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

EPA United States Environmental Protection Agency

High-throughput experimental and computational technologies at the National Center for Computational Toxicology

Chris Grulke, *Antony Williams, John Wambaugh, Richard Judson, Keith Houck and Katie Paul-Friedman*

National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC

August 2019

ACS Fall Meeting, San Diego

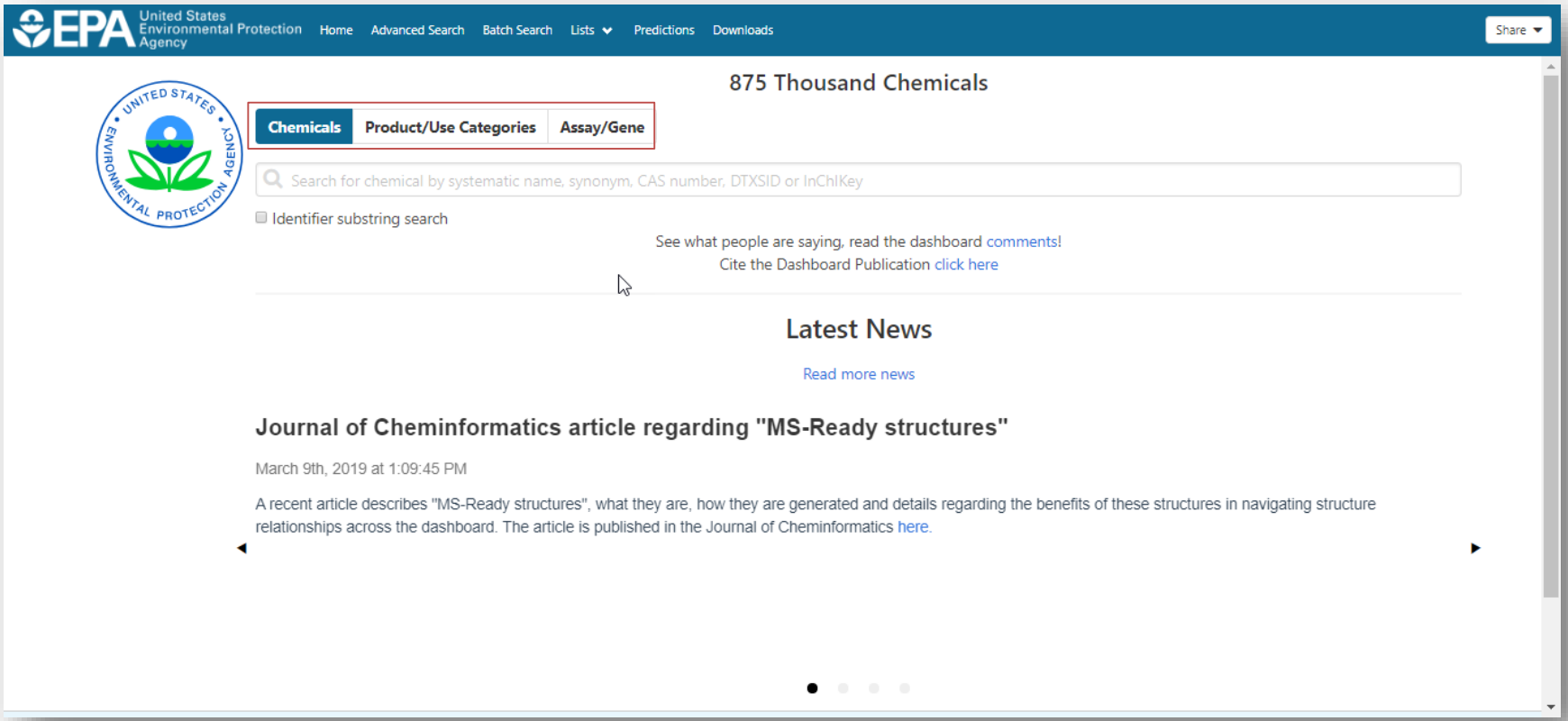
Overview

- The CompTox Chemicals Dashboard - web-based database of 875k substances
- Provides access to “high-throughput data”
  - High-throughput bioactivity “ToxCast” data
  - High-throughput prediction “physchem, fate & transport”
  - High-throughput “exposure modeling”
  - High-throughput retrieval of data – “batch searching”
- Work in progress
  - High-throughput standardization
  - High-throughput toxicokinetics
  - High-throughput transcriptomics

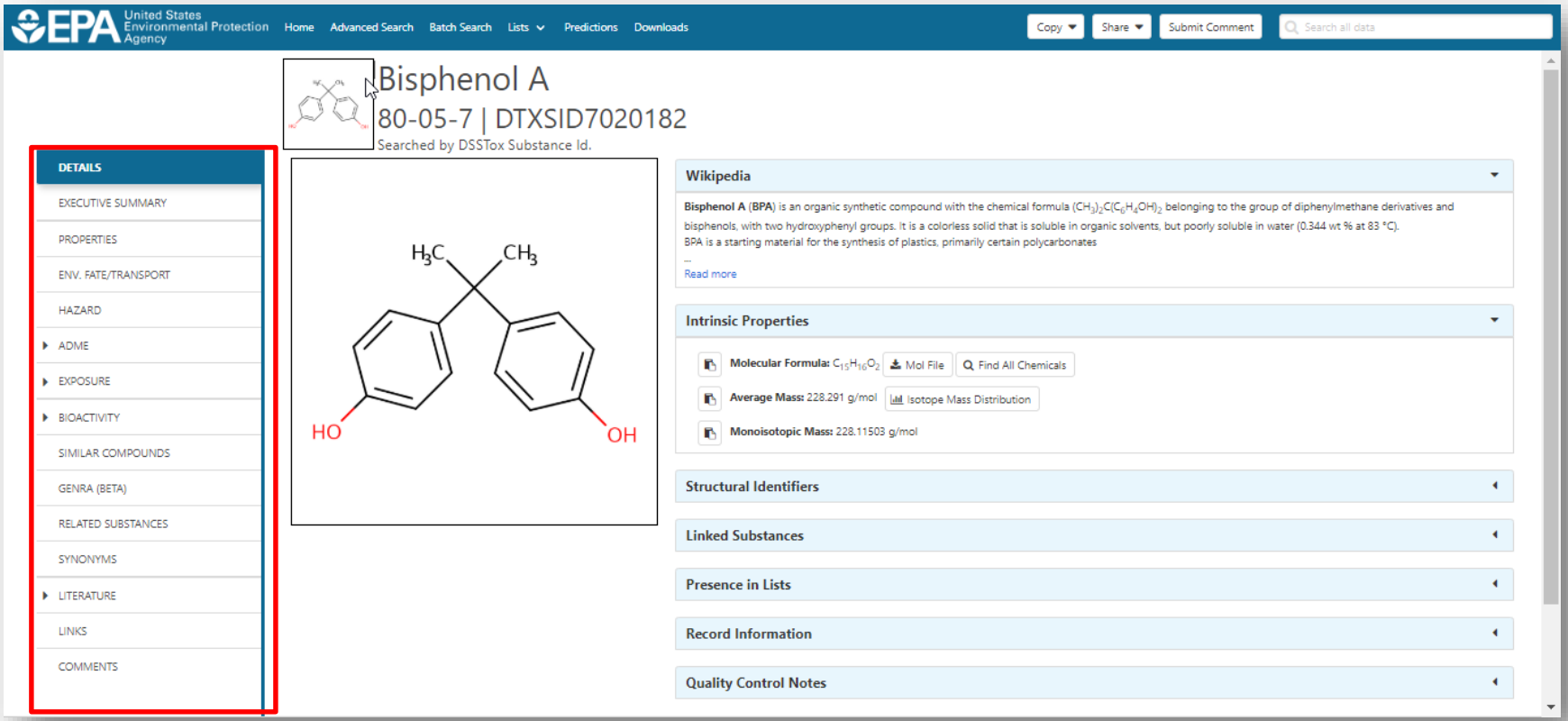
CompTox Chemicals Dashboard

https://comptox.epa.gov/dashboard

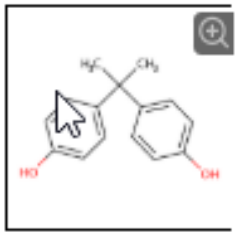
875k Chemical Substances



Detailed Chemical Pages



Experimental and Predicted Data



Bisphenol A

80-05-7 | DTXSID7020182

Searched by DSSTox Substance Id.

Property

Summary

Download Columns

Property	Experimental average	Predicted average
LogP: Octanol-Water	3.32 (1)	3.29
Melting Point	155 (7)	139
Boiling Point	200 (1)	363
Water Solubility	5.26e-4 (1)	9.62e-4
Vapor Pressure	-	8.37e-7
Flash Point	-	190

TEST and OPERA Predictions

EPA United States Environmental Protection Agency

- DETAILS
- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS
- COMMENTS

Toxicity Estimation Software Tool (TEST)

On this page:

- QSAR Methodologies
- What's New in Version 4.2.1?
- Prior Version History
- System Requirements
- Installation Instructions
- Publications
- Get Email Alerts

The Toxicity Estimation Software Tool (TEST) was developed to allow users to easily estimate the toxicity of chemicals using Quantitative Structure Activity Relationships (QSARs) methodologies. QSARs are mathematical models used to predict measures of toxicity from the physical characteristics of the structure of chemicals (known as molecular descriptors). Simple QSAR models calculate the toxicity of chemicals using a simple linear function of molecular descriptors:

Ask a Technical Expert

Got a question about our research model? Want to give us feedback? Contact a technical expert about TEST.

Mansouri *et al. J Cheminform (2018) 10:10*
https://doi.org/10.1186/s13321-018-0263-1

Journal of Cheminformatics

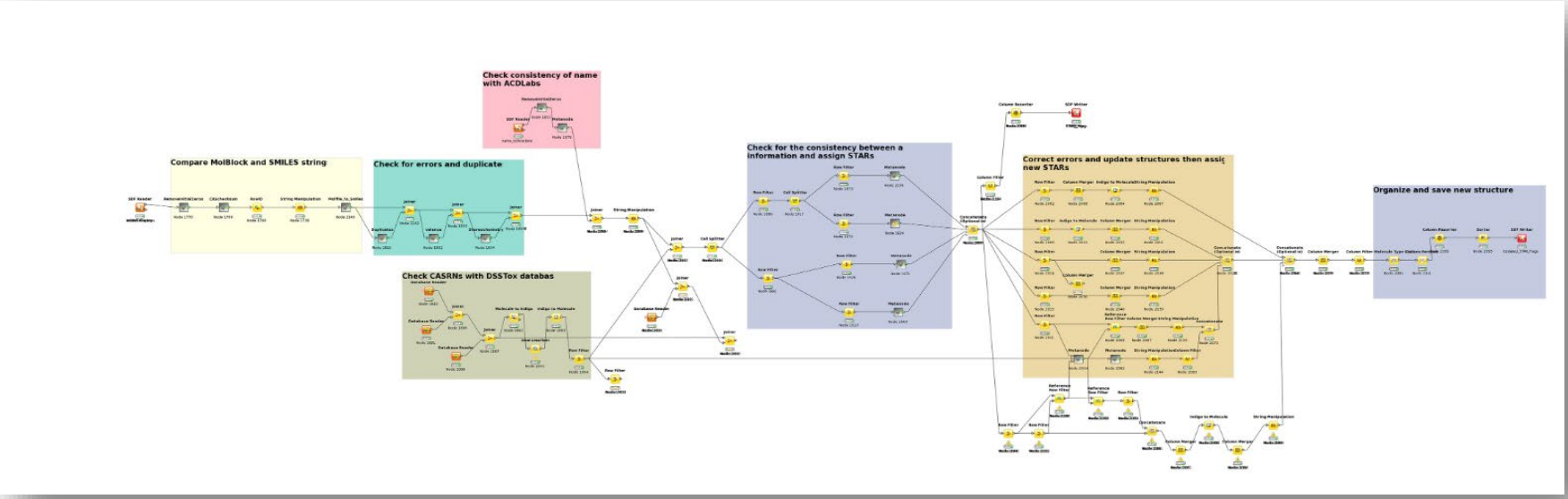
RESEARCH ARTICLE Open Access

OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri[1,2,3*], Chris M. Grulke[1], Richard S. Judson[1] and Antony J. Williams[1]

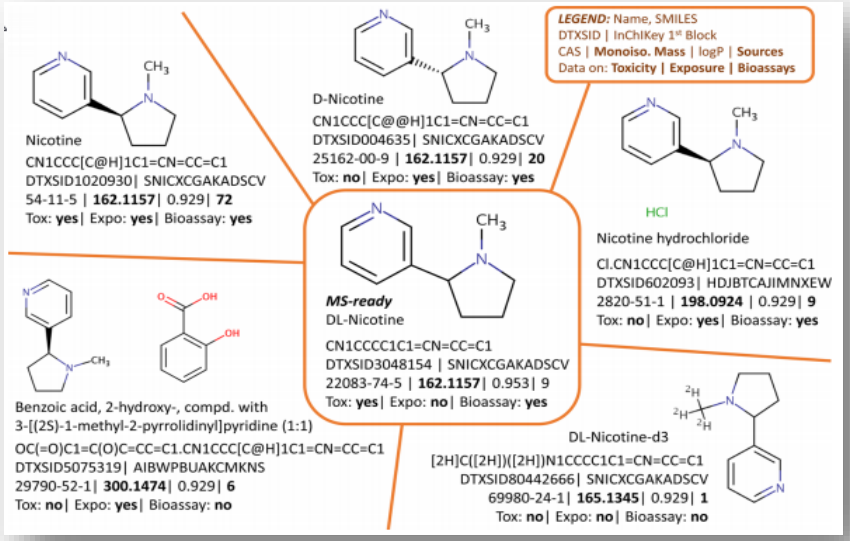
Standardization Approaches

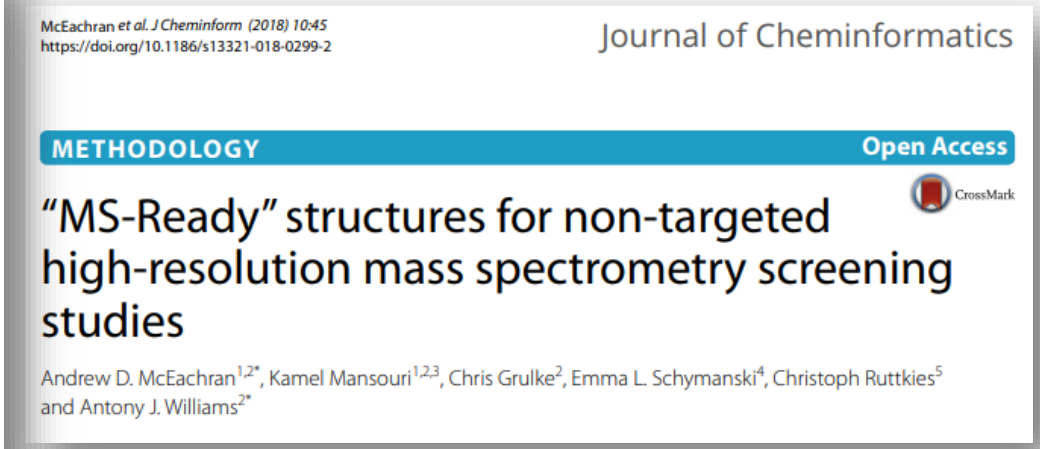
- "QSAR-ready" forms of chemicals are need for pushing through predictive models
  - OPERA predictions
  - TEST predictions
  - All QSAR models we build



Standardization Approaches

- We standardize chemicals to enable searches
- "MS-ready" forms to facilitate structure ID
  - Separate multi-component chemicals
  - Destereo, desalt, de-isotope





- Tweak algorithms and bulk process all data

Present status
For every data release…

- Presently every data release requires
  - Produce standardized data (MS and QSAR-ready)
  - QSAR predictions – TEST and OPERA
  - Systematic name generation
- Registration of data into databases
- Time-consuming and lots of overhead
- Not yet supporting other models for high-throughput

TEST Web Services

https://www.epa.gov/sites/production/files/2018-08/documents/webtest_users_guide.pdf

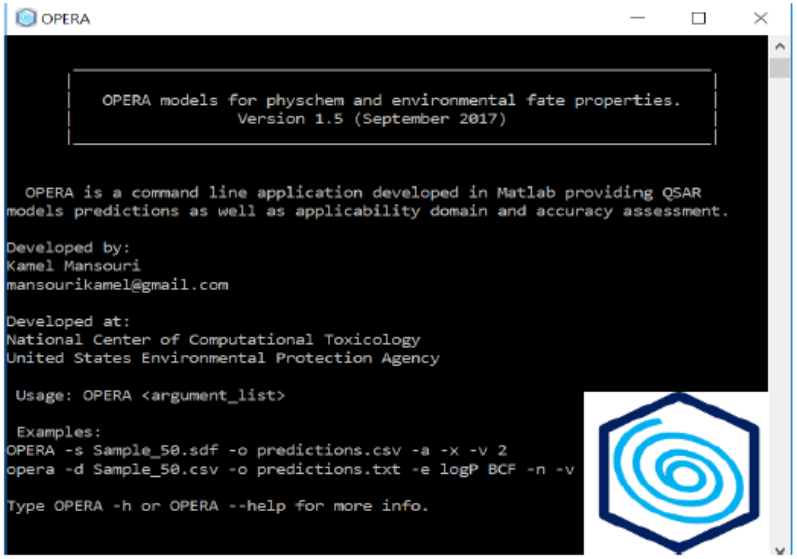
Integration to registration system required for high-throughout processing

OPERA Standalone Application

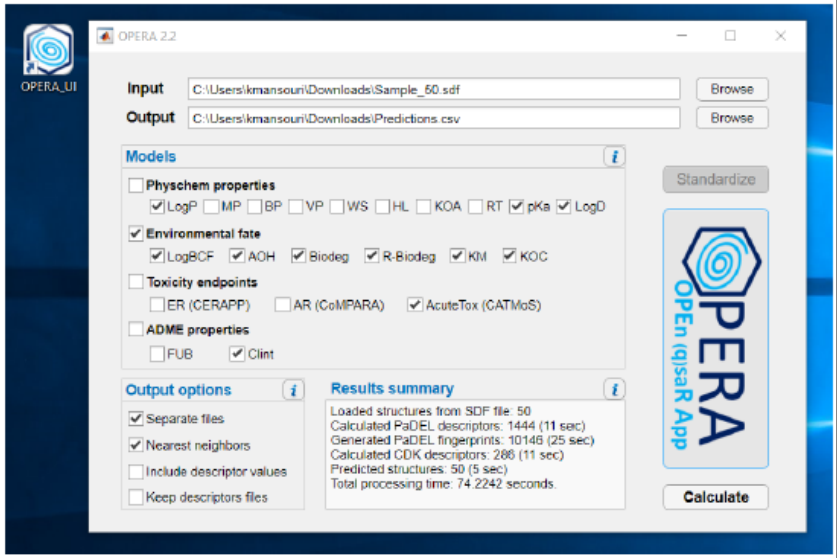
But services are required

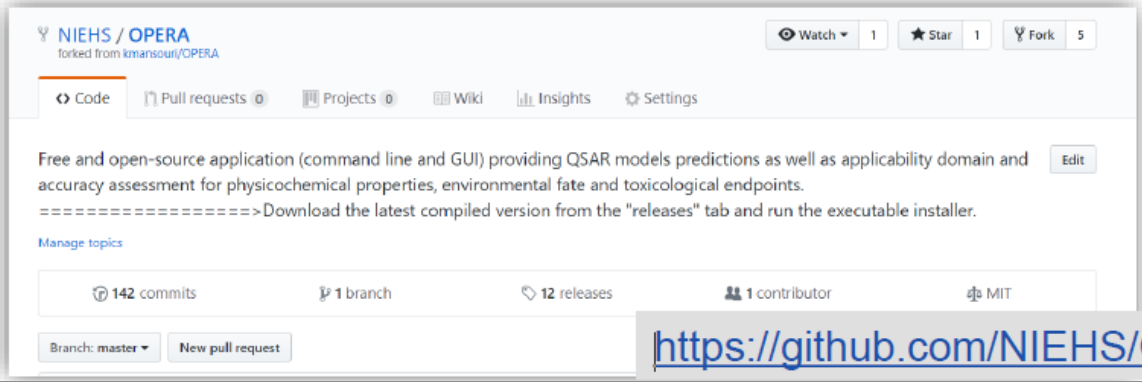
United States Environmental Protection Agency

Command line



Graphical User Interface

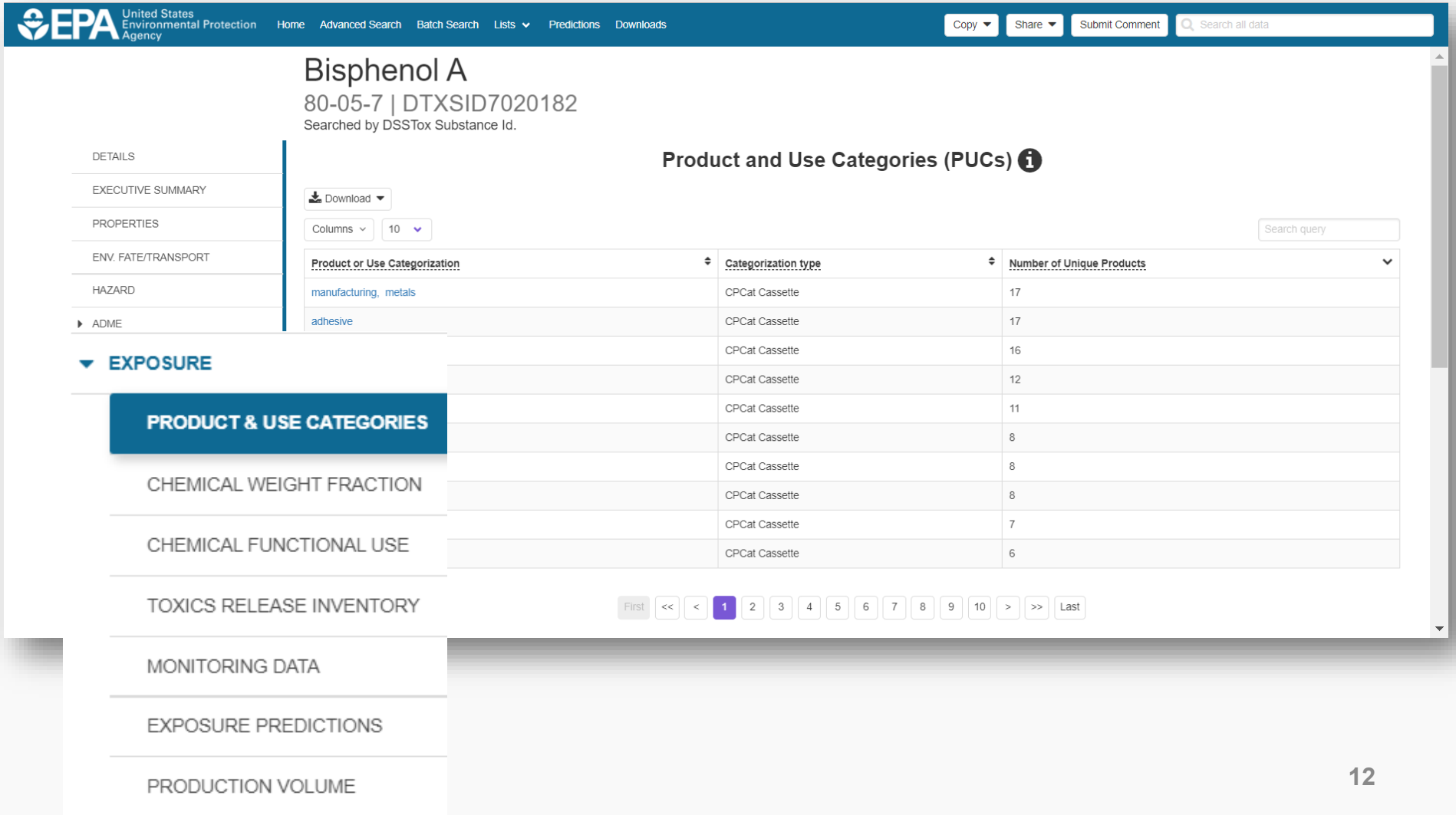




https://github.com/NIEHS/OPERA

Exposure Data

Sources of Exposure to Chemicals



Exposure Modeling

Rapid Exposure Predictions

Pictured Above: Farfield Exposure Examples

Pictured Above: Examples of Nearfield Exposure, Consumer Use and Indoor

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, enhances and evaluates two well-known exposure models to provide exposure predictions.

Farfield Exposure Models

Farfield exposure models are used to predict exposure from chemicals that are released into the outdoor environment through industrial releases. ExpoCast uses "off-the-shelf" models, USETox and RAIDAR, to estimate outdoor environment exposures. These models estimate the average amount of chemical that gets into the air, water, and soil. The estimates from these models are used in combination with the estimates from the nearfield models to make exposure predictions.

Nearfield Exposure Models

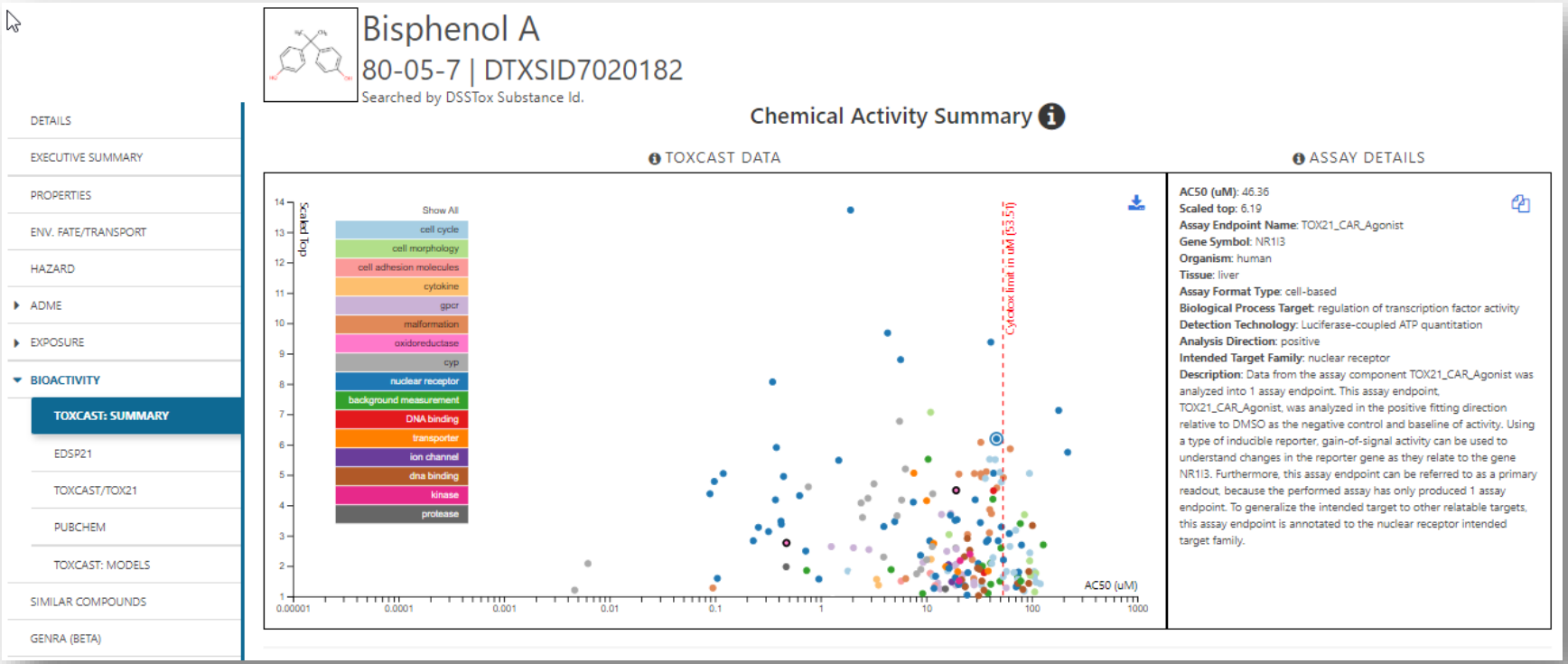
Nearfield exposure models provide estimates of exposure to chemicals used in consumer products and in-home products. The model used to estimate the range of total chemical exposures in a population is the EPA's Stochastic Human Exposure and Dose Simulation (SHEDS) model. There is a SHEDS high-throughput model that estimates exposure for thousands of chemicals, and a more precise traditional SHEDS model which needs more input data to make more accurate exposure predictions.

In vitro bioactivity data

In Vitro Bioassay Screening

ToxCast and Tox21

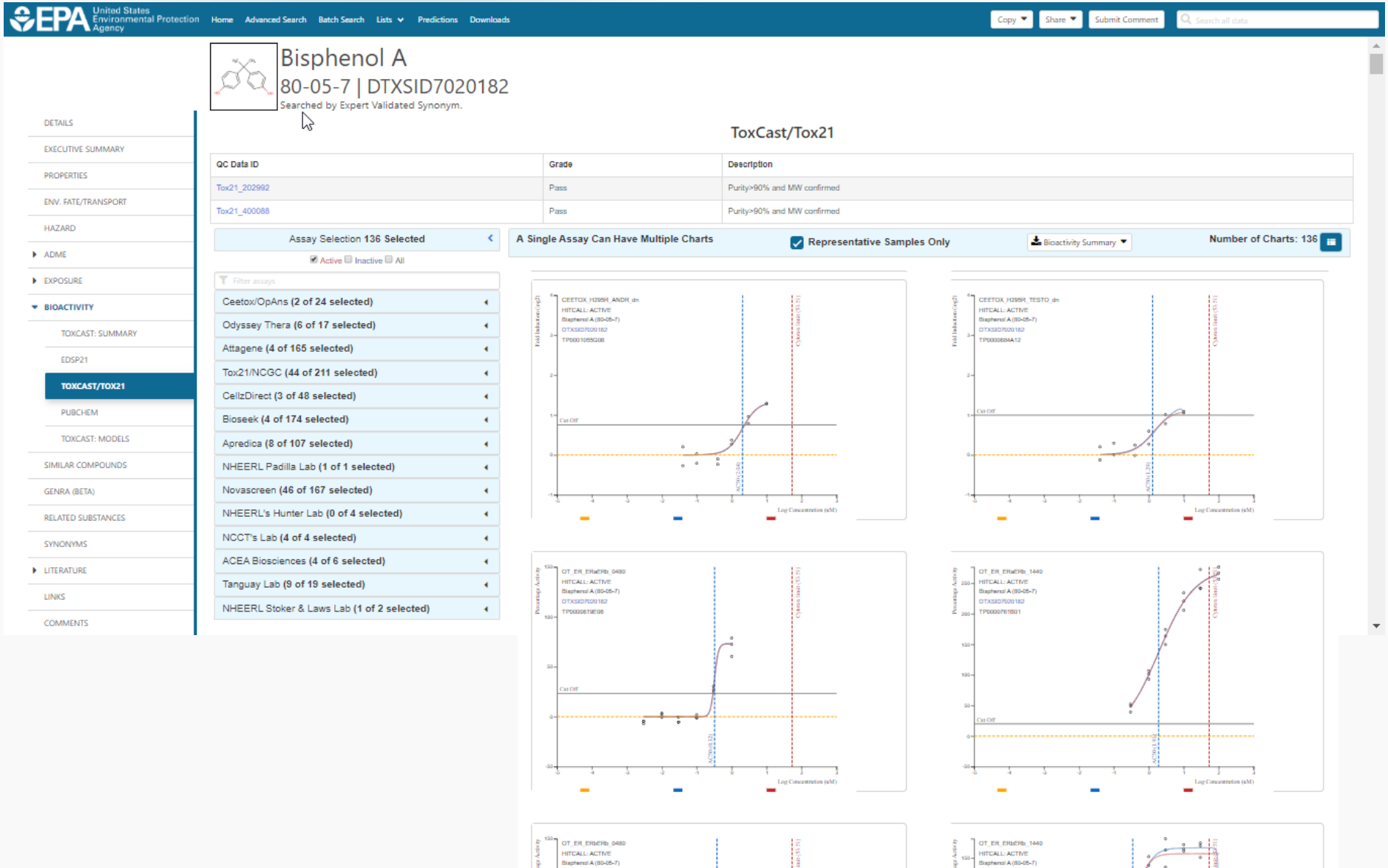
EPA United States Environmental Protection Agency



In Vitro Bioassay Screening

ToxCast and Tox21

EPA United States Environmental Protection Agency



Bioactivity: Downloadable Data

https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data

EPA United States Environmental Protection Agency

Exploring ToxCast Data: Downloadable Data

The results after processing through the Pipeline are available on the ToxCast Dashboard, and for most users EPA recommends accessing the data there.

- ToxCast Chemicals
- ToxCast Assays

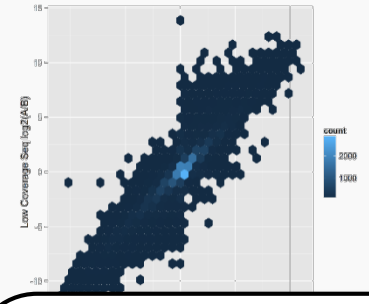
Resources

- Toxicity Forecaster (ToxCast) Fact Sheet
- ToxCast Publications
- ToxCast Citation
- About ToxCast

ToxCast Data and Information

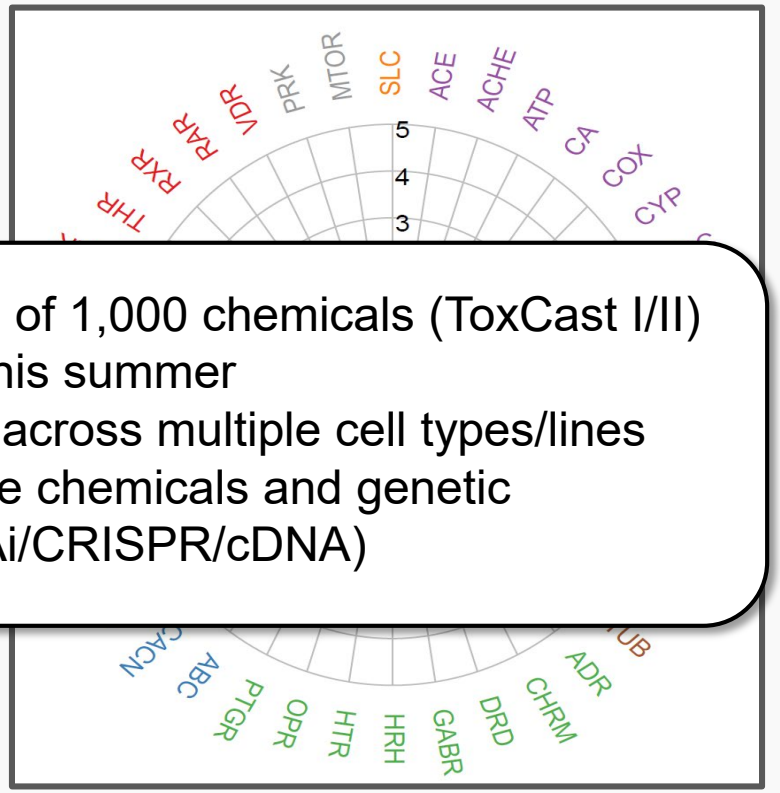
- **ToxCast & Tox21 Summary Files.** Data for a single chemical endpoint pair for thousands of chemicals and assay endpoints for 20 variables such as the activity or hit call, activity concentrations, whether the chemical was tested in a specific assay, etc.
  - Download ToxCast Summary Information
  - Download ReadMe
- **ToxCast & Tox21 Data Spreadsheet.** A spreadsheet of EPA's analysis of the chemicals screened through ToxCast and the Tox21 collaboration which includes EPA's activity calls from the screening of over 1,800 chemicals.
  - Download Data
  - Download ReadMe
- **ToxCast Data Pipeline R Package.** The R computer programming package used to process and model all EPA ToxCast and Tox21 chemical screening data. The files include the R programming package as well as documents that provide overviews of the data analysis pipeline used and the R package. Users will need experience with R to use these files.
  - Download Package
  - TCPL Overview

Transcriptomics Data will Deliver Terabytes of Data for Analysis

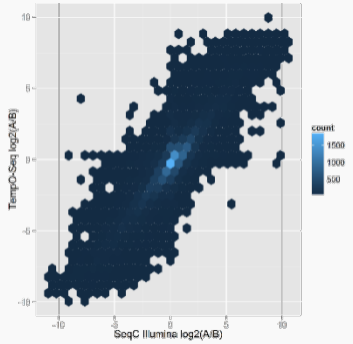


TruSeq
r^2 0.74

MOA Analysis Pipeline



- 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

"High-Throughput Searching"

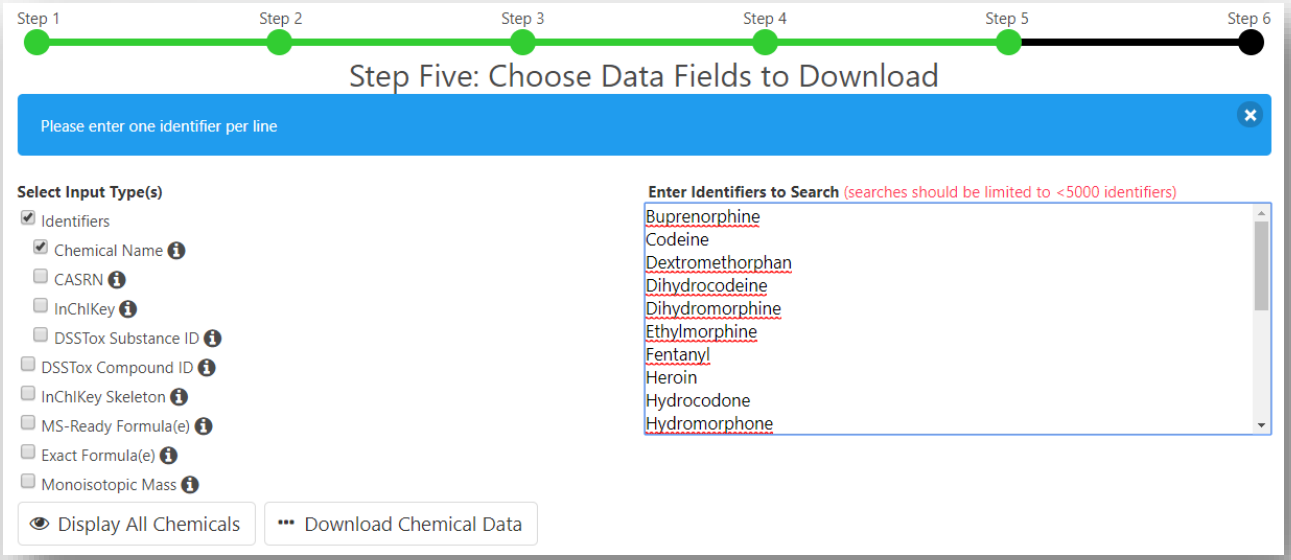
Batch Searching

- Singleton searches are useful but people generally want data on LOTS of chemicals!

- Typical questions
  - What is the list of chemicals for the formula $C_xH_yO_z$
  - What is the list of chemicals for a mass +/- error
  - Can I get chemical lists in Excel files? In SDF files?
  - Can I include properties in the download file?

Batch Search Names

Buprenorphine
Codeine
Dextromethorphan
Dihydrocodeine
Dihydromorphine
Ethylmorphine
Fentanyl
Heroin
Hydrocodone
Hydromorphone
Ketamine
Meperidine
Methadone
Morphine
Morphinone
Naloxone
Naltriben
Oxycodone
Oxymorphone
Propoxyphene
Sufentanil
Tramadol



Excel Download

INPUT	FOUND_BY	DTXSID
Buprenorphine	Approved Name	DTXSID2022705
Codeine	Approved Name	DTXSID2020341
Dextromethorphan	Approved Name	DTXSID3022908
Dihydrocodeine	Approved Name	DTXSID5022936
Dihydromorphine	Approved Name	DTXSID7048908
Ethylmorphine	Approved Name	DTXSID1046760
Fentanyl	Approved Name	DTXSID9023049
Heroin	Synonym	DTXSID6046761
Hydrocodone	Approved Name	DTXSID8023131
Hydromorphone	Approved Name	DTXSID8023133
Ketamine	Approved Name	DTXSID8023187
Meperidine	Approved Name	DTXSID9023253
Methadone	Approved Name	DTXSID7023273
Morphine	Approved Name	DTXSID9023336

Add Other Data of Interest

Chemical Identifiers

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

Structures

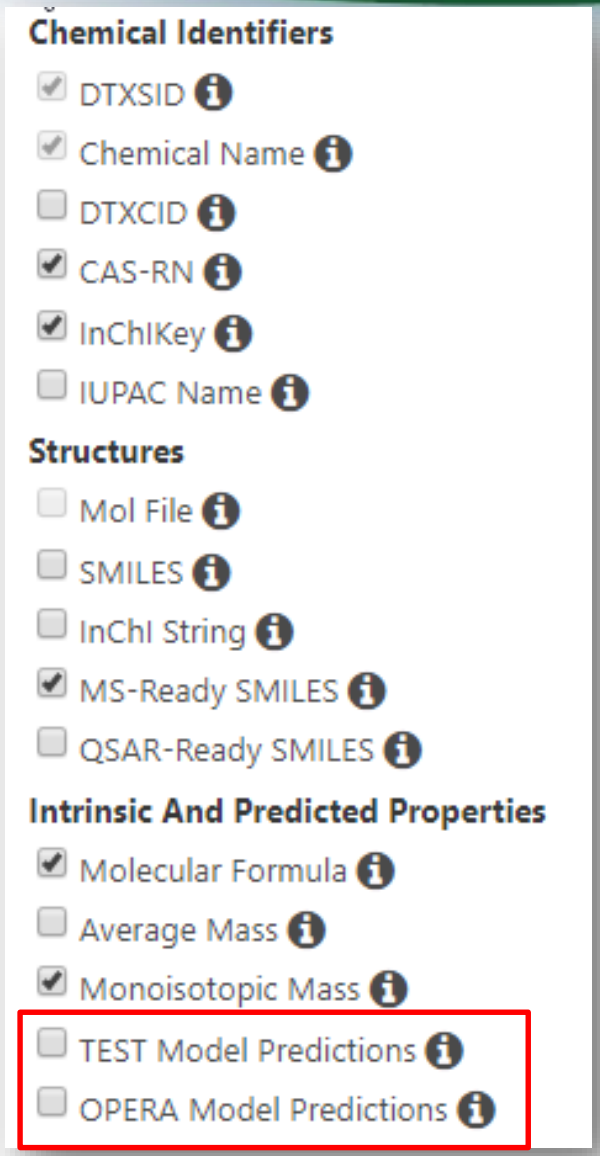
- [ ] Mol File
- [ ] SMILES
- [ ] InChI String
- [x] MS-Ready SMILES
- [ ] QSAR-Ready SMILES

Intrinsic And Predicted Properties

- [x] Molecular Formula
- [ ] Average Mass
- [x] Monoisotopic Mass
- [ ] TEST Model Predictions
- [ ] OPERA Model Predictions

INPUT	DTXSID	CASRN	MOLECULAR_F	MONOISOTOPIC	MS_READY_SMI
Buprenorph	DTXSID202	52485-79-7	C29H41NO4	467.3035588	[H]C12CC3=C4C(
Codeine	DTXSID202	76-57-3	C18H21NO3	299.1521435	[H]C12CC3=C4C(
Dextrometh	DTXSID302	125-71-3	C18H25NO	271.1936144	[H]C12CC3=C(C=
Dihydrocod	DTXSID502	125-28-0	C18H23NO3	301.1677936	[H]C12CC3=C4C(
Dihydromor	DTXSID704	509-60-4	C17H21NO3	287.1521435	[H]C12CC3=C4C(
Ethylmorph	DTXSID104	76-58-4	C19H23NO3	313.1677936	[H]C12CC3=C4C(
Fentanyl	DTXSID902	437-38-7	C22H28N2O	336.2201635	CCC(=O)N(C1CC
Heroin	DTXSID604	561-27-3	C21H23NO5	369.1576228	[H]C12CC3=C4C(
Hydrocodor	DTXSID802	125-29-1	C18H21NO3	299.1521435	[H]C12CC3=C4C(
Hydromorph	DTXSID802	466-99-9	C17H19NO3	285.1364935	[H]C12CC3=C4C(
Ketamine	DTXSID802	6740-88-1	C13H16ClNO	237.0920418	CNC1(CCCCC1=(
Meperidine	DTXSID902	57-42-1	C15H21NO2	247.1572289	CCOC(=O)C1(CC
Methadone	DTXSID702	76-99-3	C21H27NO	309.2092645	CCC(=O)C(CC(C)
Morphine	DTXSID902	57-27-2	C17H19NO3	285.1364935	[H]C12CC3=C4C(
Morphinone	DTXSID501	467-02-7	C17H17NO3	283.1208434	[H]C12CC3=C4C(
Naloxone	DTXSID802	465-65-6	C19H21NO4	327.1470582	[H]C12CC3=C4C(
Naltriben	-	-	-	-	-
Oxycodone	DTXSID502	76-42-6	C18H21NO4	315.1470582	[H]C12CC3=C4C(
Oxymorpho	DTXSID502	76-41-5	C17H19NO4	301.1314081	[H]C12CC3=C4C(
Propoxyphe	DTXSID102	469-62-5	C22H29NO2	339.2198292	CCC(=O)OC(CC1
Sufentanil	DTXSID602	56030-54-7	C22H30N2O2S	386.2027994	CCC(=O)N(C1=C
Tramadol	DTXSID908	27203-92-5	C16H25NO2	263.188529	COC1=CC=CC(=

Access to other data

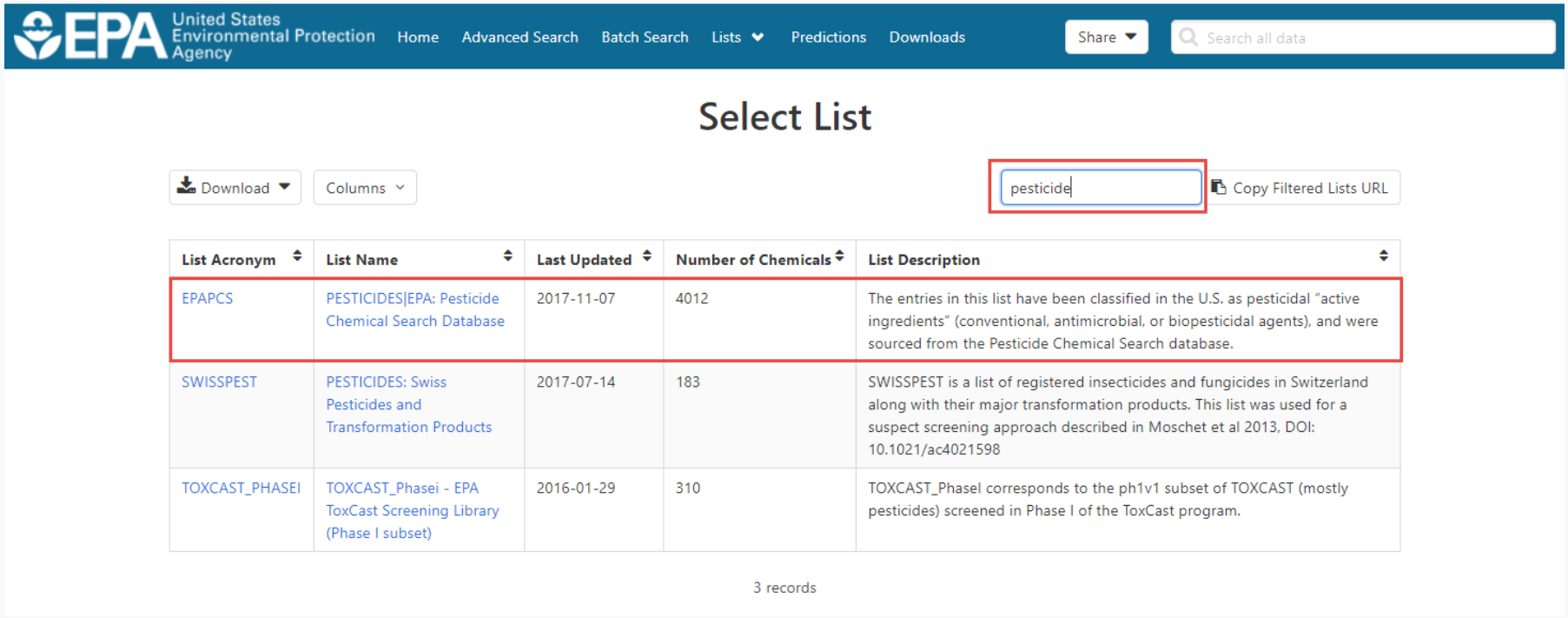


- Pre-predicted TEST and OPERA data

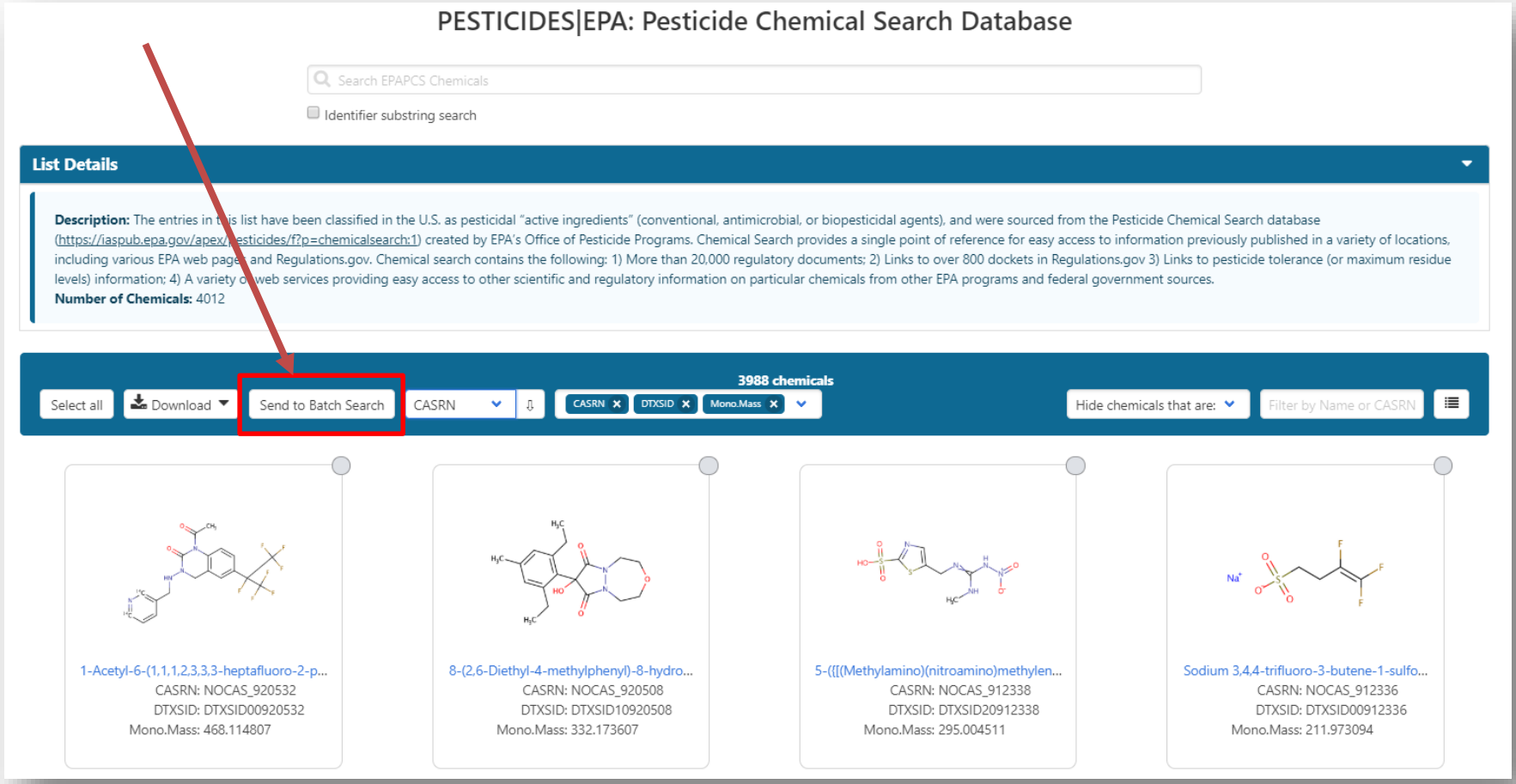
C	D	E	F	G	H	I	J	K
DTXSID	PREFERRE	BIOCONCE	BOILING_P	48HR_DAPI	DENSITY_C	DEVTOX_T	96HR_FATH	FLASH_PO
DTXSID7020009	Acetonitrile	2.04	74.44	0.0064565	0.92	0.54	0.0075858	22.46
DTXSID5020023	Acrolein	-	52.443	0.0006427	0.841	0.641	0.0001581	8.913
DTXSID5020029	Acrylonitrile	-	107.213	0.0002877	0.901	0.642	0.0002193	19.044
DTXSID8020044	Allyl alcohol	1.87932	83.788	0.000975	0.858	0.719	0.0011912	21.313
DTXSID8020090	Aniline	5.3827	192.271	3.184E-05	1.108	0.529	0.000869	98.93
DTXSID8020173	bis(Chlorom	3.47536	102.976	0.001205	1.233	0.308	0.0017824	35.353
DTXSID8020250	Carbon tetra	17.2982	91.367	0.000104	1.601	0.777	0.00014	21.094
DTXSID1020273	Chlorine	-	-	-	-	-	-	-
DTXSID4020292	Chloroaceta	-	73.604	0.0011169	1.162	0.797	0.0002897	49.401
DTXSID9020297	(2-Chlorobe	-	278.468	4.808E-06	1.286	0.859	2.559E-06	164.565
DTXSID4020298	Chlorobenze	108.393	150.5	6.699E-05	1.123	0.42	0.0001762	49.082
DTXSID1020306	Chloroform	6.71429	61.07	0.0008395	1.425	0.794	0.0007079	24.417
DTXSID6020307	Chloromethy	2.29087	69.641	0.0038726	1.126	0.426	0.0048306	16.663
DTXSID6020351	(E)-Crotonal	-	97.879	0.0004764	0.871	0.6	0.0001318	19.542
DTXSID2020422	1,1-Dichloro	8.97429	33.566	0.0006138	1.214	0.464	0.0007638	7.323
DTXSID6020515	N,N-Dimethy	0.677642	166.494	0.0048195	0.947	0.94	0.0141906	77.204
DTXSID1020516	1,1-Dimethy	-	97.361	0.002958	0.914	0.515	0.0185353	21.051
DTXSID1020566	Epichlorohyc	3.12608	102.31	0.000743	1.214	0.66	0.0005023	21.312
DTXSID8020599	Ethyleneimir	-	28.496	-	-	0.73	-	-2.929
DTXSID0020600	Ethylene oxi	0.724436	34.921	0.0043551	0.968	0.758	0.0037497	15.445
DTXSID6020646	Furan	4.88652	73.387	0.0020845	0.997	0.394	0.0021232	-1.928

So what about Pesticides???

Curated List of Pesticides



Curated List of Pesticides



Send to batch and select….

Intrinsic And Predicted Properties

- [ ] Molecular Formula
- [ ] Average Mass
- [ ] Monoisotopic Mass
- [ ] TEST Model Predictions
- [ ] OPERA Model Predictions

Metadata

- [ ] Curation Level Details
- [ ] NHANES/Predicted Exposure
- [ ] Data Sources
- [ ] Include ToxVal Data Availability
- [ ] Assay Hit Count
- [ ] Number of PubMed Articles
- [ ] PubChem Data Sources
- [ ] CPDat Product Occurrence Count
- [ ] IRIS
- [ ] PPRTV

3988 Pesticide….Batch Search

- A few seconds to assemble
  - ToxCast data - #actives/#assays and % active
  - # articles in PubMed
  - Links to IRIS or PPRTV reports
  - TEST or OPERA predictions
  - Availability of exposure data, presence in consumer products, EXPOCAST predictions

A	B	C	D	E	F	G	H	I	J	K	L	M
DTXSID	PREFERRED_NAME	EXPOCAST_MEI	EXPOCAST	NHANES	TOXVAL_DATA	TOXCAST_%_ACT	TOXCAST_	#PUBMED_	PUBCHEM_	CPDAT_COUI	IRIS_LINK	PPRTV_LINK
DTXSID2021105	Pentachloronitrobenzene	1.14e-07	Y	Y	Y	11.8	99/839	69	96	164	Y	-
DTXSID4022527	Propylparaben	1.4e-05	Y	Y	Y	13.77	99/719	201	121	1476	-	-
DTXSID4024064	Dinex	8.29e-08	Y	-	Y	42.13	99/235	-	35	5	Y	-
DTXSID0032493	Triadimenol	1.73e-08	Y	-	Y	10.54	98/930	163	74	83	-	-
DTXSID4032667	Esfenvalerate	1.7e-06	Y	-	Y	11.45	98/856	483	45	198	-	-
DTXSID6020561	Endrin	1.29e-07	Y	-	Y	14.02	98/699	284	16	98	Y	Y
DTXSID6025355	Glutaraldehyde	2.03e-05	Y	-	Y	14.35	98/683	6515	139	1144	-	-
DTXSID8032417	Isofenphos	1.87e-08	Y	-	Y	16.28	98/602	30	42	60	-	-
DTXSID6032352	Chlorpyrifos-methyl	1.07e-07	Y	Y	Y	11.27	97/861	72	50	116	-	-
DTXSID8020620	Fenthion	8.99e-08	Y	Y	Y	11.56	97/839	354	100	99	-	-
DTXSID2020189	FD&C Blue No. 1	0.000178	Y	-	Y	13.72	97/707	174	49	672	-	-
DTXSID7044843	Erythrosin B	6.3e-07	Y	-	-	24.25	97/400	14843	51	7	-	-
DTXSID5041778	Chloropropylate	1.05e-07	Y	-	Y	40.93	97/237	-	36	12	-	-
DTXSID5023900	Benomyl	1.11e-07	Y	-	Y	11.23	96/855	476	91	105	Y	-
DTXSID9020247	Carbaryl	5.61e-08	Y	Y	Y	11.51	96/834	1135	117	245	Y	-
DTXSID8024109	Flutolanil	1.63e-08	Y	-	Y	11.4	95/833	6	59	80	-	-
DTXSID1023998	Cypermethrin	1.62e-06	Y	Y	Y	10.78	94/872	1148	148	246	-	-
DTXSID2024242	Paclobutrazol	9.19e-08	Y	-	Y	11.11	94/846	139	-	40	Y	-
DTXSID1020807	2-Mercaptobenzothiazole	4.7e-05	Y	-	Y	12.82	94/733	111	181	86	-	Y

Other packages available from NCCT

High-Throughput Toxicokinetics

https://cran.r-project.org/web/packages/httk/index.html

EPA United States Environmental Protection Agency

Authors:	Robert G. Pearce, R. Woodrow Setzer, Cory L. Strope, Nisha S. Sipes, John F. Wambaugh
Title:	**httk: R Package for High-Throughput Toxicokinetics**

`httk: High-Throughput Toxicokinetics`

Functions and data tables for simulation and statistical analysis of chemical toxicokinetics ("TK") as in Pearce et al. (2017) <doi:10.18637/jss.v079.i04>. Chemical-specific in vitro data have been obtained from relatively high throughput experiments. Both physiologically-based ("PBTK") and empirical (e.g., one compartment) "TK" models can be parameterized for several hundred chemicals and multiple species. These models are solved efficiently, often using compiled (C-based) code. A Monte Carlo sampler is included for simulating biological variability (Ring et al., 2017 <doi:10.1016/j.envint.2017.06.004>) and measurement limitations. Calibrated methods are included for predicting tissue:plasma partition coefficients and volume of distribution (Pearce et al., 2017 <doi:10.1007/s10928-017-9548-7>). These functions and data provide a set of tools for in vitro-in vivo extrapolation ("IVIVE") of high throughput screening data (e.g., Tox21, ToxCast) to real-world exposures via reverse dosimetry (also known as "RTK") (Wetmore et al., 2015 <doi:10.1093/toxsci/kfv171>).

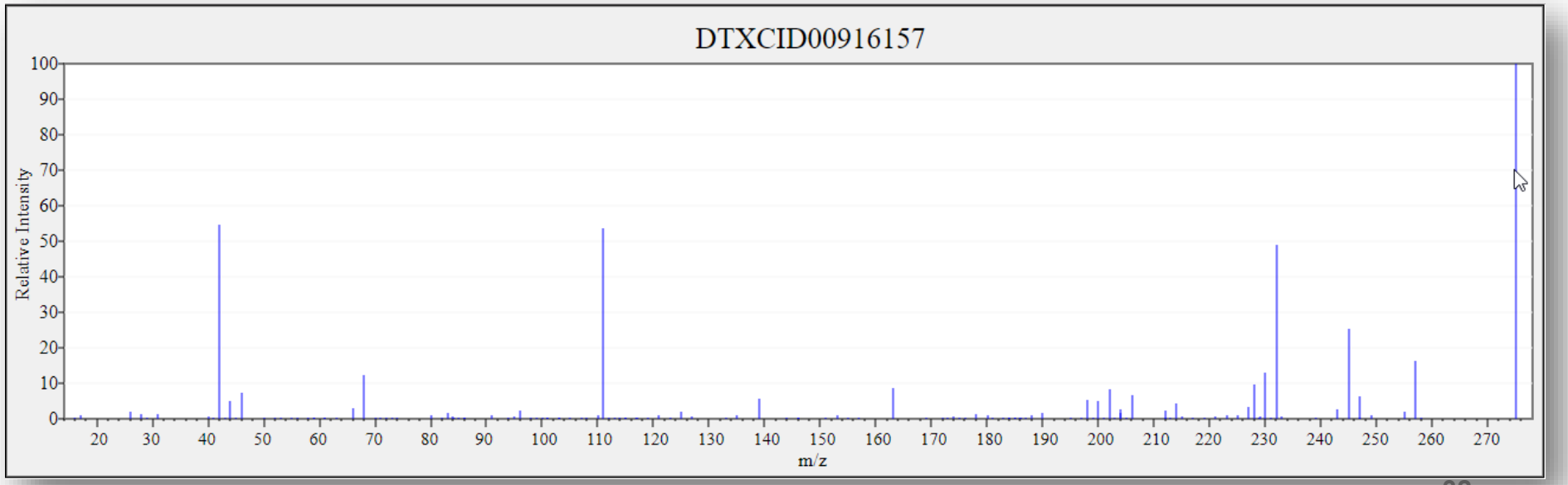
EPA United States Environmental Protection Agency

Work in Progress

Predicted Mass Spectra

http://cfmid.wishartlab.com/

- MS/MS spectra prediction for ESI+, ESI-, and EI to support candidate ranking for mass/formula searches



High-Throughput Spectral Prediction CFM-ID Predicted Library

- Predictions already generated for >700k structures

Data Descriptor | OPEN | Published: 02 August 2019

Linking *in silico* MS/MS spectra with chemistry data to improve identification of unknowns

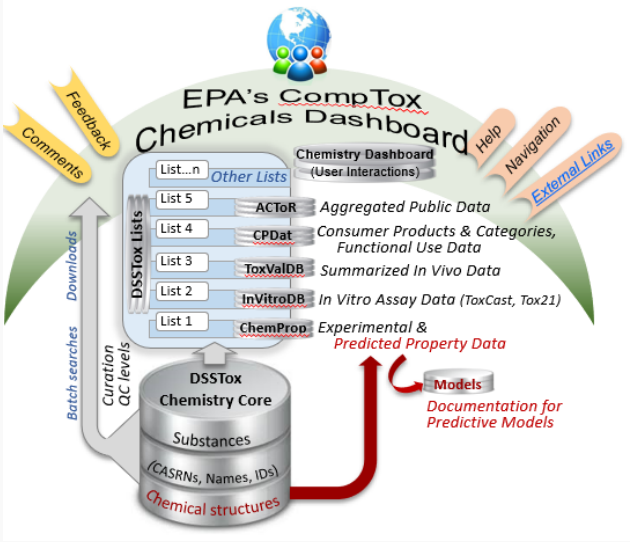
Andrew D. McEachran, Ilya Balabin, Tommy Cathey, Thomas R. Transue, Hussein Al-Ghoul, Chris Grulke, Jon R. Sobus & Antony J. Williams

Scientific Data **6**, Article number: 141 (2019) | Download Citation

- Ideally CFM-ID spectra will be generated for all chemicals added to our registration system

Conclusion

- High-throughput data generation comes in many forms at the National Center for Computational Toxicology
  - High-throughput bioactivity screening – ToxCast/Tox21
  - High-throughput transcriptomics data
  - High-throughput standardization (MS and QSAR-ready processing)
  - High-throughput QSAR prediction (e.g. TEST and OPERA models)
  - High-throughput Mass Spec fragmentation prediction (CFM-ID)
- Pipelining and versioning of data is in scope as future work



Acknowledgements

EPA-RTP

- *An enormous team of contributors from NCCT, especially the IT software development team*
- *Our curation team for their care and focus on data quality*
- *Multiple centers and laboratories across the EPA*
- *Many public domain databases and open data contributors*