - Multiple projects, loose connectivity
- Migration to homogeneity for production Apps
  - MySQL
  - Java
  - JQuery/JavaScript
  - Versioned microservices
- Sprint methodology with Atlassian management tools
  - **Confluence** – knowledge management
  - **Jira** – ticketing/stories (work assignments)
  - **Fisheye** – collaborative peer code review
  - **Bitbucket** – code version control system

Computational Infrastructure

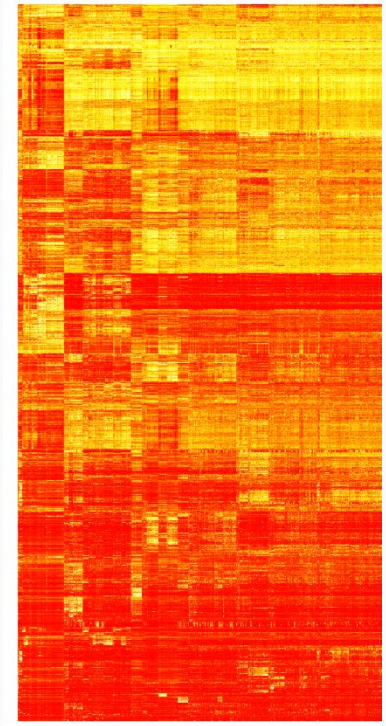


The NCCT Team

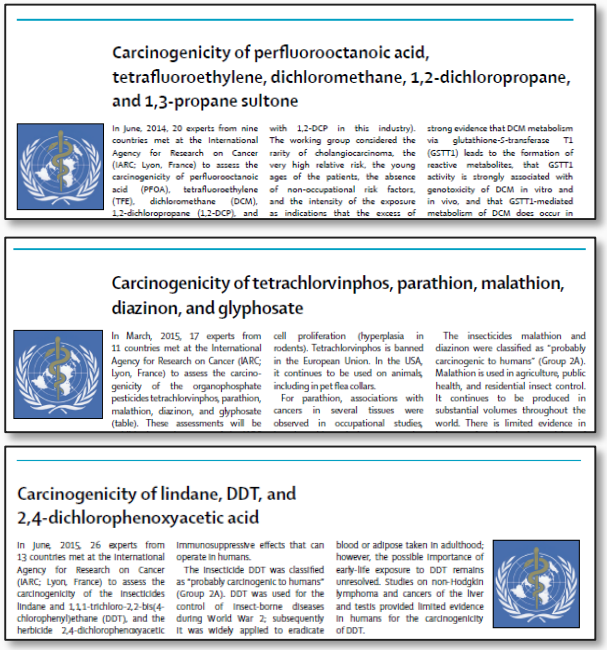
- Our team is broad in skills
  - Biologists, chemists and toxicologists
  - Bioinformaticians and cheminformaticians - modelers (QSAR, Deep Learning), text miners
  - Information technology team, software developers
  - A large IT support team for production applications (National Computing Center)

Many Successes from Our Approaches

Group Chemicals by Similar Bioactivity and Predictive Modeling



Provide **Mechanistic Support** for Hazard ID



Carcinogenicity of perfluorooctanoic acid, tetrafluoroethylene, dichloromethane, 1,2-dichloropropane, and 1,3-propane sultone

In June, 2014, 20 experts from nine countries met at the International Agency for Research on Cancer (IARC; Lyon, France) to assess the carcinogenicity of perfluorooctanoic acid (PFOA), tetrafluoroethylene (TFE), dichloromethane (DCM), 1,2-dichloropropane (1,2-DCP), and with 1,2-DCP in this industry). The working group considered the rarity of cholangiocarcinoma, the very high relative risk, the young ages of the patients, the absence of non-occupational risk factors, and the intensity of the exposure as indications that the excess of strong evidence that DCM metabolism via glutathione-S-transferase T1 (GSTT1) leads to the formation of reactive metabolites, that GSTT1 activity is strongly associated with genotoxicity of DCM in vitro and in vivo, and that GSTT1-mediated metabolism of DCM does occur in

Carcinogenicity of tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate

In March, 2015, 17 experts from 11 countries met at the International Agency for Research on Cancer (IARC; Lyon, France) to assess the carcinogenicity of the organophosphate pesticides tetrachlorvinphos, parathion, malathion, diazinon, and glyphosate (table). These assessments will be cell proliferation (hyperplasia in rodents). Tetrachlorvinphos is banned in the European Union. In the USA, it continues to be used on animals, including in pet flea collars.

For parathion, associations with cancers in several tissues were observed in occupational studies, The insecticides malathion and diazinon were classified as "probably carcinogenic to humans" (Group 2A). Malathion is used in agriculture, public health, and residential insect control. It continues to be produced in substantial volumes throughout the world. There is limited evidence in

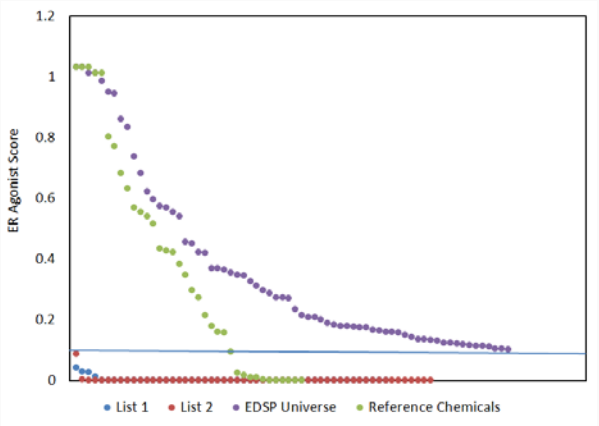
Carcinogenicity of lindane, DDT, and 2,4-dichlorophenoxyacetic acid

In June, 2015, 26 experts from 13 countries met at the International Agency for Research on Cancer (IARC; Lyon, France) to assess the carcinogenicity of the insecticides lindane and 1,1,1-trichloro-2,2-bis(4-chlorophenyl)ethane (DDT), and the herbicide 2,4-dichlorophenoxyacetic Immunosuppressive effects that can operate in humans.

The insecticide DDT was classified as "probably carcinogenic to humans" (Group 2A). DDT was used for the control of insect-borne diseases during World War 2; subsequently it was widely applied to eradicate blood or adipose taken in adulthood; however, the possible importance of early-life exposure to DDT remains unresolved. Studies on non-Hodgkin lymphoma and cancers of the liver and testis provided limited evidence in humans for the carcinogenicity of DDT.

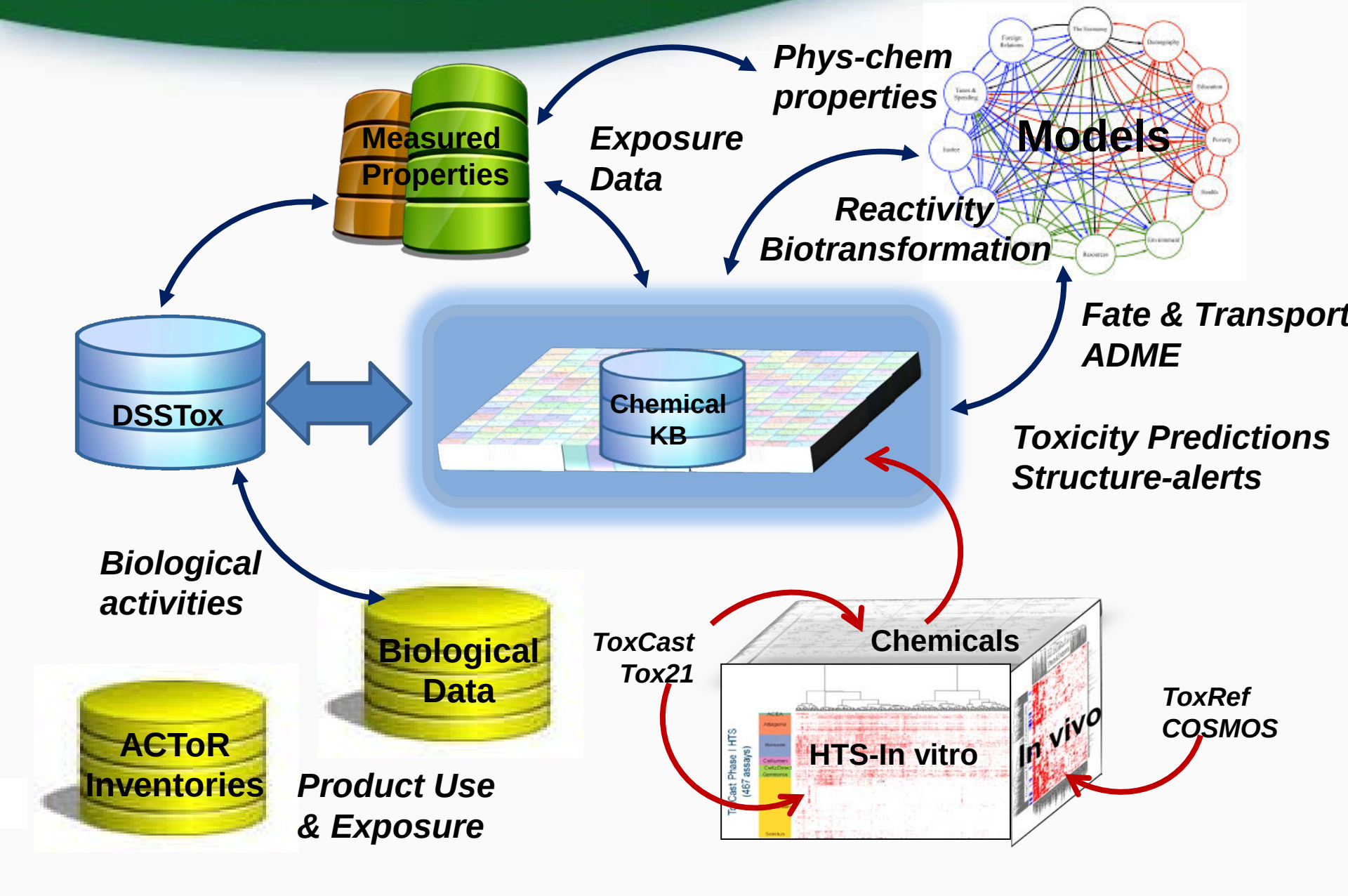
IARC Monographs 110, 112, 113

Prioritization of Chemicals for Further Testing

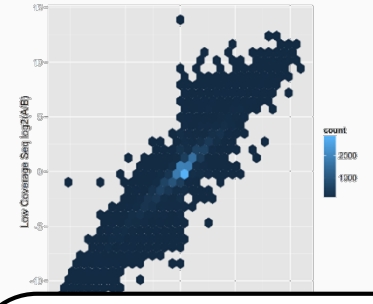


FIFRA SAP, Dec 2014

We're not done yet…

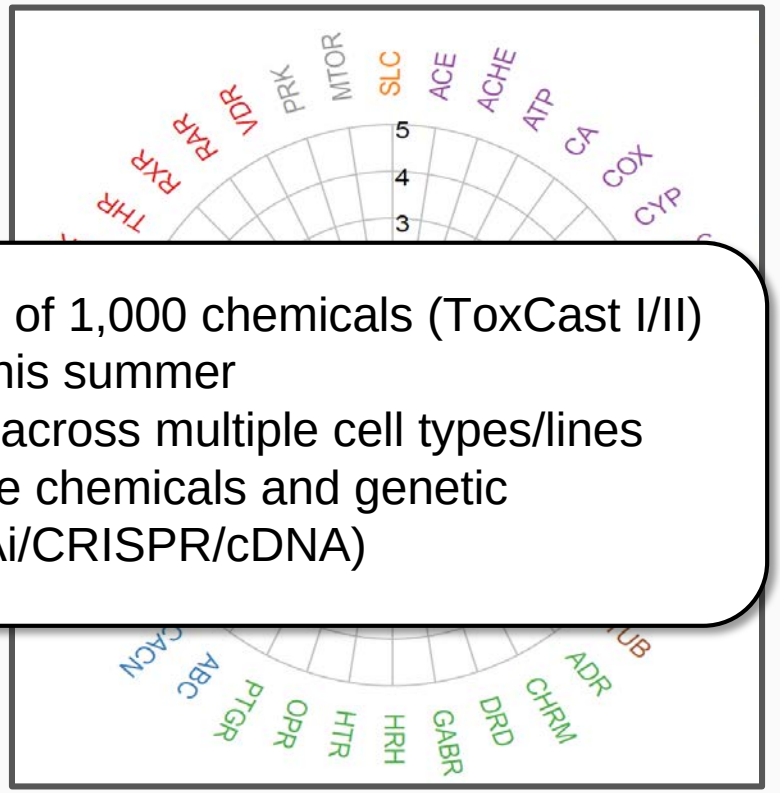


Transcriptomics Data will Deliver Terabytes of Data for Analysis



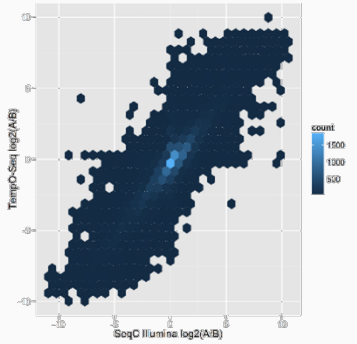
TruSeq
r^2 0.74

MOA Analysis Pipeline



COMING SOON!

- Large scale screen of 1,000 chemicals (ToxCast I/II) in single cell type this summer
- Additional screens across multiple cell types/lines
- Additional reference chemicals and genetic perturbations (RNAi/CRISPR/cDNA)



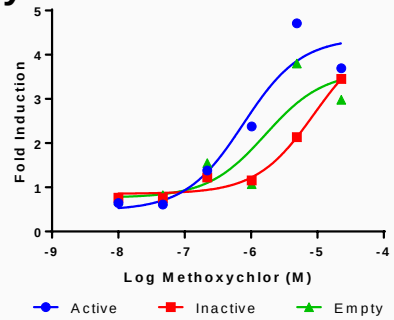
Low Coverage
r^2 0.83

Currently capable of assigning to >40 MOAs based on transcriptional responses

Beginning to Address Metabolic Competence

"Extracellular" Approach

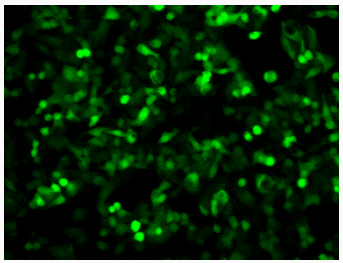
Chemicals metabolism in the media or buffer of cell-based and cell-free assays

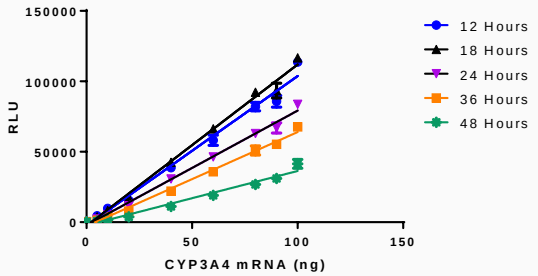


More closely models effects of hepatic metabolism and generation of circulating metabolites

"Intracellular" Approach

Capable of metabolizing chemicals inside the cell in cell-based assays





More closely models effects of target tissue metabolism

Integrated approach to model *in vivo* metabolic bioactivation and detoxification

Future Work

- Existing dashboards functionality to be added to the “CompTox ~~Chemistry~~ Dashboard”
- RapidTox functionality tested and released
- Open Application Programming Interface and web services
- Integration to other agency databases
- Deliver our data to the semantic web
  - Use accepted ontologies, chemical structures first

Acknowledgements

EPA-RTP

Russell Thomas
Kevin Crofton
John Cowden
Chris Grulke
Ann Richard
John Wambaugh
Grace Patlewicz
Imran Shah
Richard Judson
Jeff Edwards
Nancy Baker
NERL Colleagues

Contact

Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821