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

Incorporating new technologies and high-throughput approaches in the design and selection of chemical alternatives

Antony Williams

U.S. Environmental Protection Agency, RTP, NC

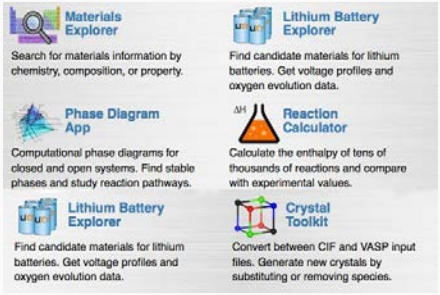
June 15th, 2017

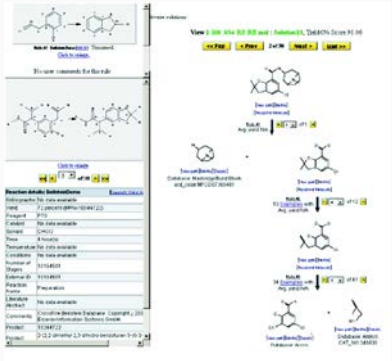
21st Annual Green Chemistry & Engineering Conference, Reston, VA

Paradigm Shift in Chemical Design (especially in regards to toxicology)

EPA United States Environmental Protection Agency

- Chemical Design in the future
  - High-throughput modeling
  - SAR/QSAR/QSTR models
  - Retrosynthetic analysis algorithms for synthesis





- Toxicology
  - FEWER animal studies. Maybe NONE??
  - Openness in data sharing
  - More confidence in computational predictive models that link chemical structures to adverse outcomes

It will depend on lots of DATA

Chemistry International

The News Magazine of IUPAC

Volume 39, Issue 3 (Jul 2017)

‹ Previous Article Next Article ›

The Future of Chemical Information Is Now

Antony J. Williams / Harry E. Pence

Published Online: 2017-05-24 | DOI: https://doi.org/10.1515/ci-2017-0304

The Future of Chemical Information Is Now

by Antony J. Williams and Harry E. Pence

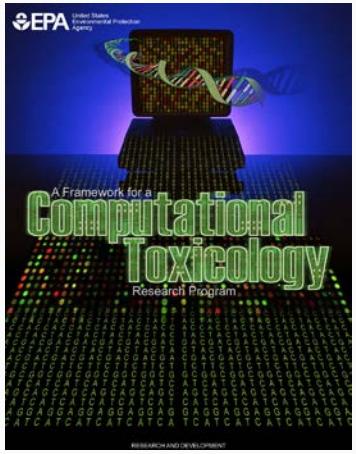
Search and retrieval of chemical information has been dramatically changed by the application of "Big Data" techniques. This development continues to be driven by the massive growth of chemical scientific literature and of online data and databases. Not only is there an expansion of the traditional avenues of publication, but many new contributing resources, such as open access journals, MOOCs (Massive Open Online Courses), Wikis, and blogs

What we need to progress

- Adequate information on potential bioactivity for as many chemicals as possible

- Read-across, QSAR and QSTR models from large high-quality curated datasets

National Center for Computational Toxicology



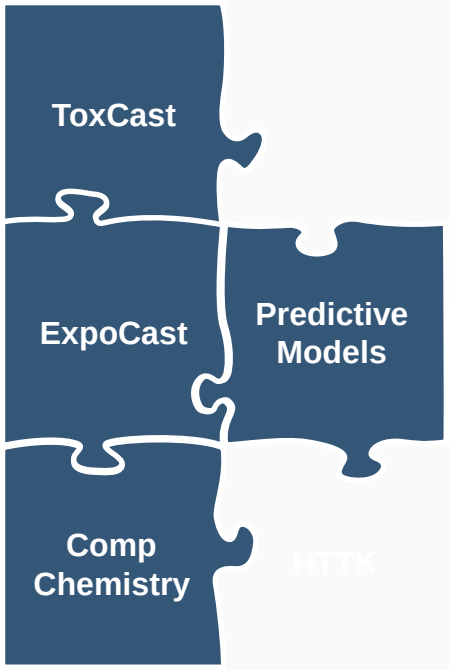
- National Center for Computational Toxicology established in 2005 to integrate:
  - High-throughput and high-content technologies
  - Modern molecular biology
  - Data mining and statistical modeling
  - Computational biology and chemistry
- Currently staffed by ~60 employees as part of EPA's Office of Research and Development
- Home of ToxCast & ExpoCast research efforts
- Key partner in U.S. Tox21 federal consortium
- Multiple cross-division collaborations

What we need to progress

- Adequate information on potential bioactivity for as many chemicals as possible
- **High Throughput Screening** ToxCast/Tox21 generating and making available bioactivity data
- GREAT progress to date – more to come!

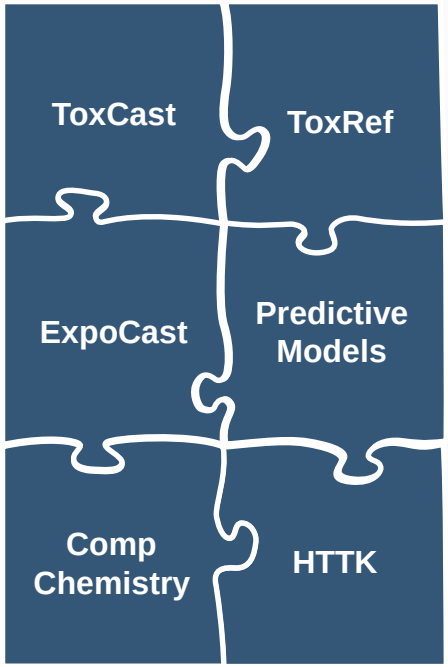
- Read-across, QSAR and QSTR models from large high-quality curated datasets
- **Chemistry data** gathering and high-throughput bioassays are providing the necessary data to generate these **models**

Data and Modeling Efforts to Date



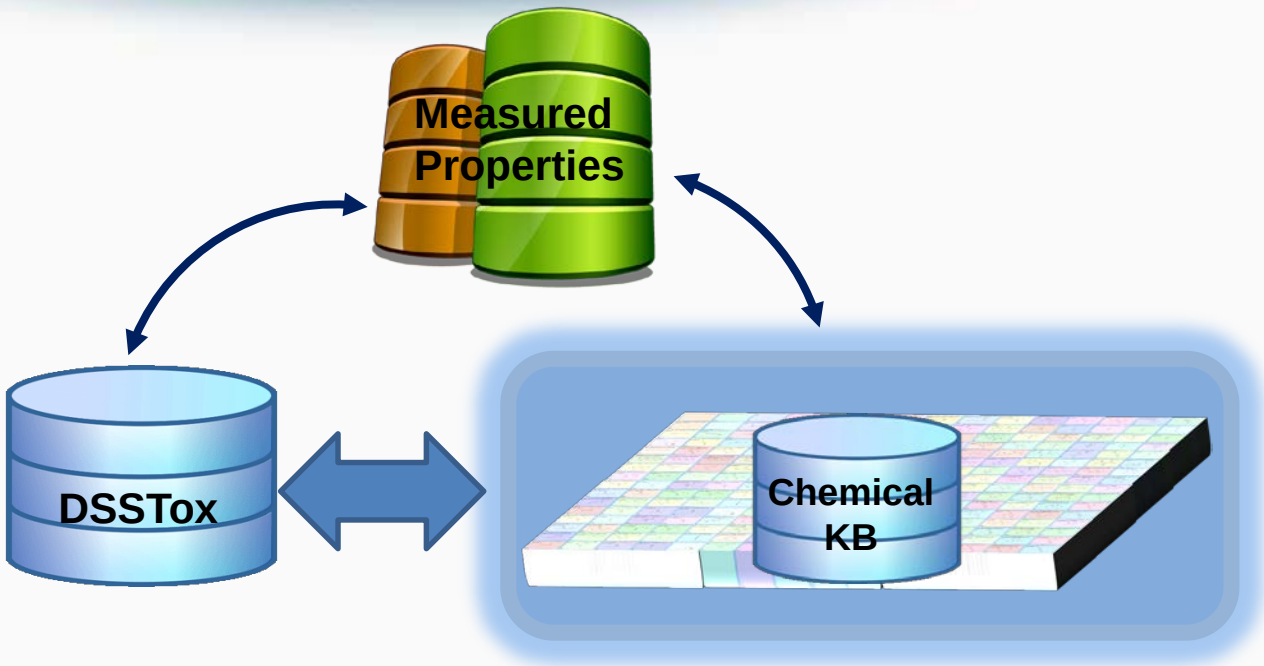
- High-throughput *in vitro* screening of ~4,000 chemicals across ~1200 assay endpoints (ToxCast) and ~8,000 chemicals in ~60 assay endpoints (Tox21)
- High quality, curated chemical structure and physical chemical properties database of ~750,000 chemicals
- High-throughput exposure estimates with uncertainty for ~ 7,000 chemicals based on production volume and chemical use

Data and Modeling Efforts to Date

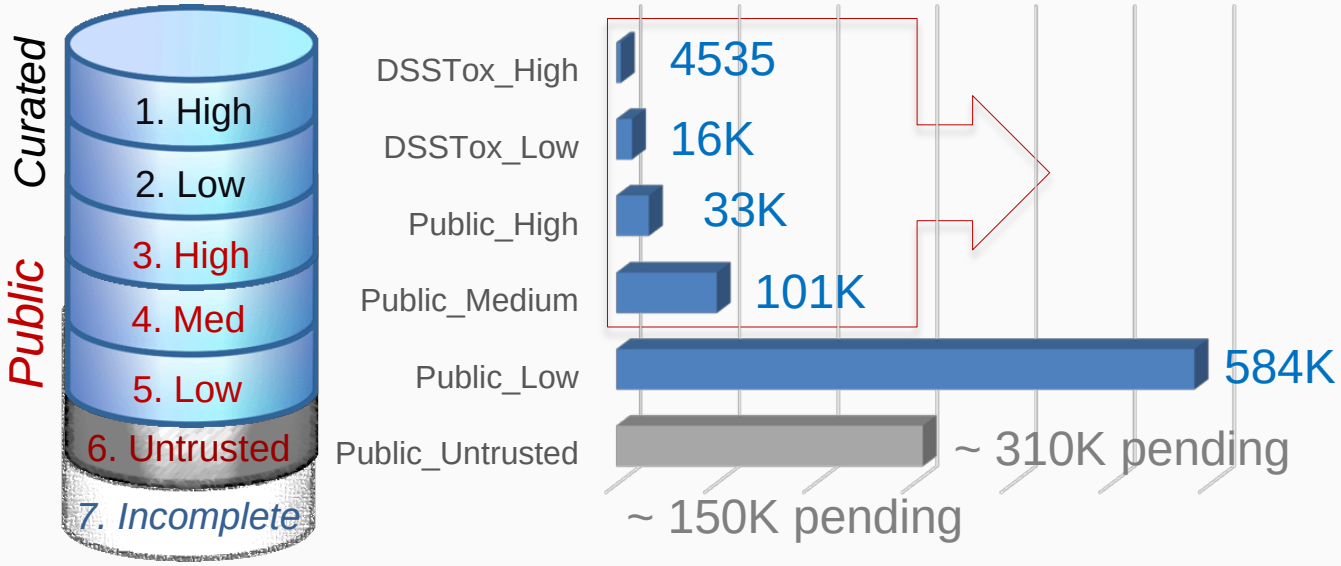


- High-throughput *in vitro* screening of ~4,000 chemicals across ~1200 assay endpoints (ToxCast) and ~8,000 chemicals in ~60 assay endpoints (Tox21)
- High quality, curated chemical structure and physical chemical properties database of ~750,000 chemicals
- High-throughput exposure estimates with uncertainty for ~ 7,000 chemicals based on production volume and chemical use
- Legacy *in vivo* data from ~6,000 animal toxicology studies on ~1,110 unique chemicals
- High-throughput toxicokinetic (HTTK) models for ~700 chemicals based on *in vitro* measurements
- AOPs and virtual tissue models for broad range of developmental toxicities

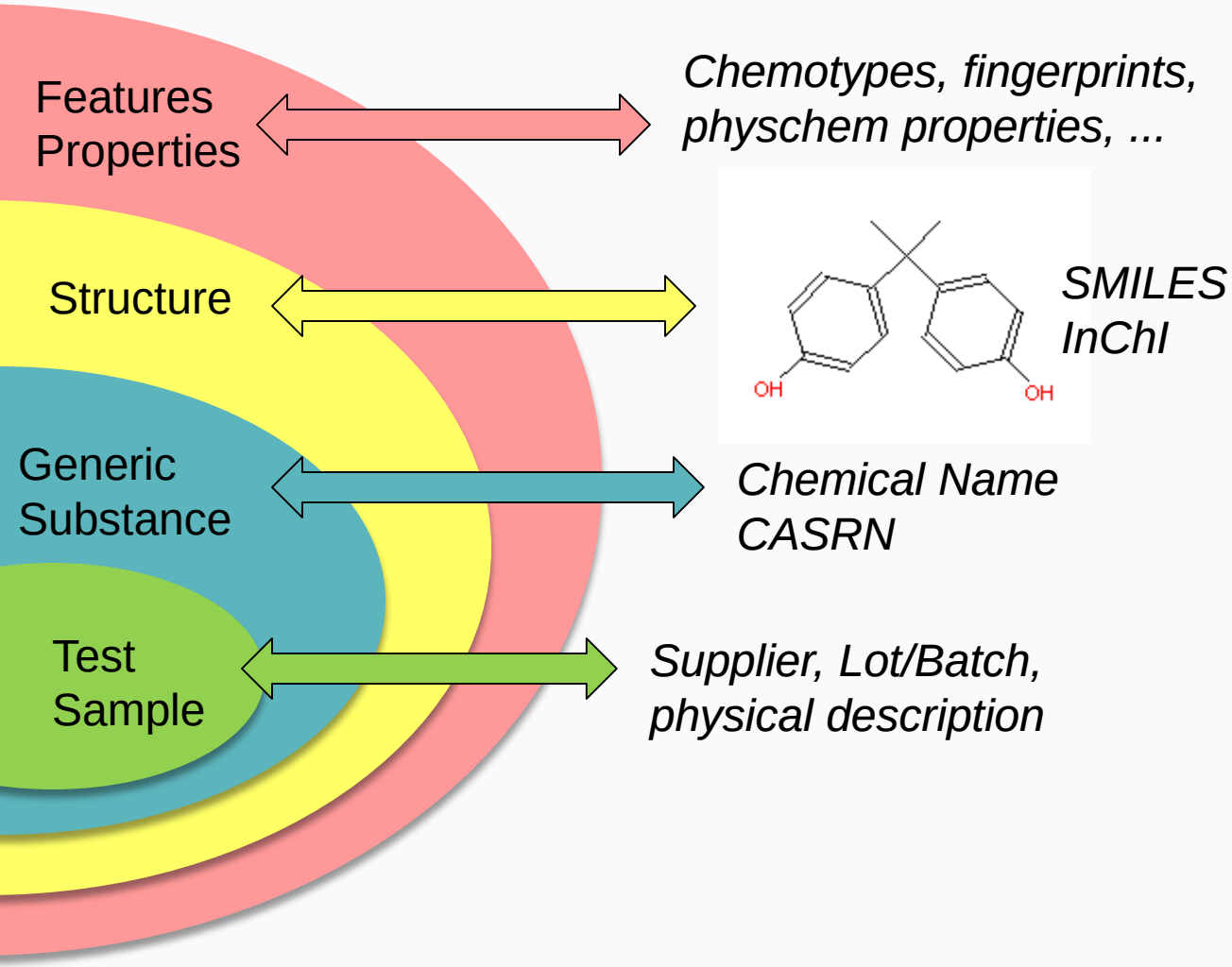
Underpinning with chemicals



Approximately 15 Years of Data… Growing with daily curation



Chemical representation levels supporting data integration



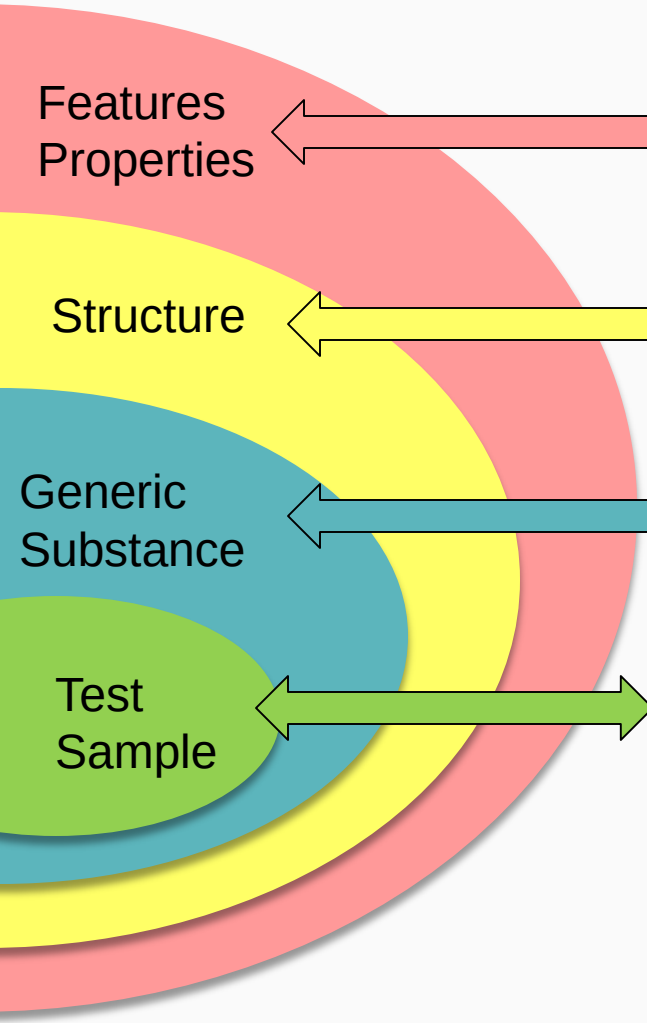
Chemotypes, fingerprints, physchem properties, ...

SMILES
InChI

Chemical Name
CASRN

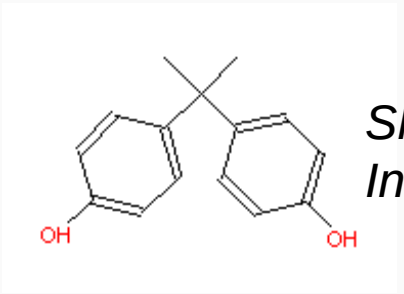
Supplier, Lot/Batch, physical description

Chemical representation levels supporting data integration



Chemotypes, fingerprints, physchem properties, ...

Chemical analogs, Read-across, SAR modeling



SMILES
InChI

Structure searching & modeling

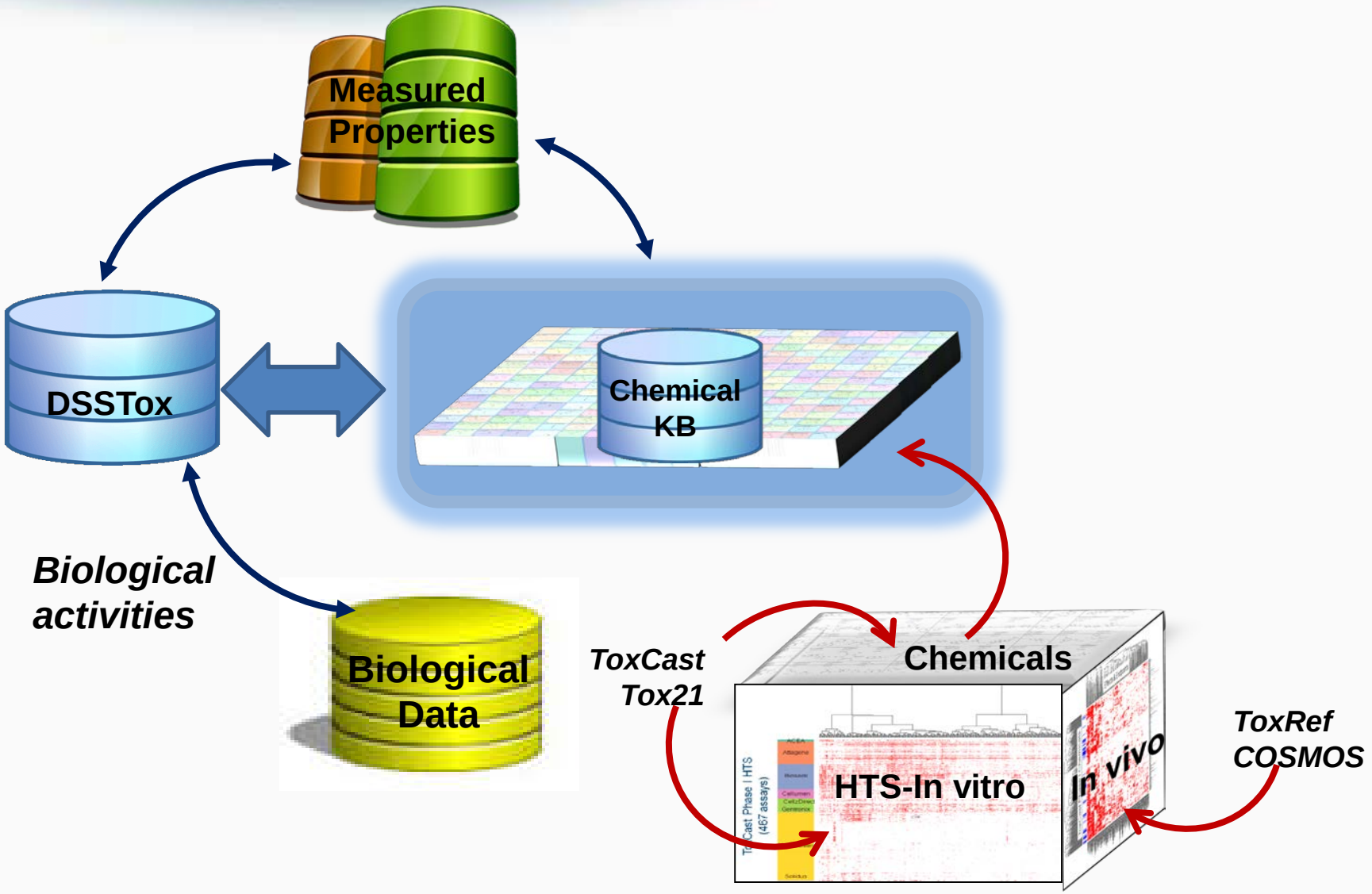
Chemical Name
CASRN

Public toxicity datasets

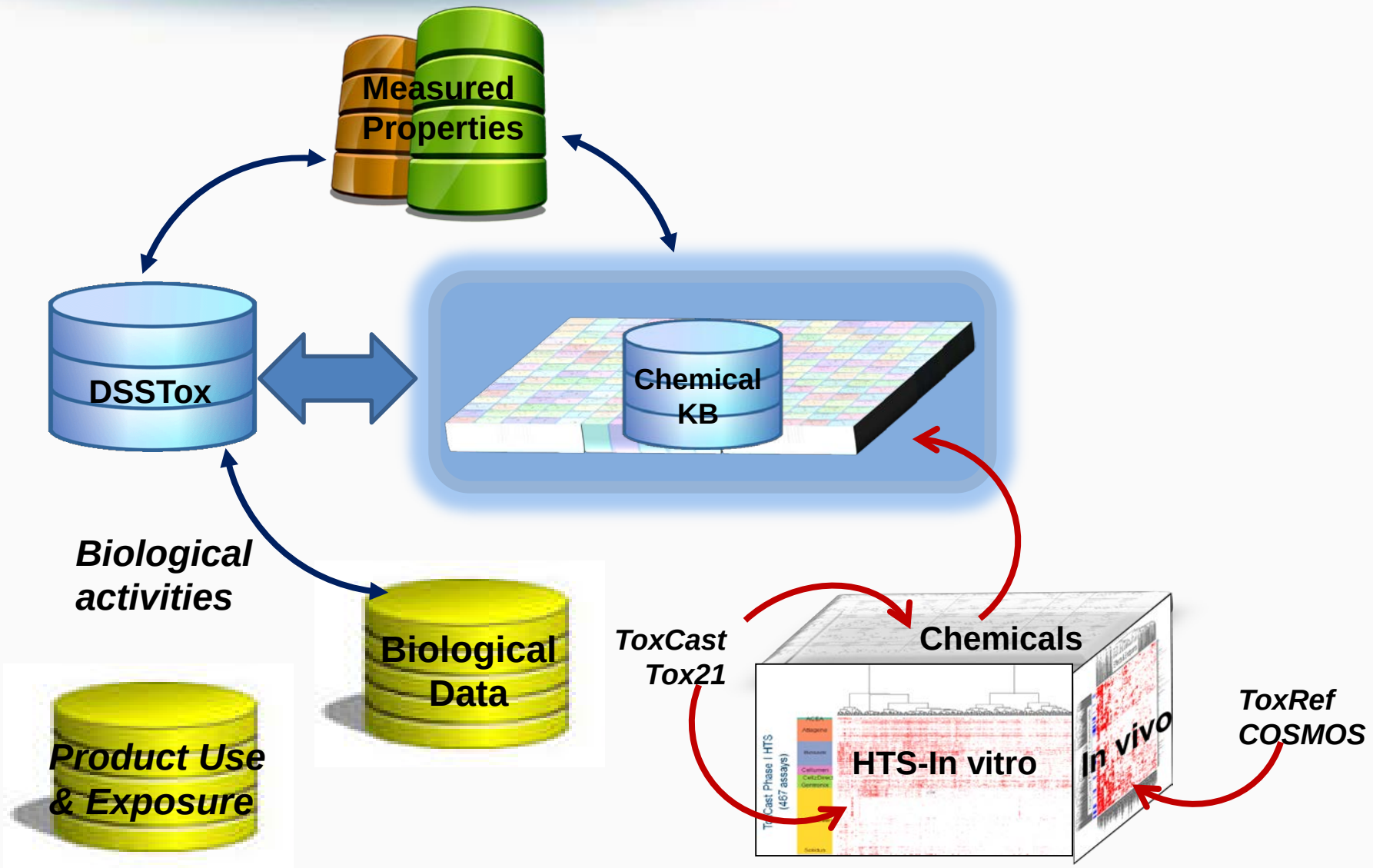
Supplier, Lot/Batch, physical description

Experimental Endpoint Data

Integrating *in vitro* and *in vivo* data



Adding Product Use and Exposure



High Throughput Measurement to Characterize **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

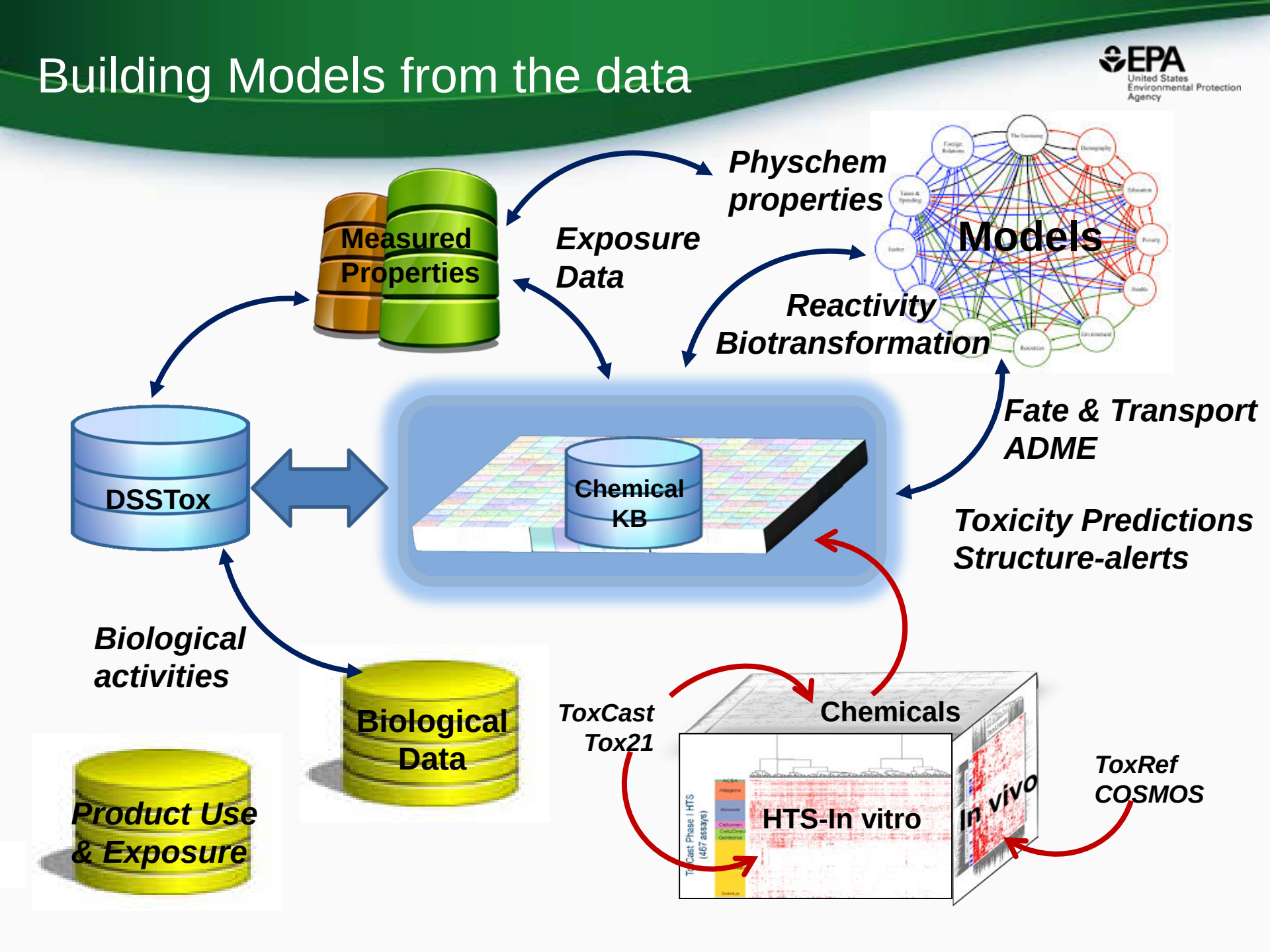
› Farfield Exposure Models

› Nearfield Exposure Models

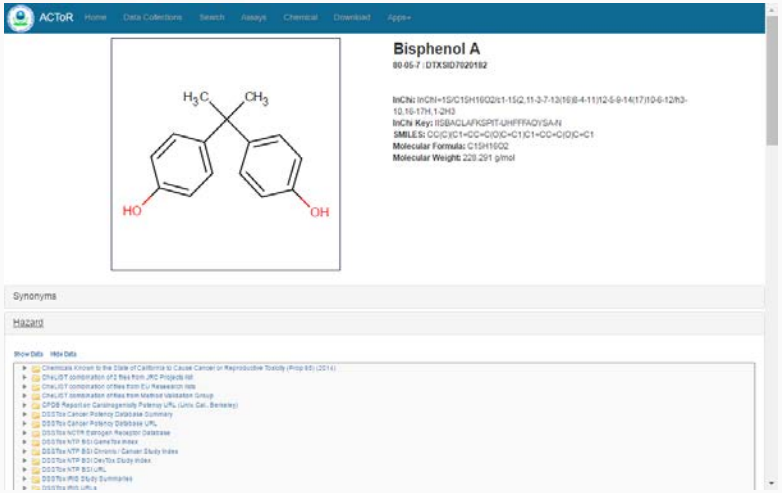
Evaluating High-throughput Exposure Predictions

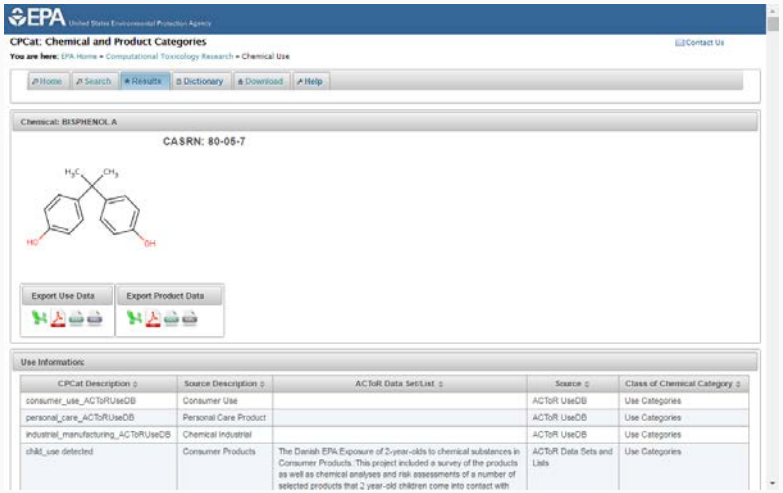
EPA is currently evaluating the effectiveness of high-throughput exposure models

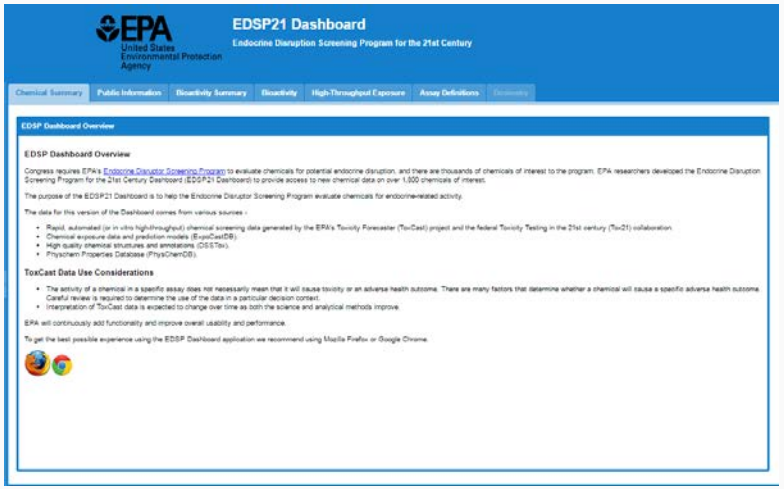
Building Models from the data



Access to Our Data and Models



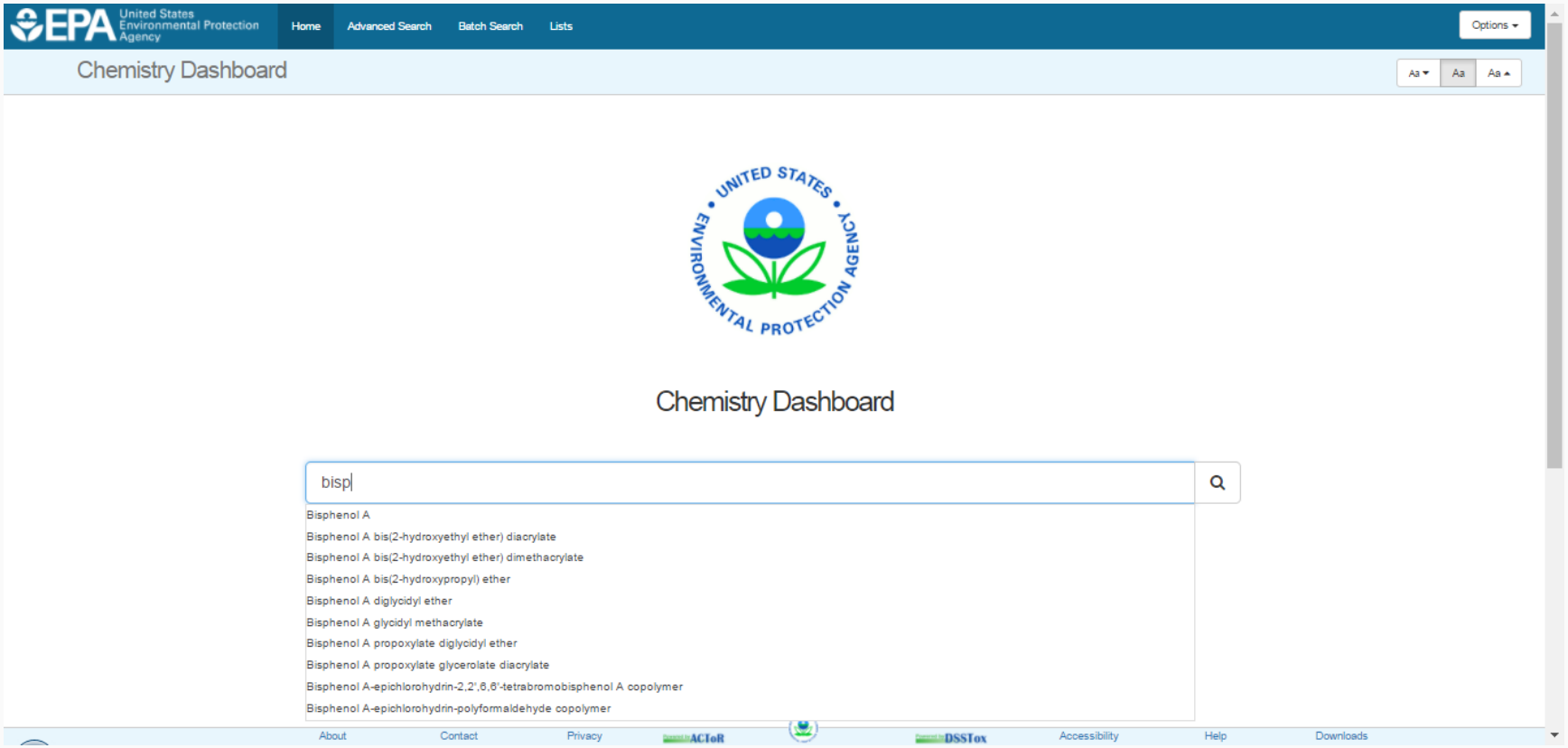




What we have learned…

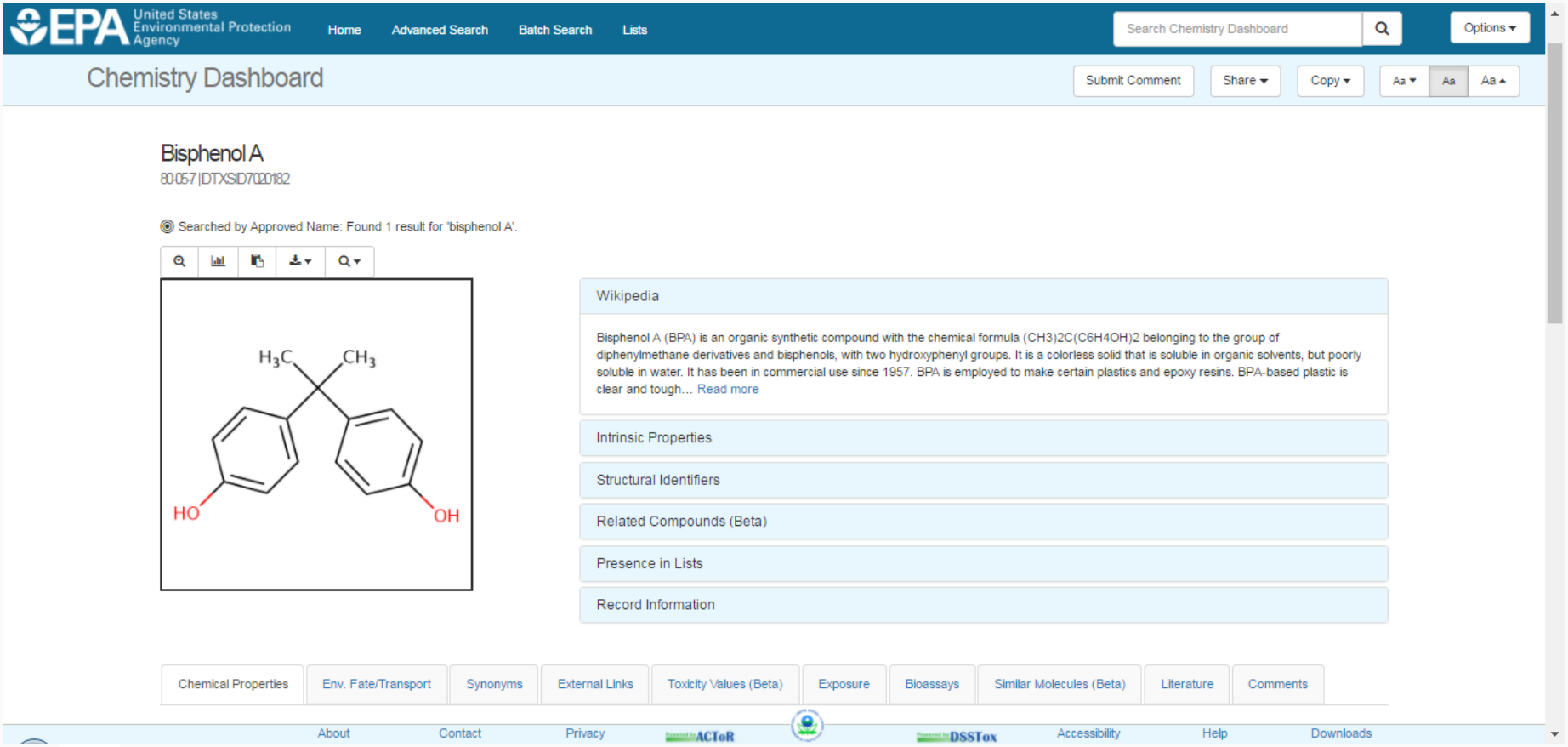
- Data **curation**, standardization and versioning is essential
- Prototype application development suffices for research projects but **production development requires managed processes**
- **ODOSOS** (Open Data, Open Source and Open Standards) endows many benefits
- With this in mind…

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



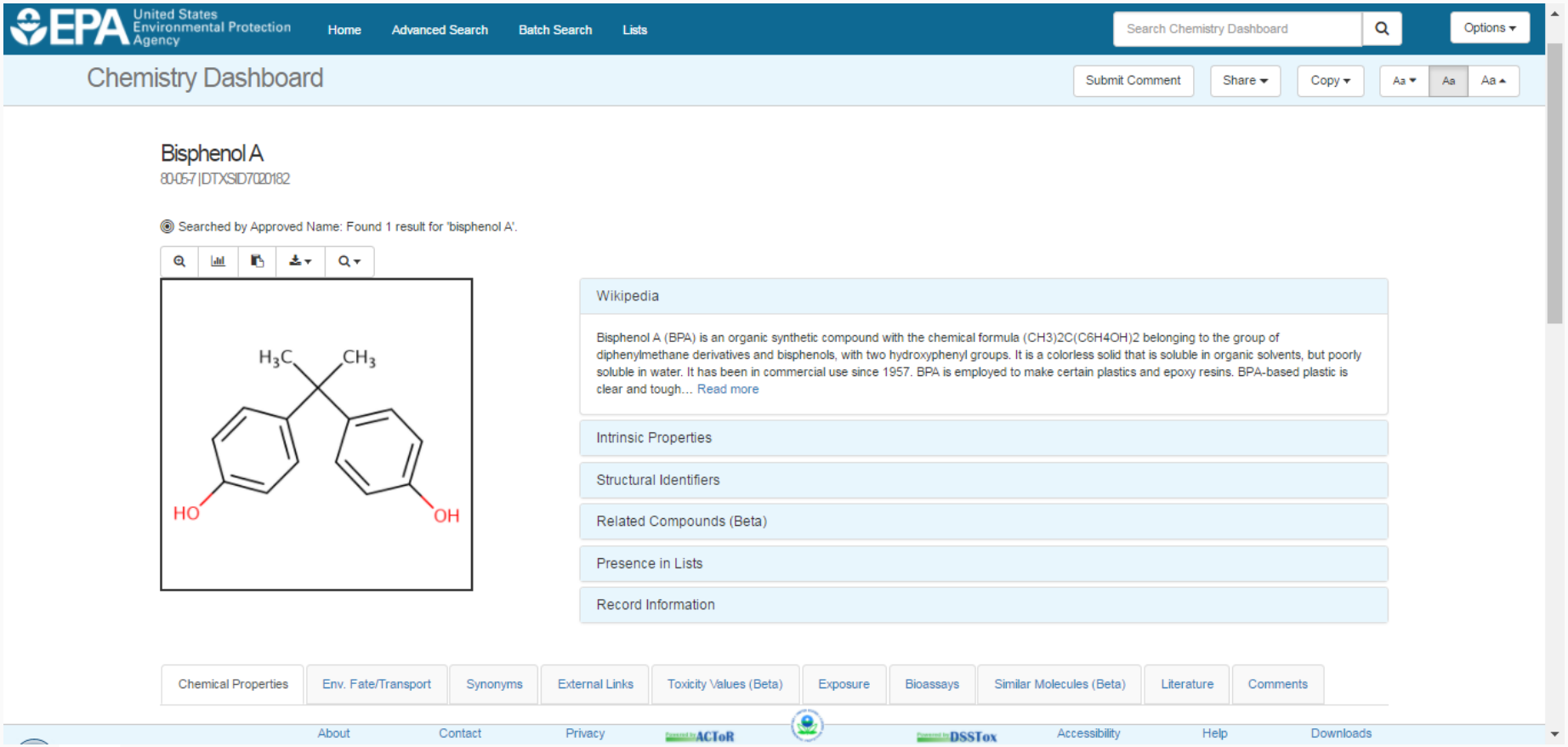
Bisphenol A

EPA United States Environmental Protection Agency



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

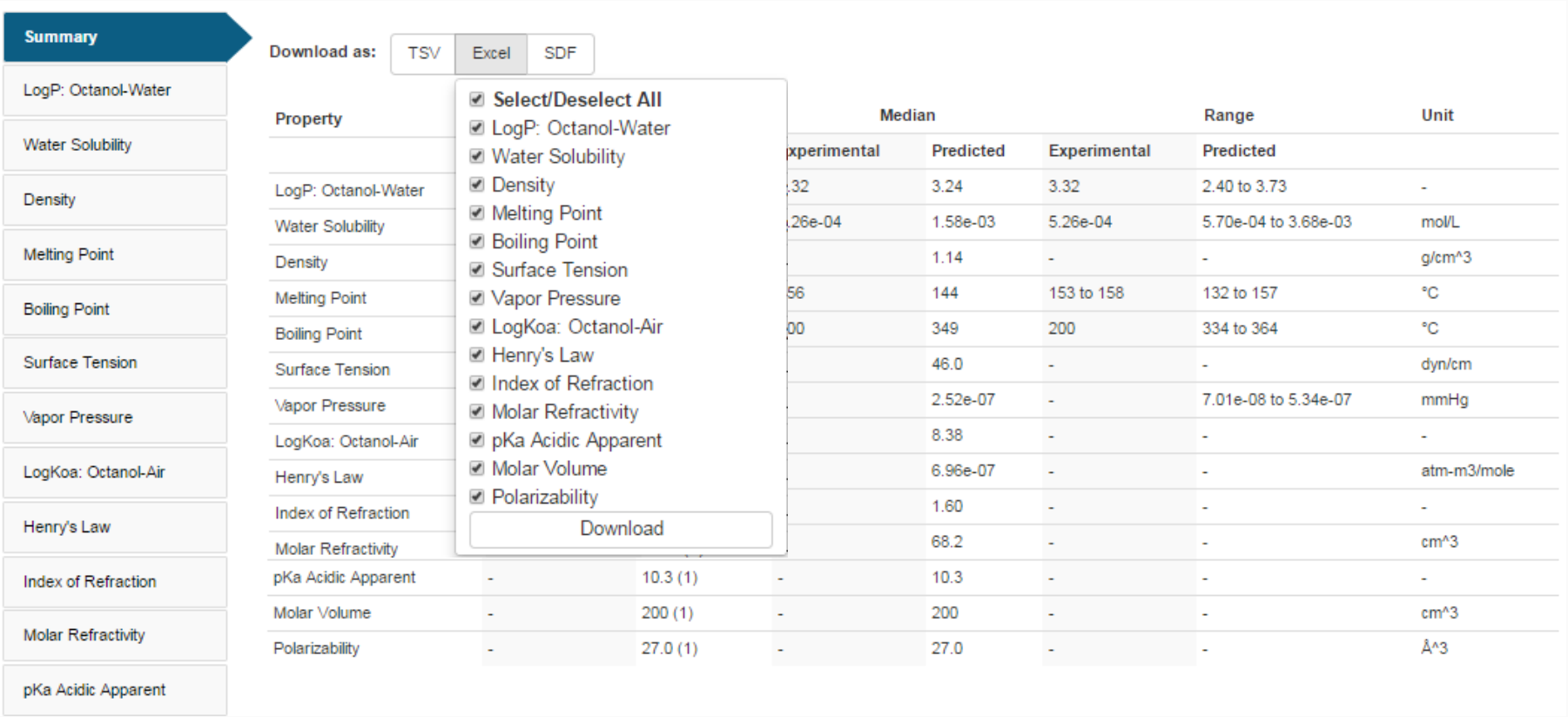
Download as: TSV | Excel | SDF

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



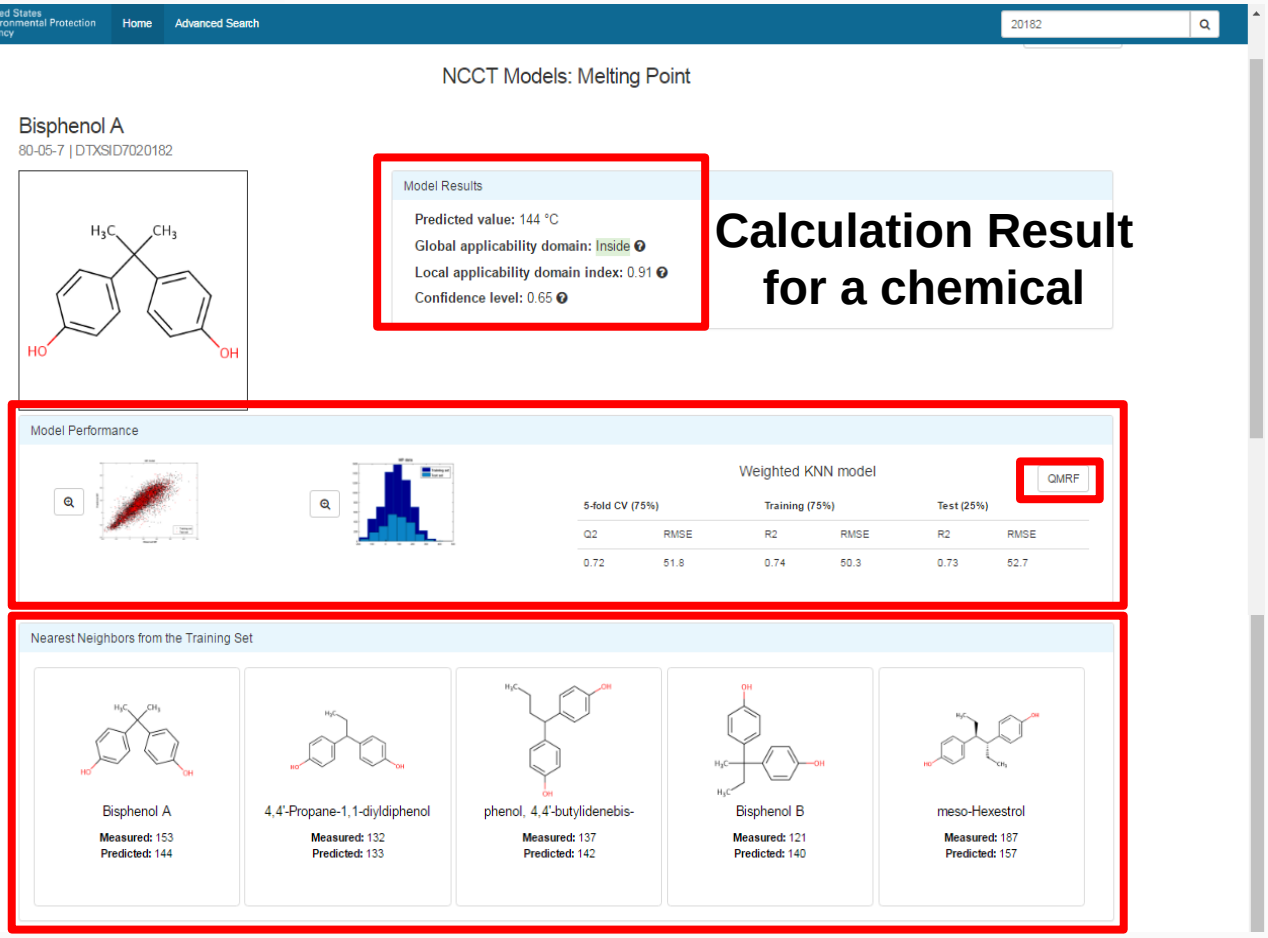
Data Distribution

	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
17								

Modeling Details

Model Performance with full QMRF

Nearest Neighbors from Training Set



Prediction Details and QMRF Report

EPA United States Environmental Protection Agency

Model Results

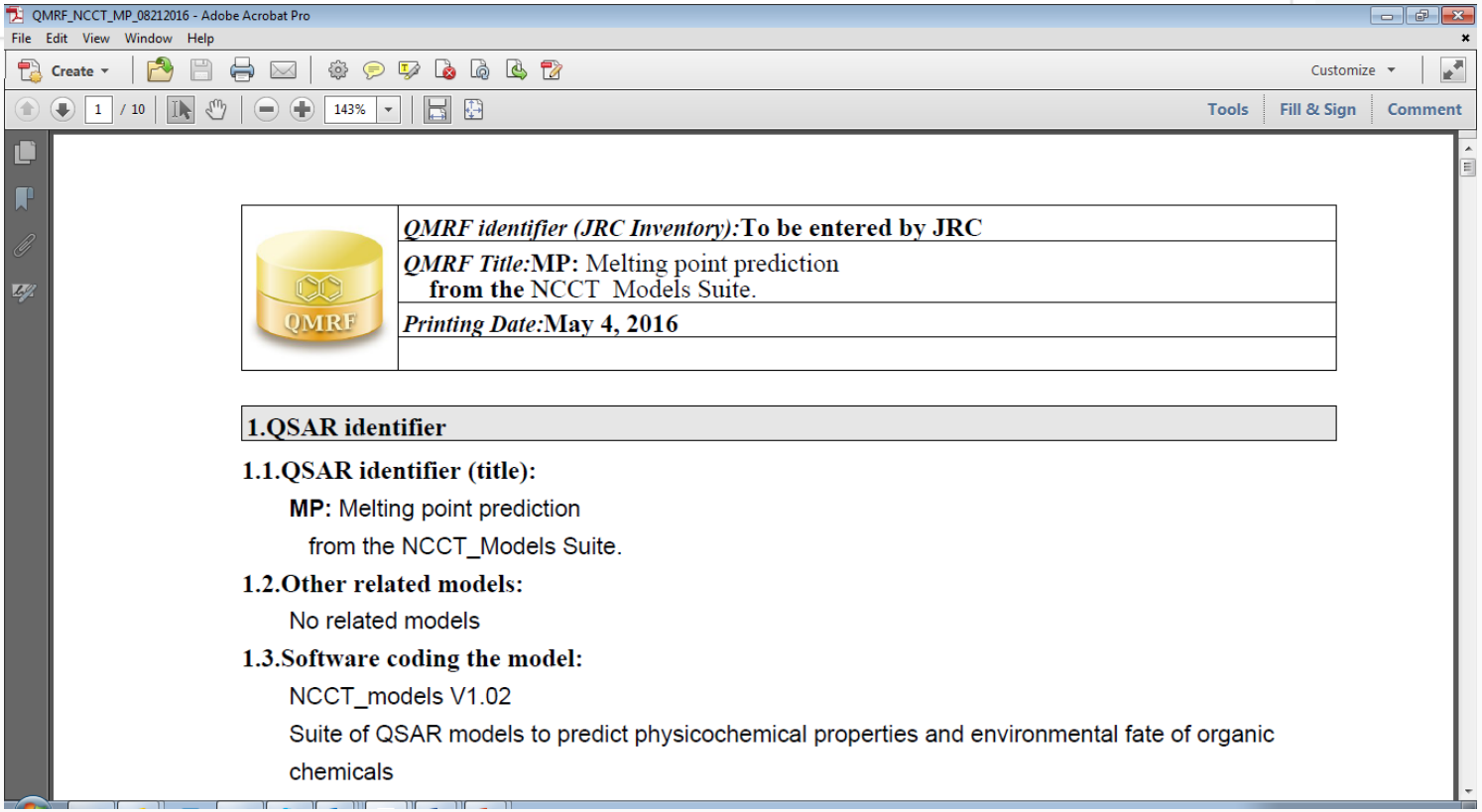
Predicted value: 144 °C

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



QMRF_NCCT_MP_08212016 - Adobe Acrobat Pro

File Edit View Window Help

Create Customize

1 / 10 143% Tools Fill & Sign Comment

QMRF	*QMRF identifier (JRC Inventory):* **To be entered by JRC**
	QMRF Title: **MP:** Melting point prediction **from the** NCCT Models Suite.
	Printing Date: **May 4, 2016**

1.QSAR identifier

1.1.QSAR identifier (title):

MP: Melting point prediction from the NCCT_Models Suite.

1.2.Other related models:

No related models

1.3.Software coding the model:

NCCT_models V1.02

Suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals

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**

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 $

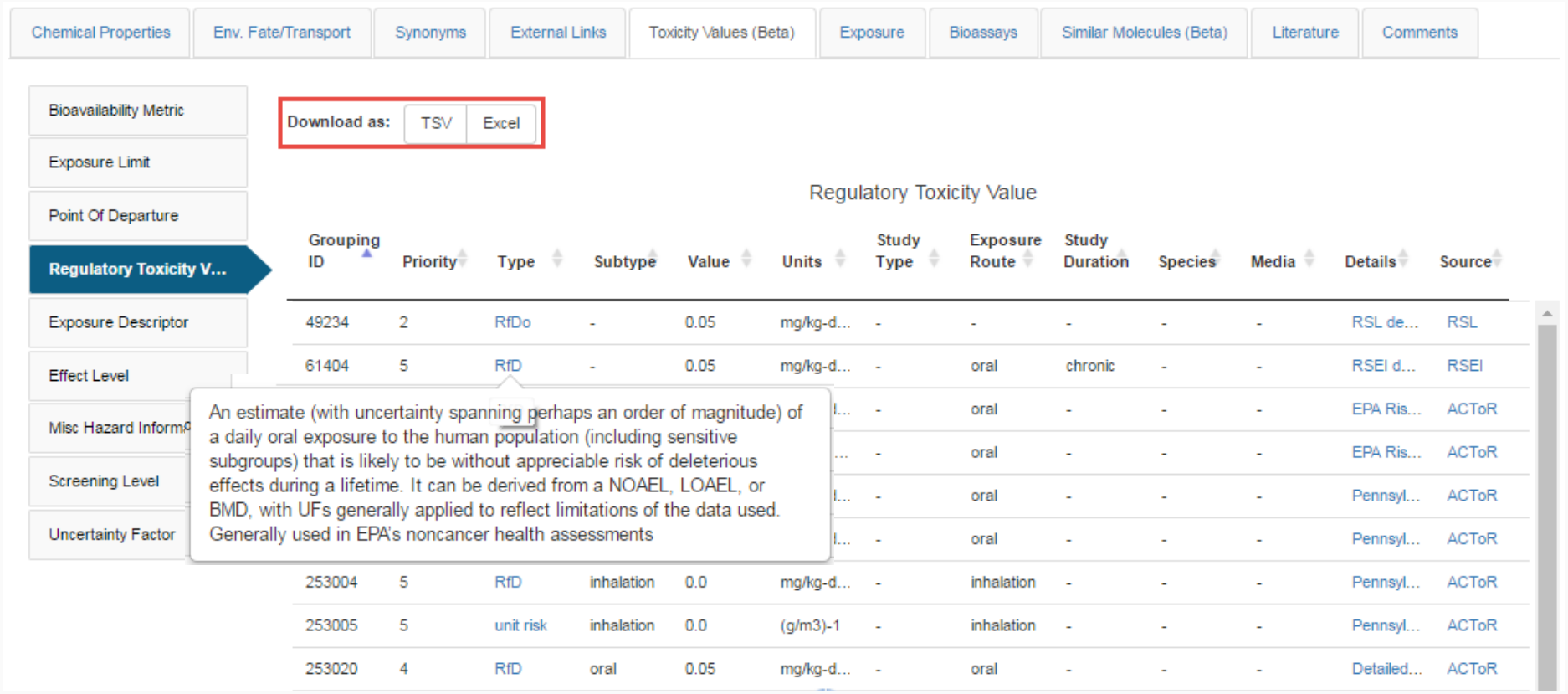
K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams

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

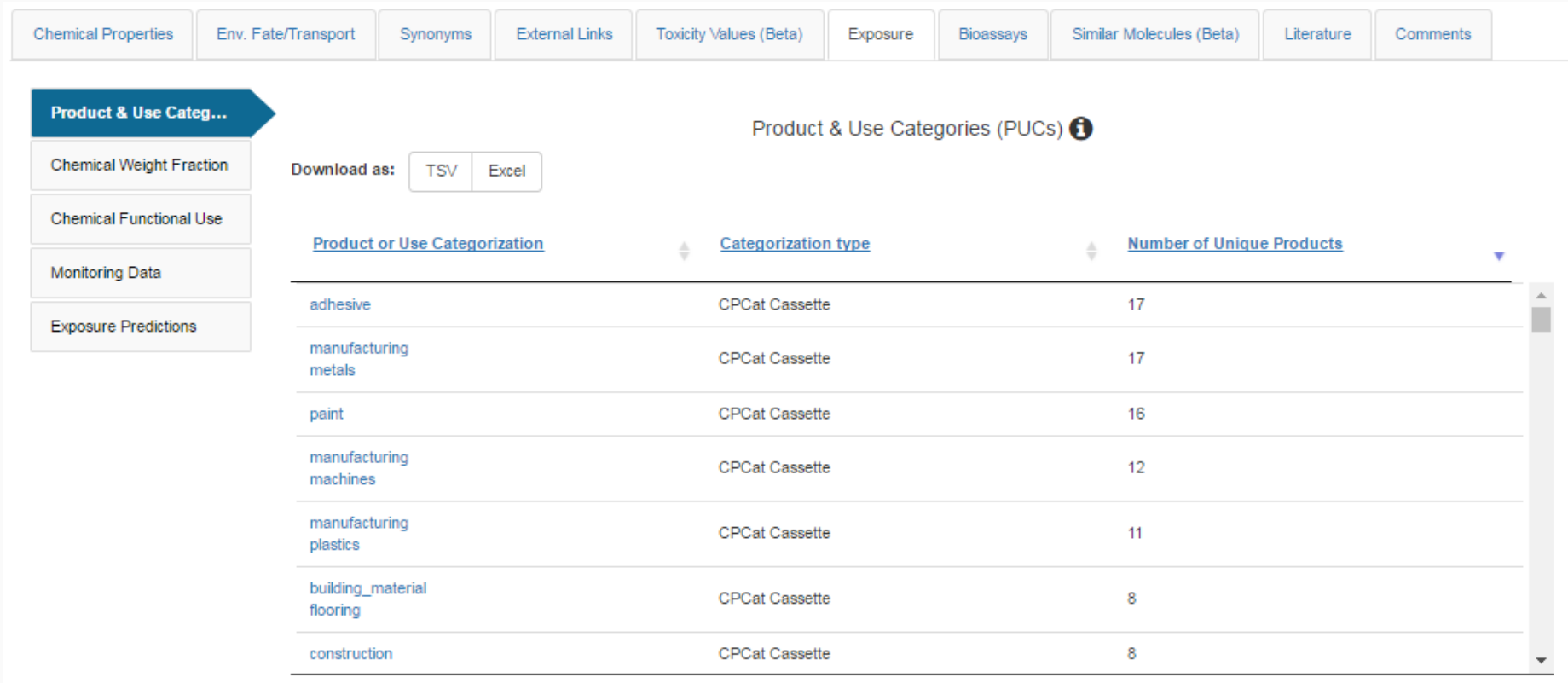
Download citation http://dx.doi.org/10.1080/1062936X.2016.1253611 Check for updates

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

Making Toxicity Value Data available



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

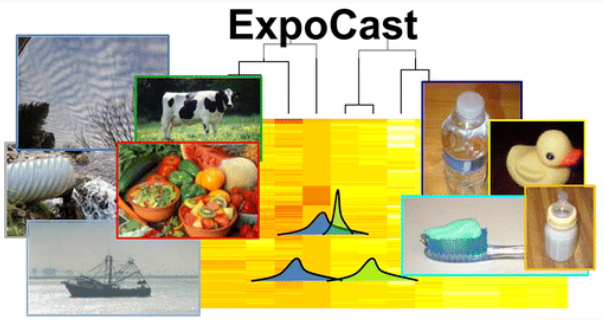
NHANES Exposure data

EPA United States Environmental Protection Agency

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

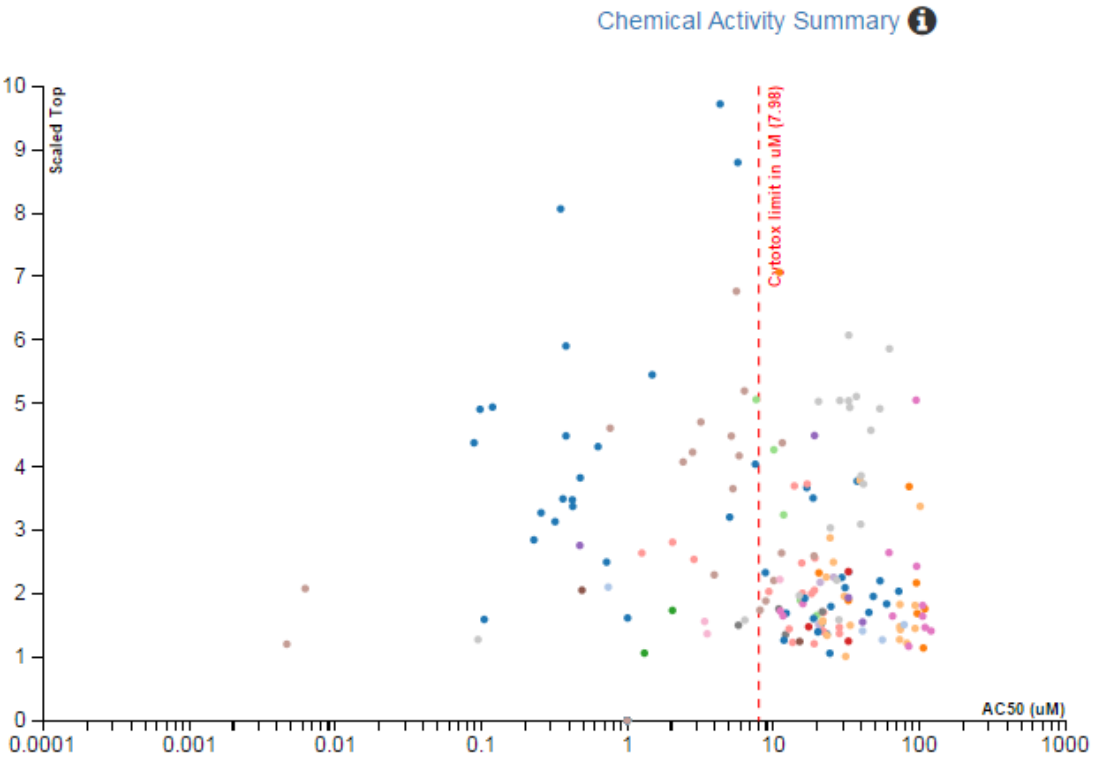
ToxCast and Tox21 Bioassays

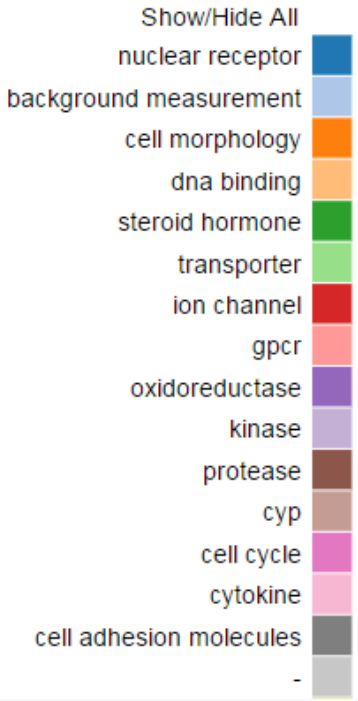
EPA United States Environmental Protection Agency

ToxCast

PubChem

ScrubChem (Beta)





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

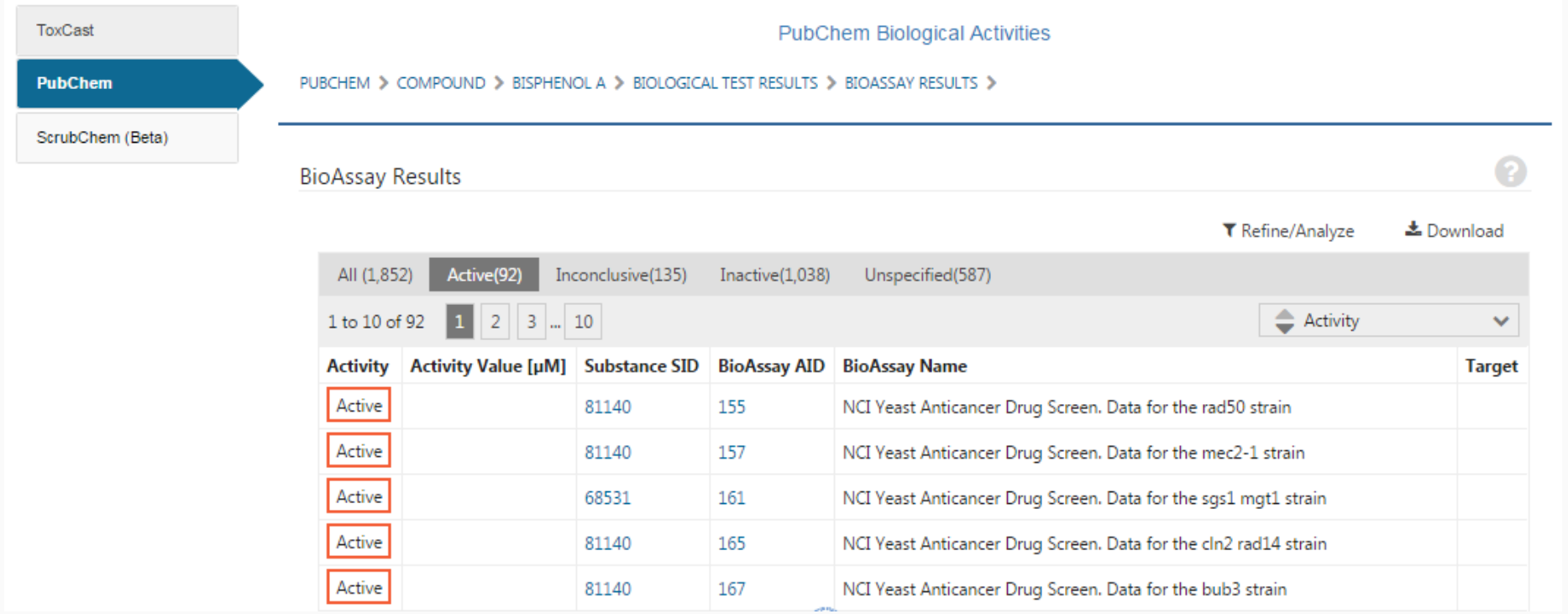
ToxCast and Tox21 Bioassays

EPA United States Environmental Protection Agency

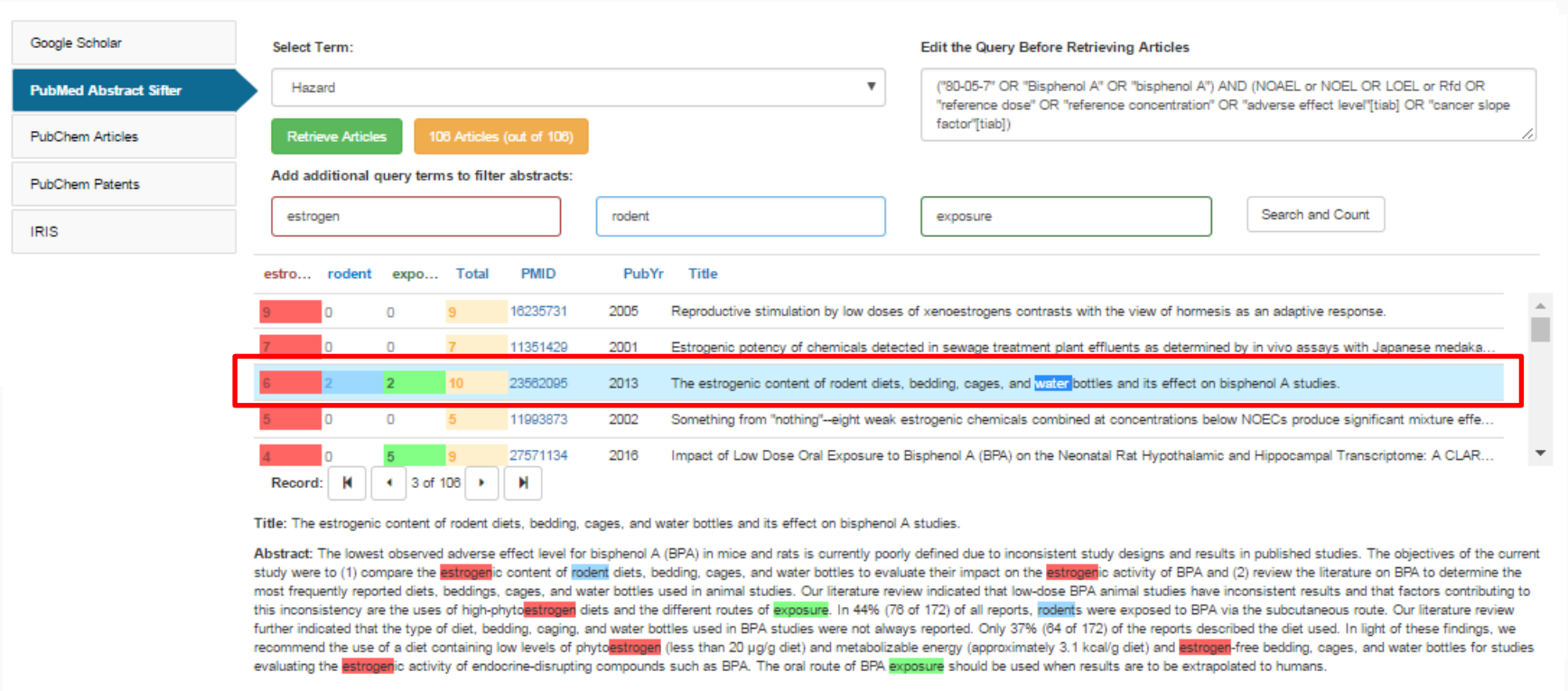
Download as: TSV | Excel **Show:** Inactive | Background

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



PubMed: BIG DATA Literature Search Integration

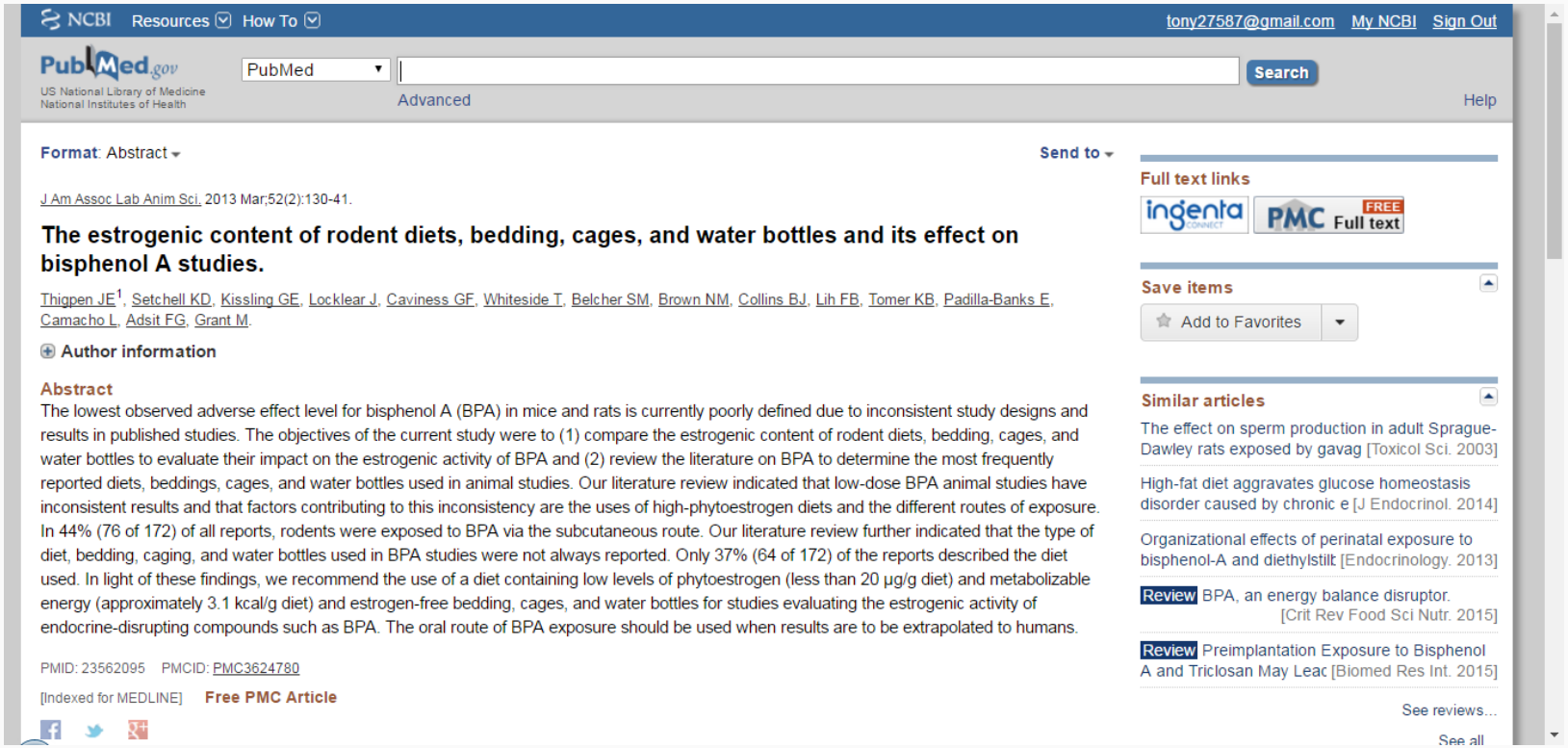


Linked Directly to PubMed

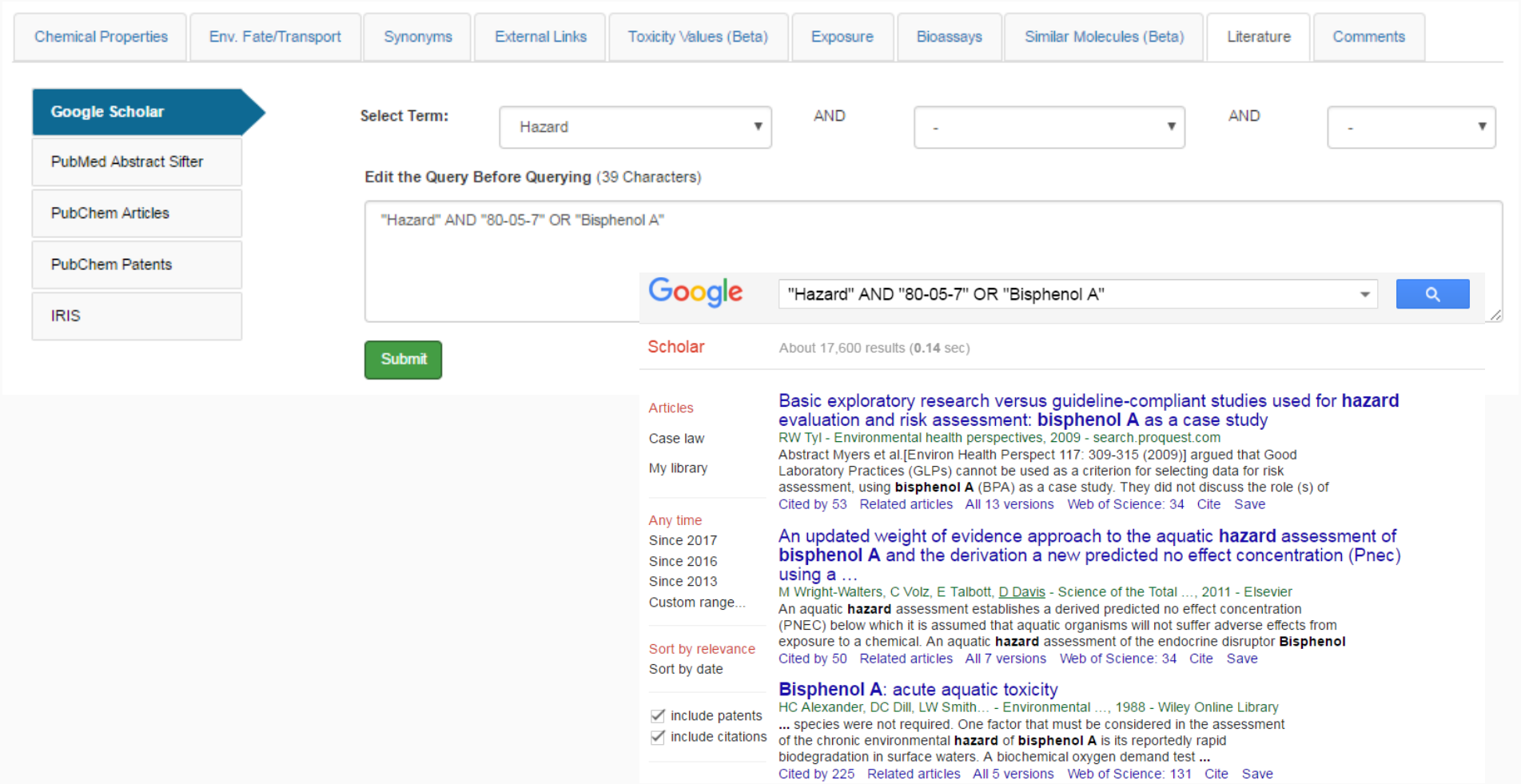
EPA United States Environmental Protection Agency

PubMed

PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

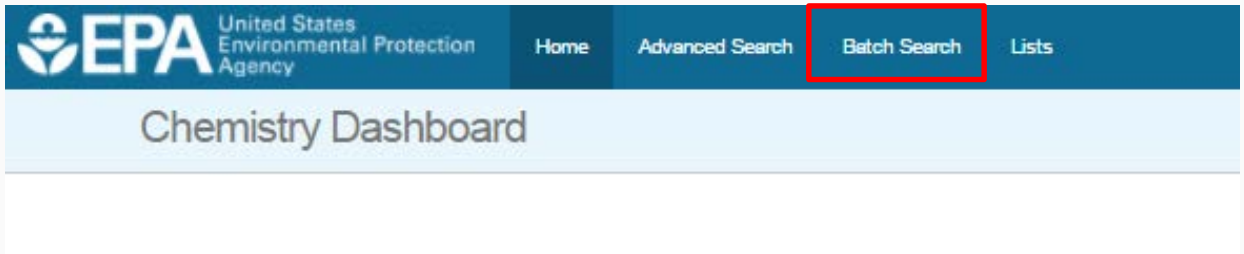


Literature Search Integration
Google Scholar



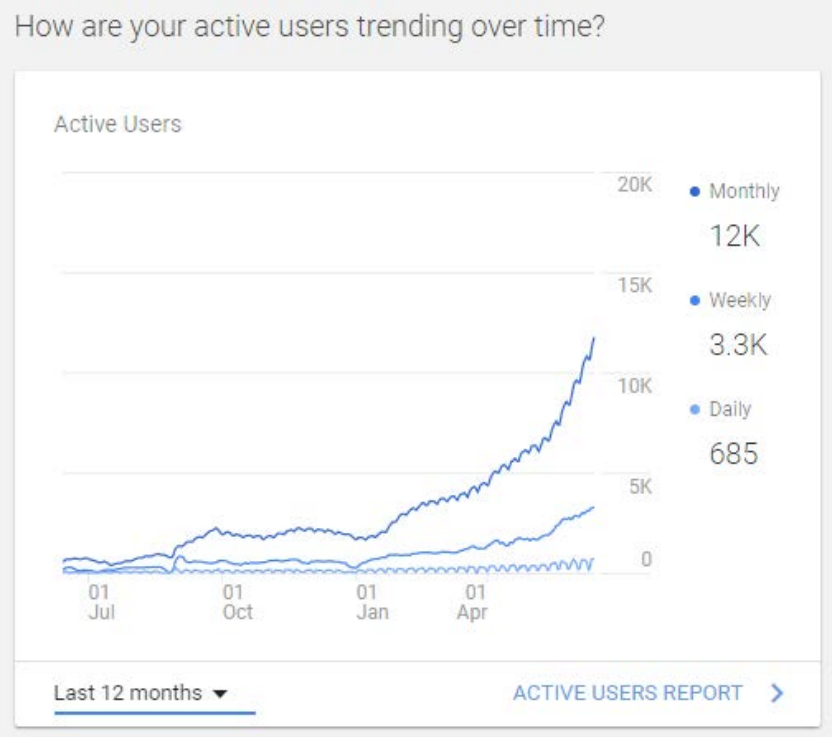
Batch Data Searching

- I have a 5000 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?
  - Can I get predicted physchem data for my model?



We get lots of user feedback..

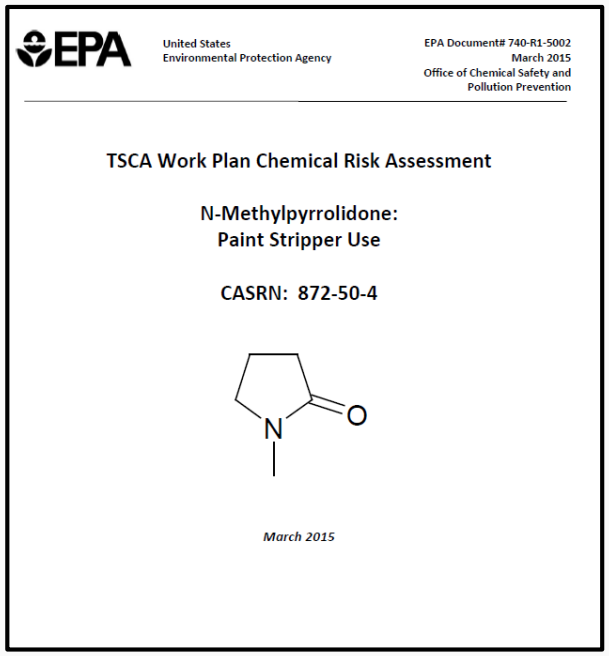
- Online since April 2016
- Feedback welcomed!



“RapidTox” in Development

- Semi-automated **decision support tool** for high-throughput risk assessments
- Use Dashboard “architecture”, existing data streams and add new data – e.g. Global Hazard Summary and ECHA data
- Combine data streams into **quantitative toxicity values** with uncertainty estimates

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

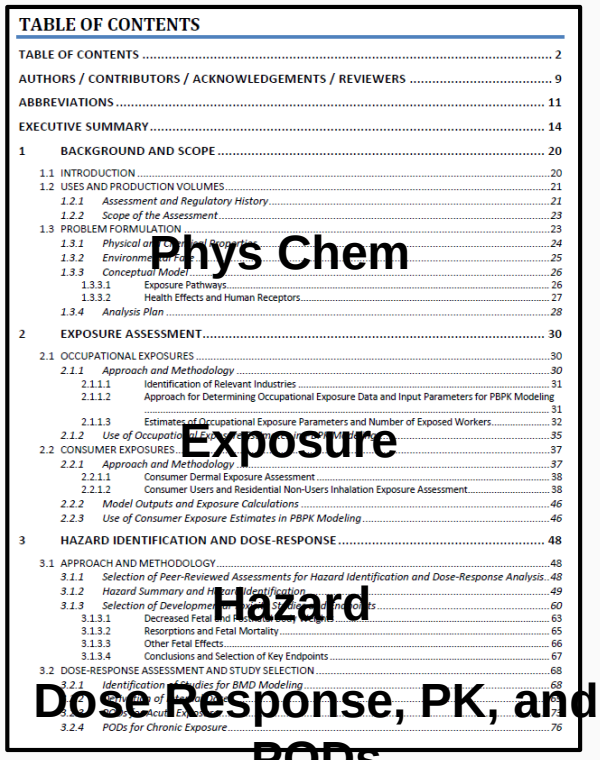
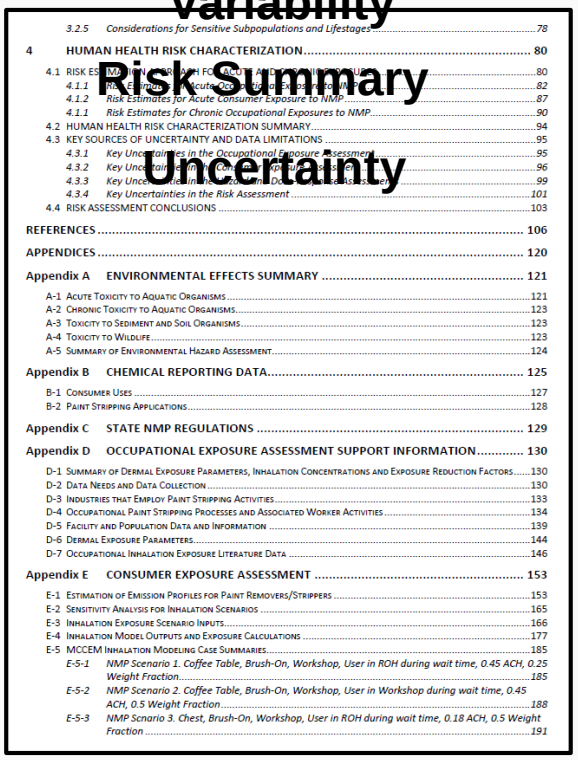


TABLE OF CONTENTS



Phys Chem

Exposure

Hazard

Dose Response, PK, and PODs

Variability

Risk Summary

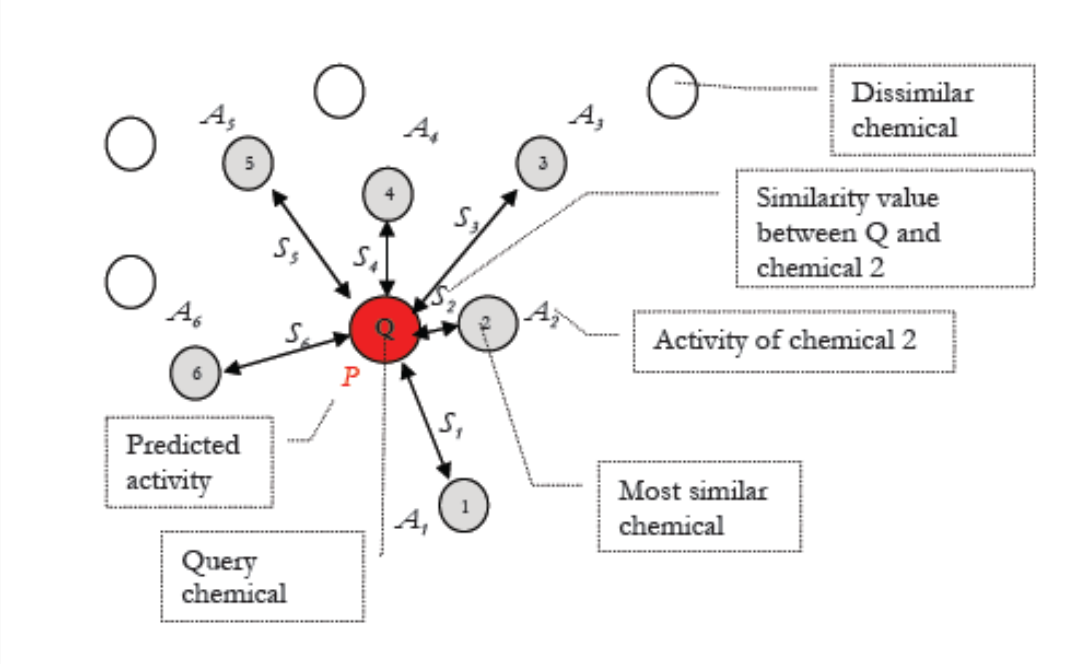
Uncertainty

We are assembling these components to deliver RapidTox

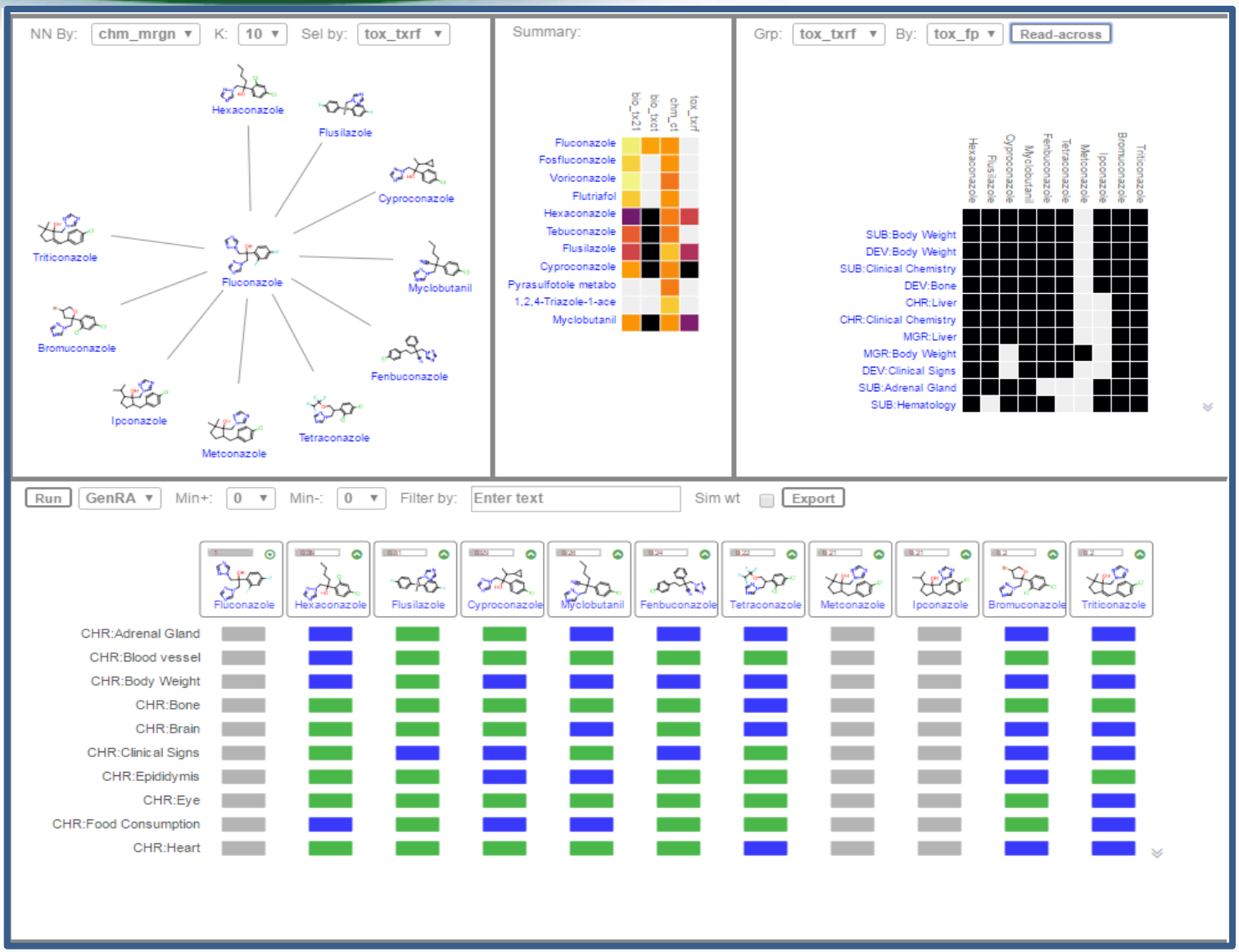
Quantifying Uncertainty & Assessing Performance of Read-Across - GenRA

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

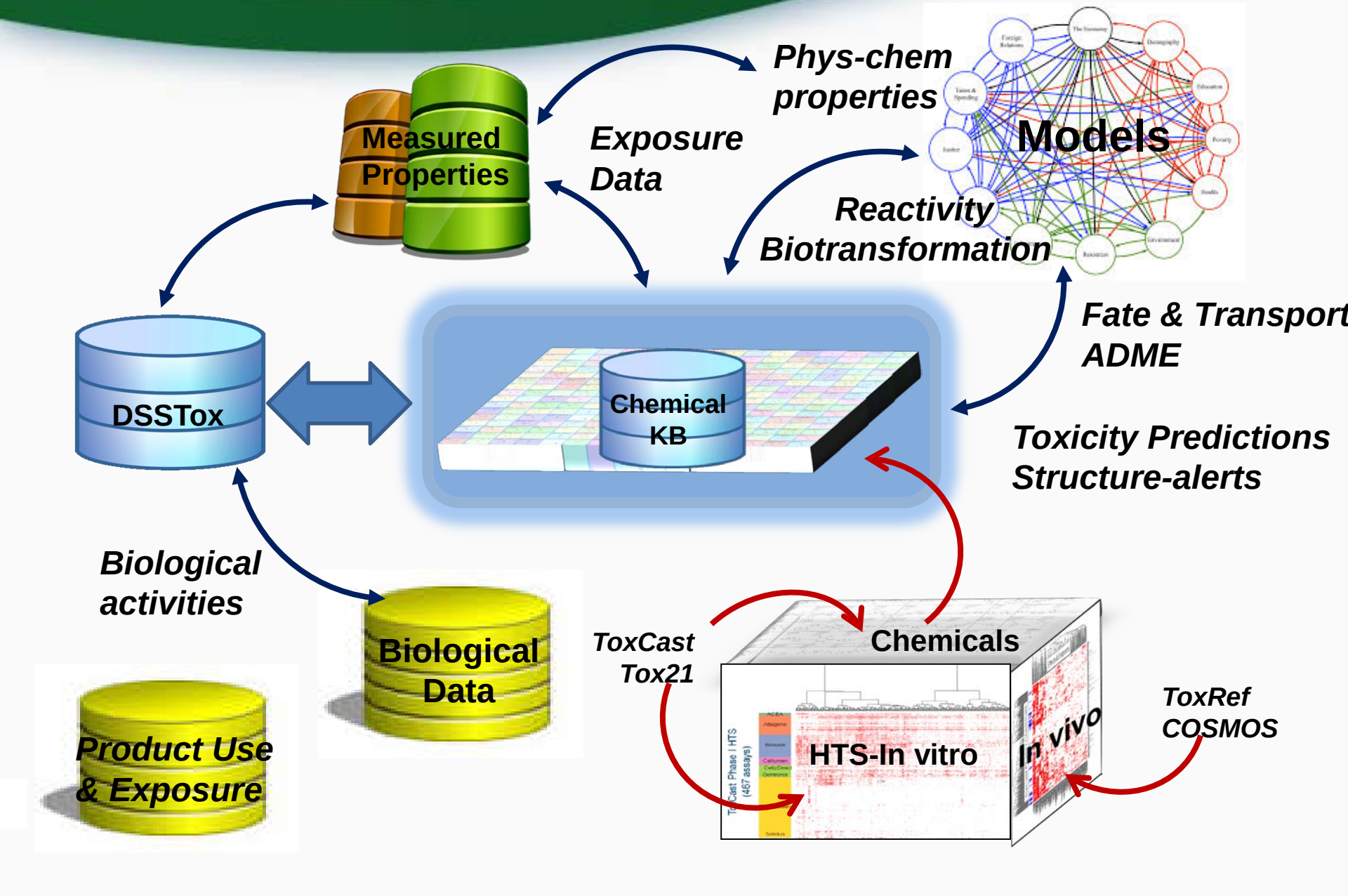
- Evaluates read-across performance and uncertainty using available data



Work-in-progress



We're not done yet…



Acknowledgements

EPA-RTP
Russell Thomas
Kevin Crofton
Chris Grulke
Ann Richard
Richard Judson
John Cowden
John Wambaugh
Imran Shah
Grace Patlewicz
and…

Many other contributors from across EPA-ORD

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