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

EPA United States Environmental Protection Agency

Progress in Delivering Transparency in Research Data by the National Center for Computational Toxicology at the US-EPA

Antony J. Williams, Jeff Edwards, Chris Grulke and John Cowden

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

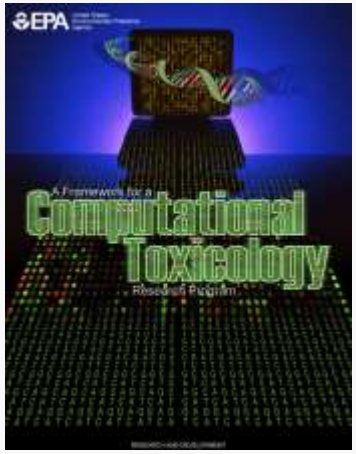
August 2018

ACS Fall Meeting, Boston

Disclaimer of Endorsement

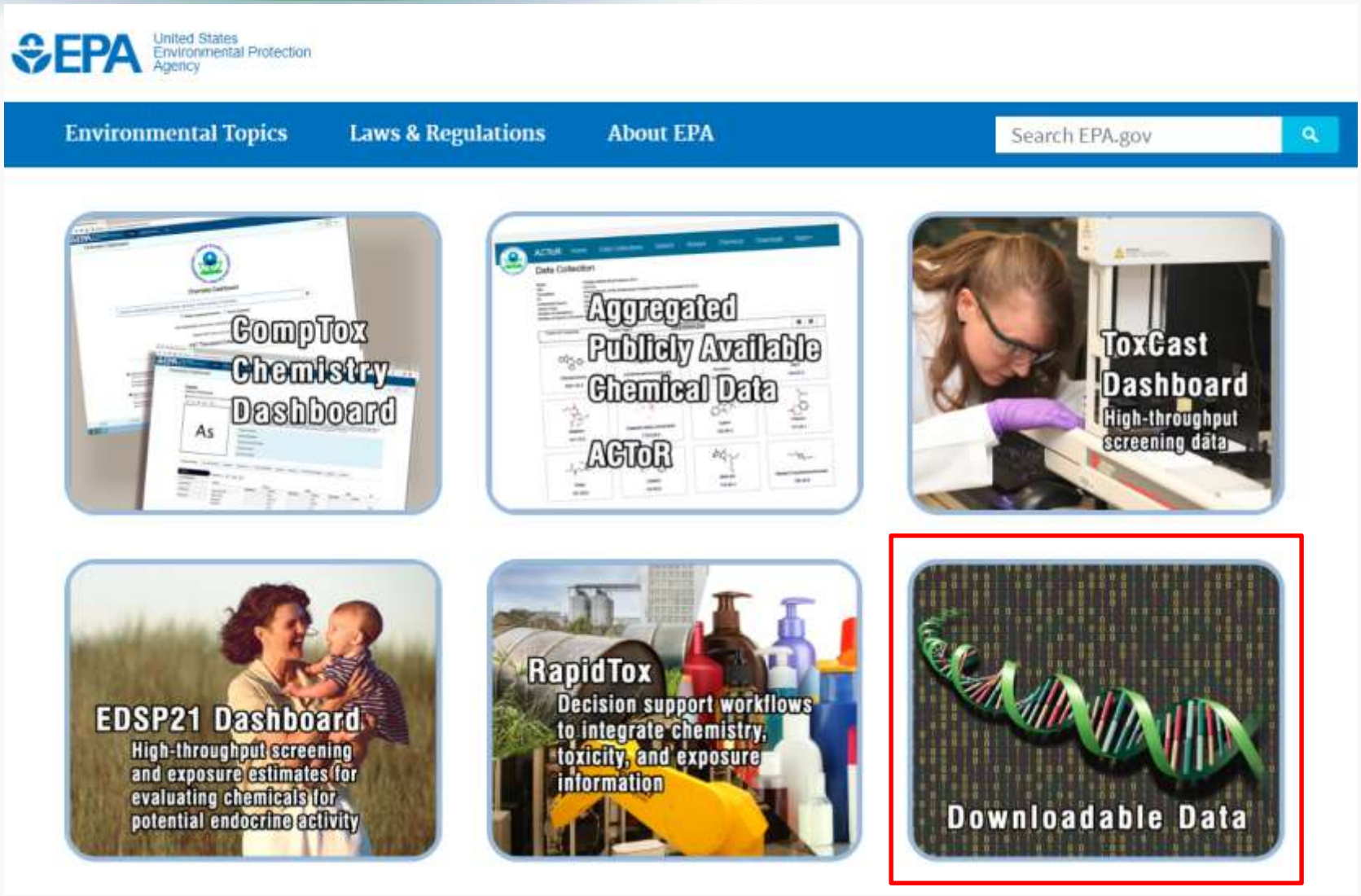
- Mention of or referral to commercial products or services, and/or links to non-EPA sites **does not imply official EPA endorsement** of or responsibility for the opinions, ideas, data, or products presented at those locations, or guarantee the validity of the information provided.

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
- Outputs: a lot of data, models, algorithms, software applications and publications
- Open Data – we want scientists to interrogate it, learn from it, develop understanding

The CompTox Portal
https://comptox.epa.gov/



Downloadable CompTox Data

https://www.epa.gov/chemical-research/downloadable-computational-toxicology-data

EPA United States Environmental Protection Agency

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. 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.

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 and metabolic clearance for hundreds of chemicals and incorporate these data into computer models. The high-throughput toxicokinetic data can be paired with the high-throughput screening data to estimate real-world exposures.

Sustainable Chemistry Data

EPA researchers use chemistry data such as chemical structures and physicochemical property information to evaluate thousands of chemicals for potential health effects.

- **Collaborative Estrogen Receptor Activity Prediction Project Data**: Data and supplemental files from CERAPP (A large-scale modeling project). CERAPP combined multiple models developed in collaboration with 17 groups in the United States and Europe to predict estrogen receptor activity of a common set of 32,464 chemical structures. *Quantitative structure-activity relationship* models and docking approaches were employed, to build a total of 40 categorical and 8 continuous models for binding, agonist, and antagonist ER activity.
  - Evaluation Set
  - Models
  - Prediction Set
  - Training Set
- **Chemistry Dashboard Data**: Data from the Chemistry Dashboard including the mappings between the DTXSIDs and the InChIStrings and Keys, SDF files containing all chemical structures and relevant information, and a file containing CAS Number, Preferred Chemical Name and DTXSID file.

Deliver Data for Reuse: DIFFERENT formats

EPA United States Environmental Protection Agency

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.

Downloads

Data Set	Description	Release Date	Database Version	Download
ToxCast & Tox21 Chemicals Distributed Structure-Searchable Toxicity Database (DSSTox files)	Chemical details for 8,599 unique substances (GSIDs) and DSSTox standard chemical fields (chemical name, CASRN, structure, etc.) for EPA ToxCast chemicals and the larger Tox21 chemical list. Also includes chemical mapping files and quality control grades for chemicals.	October 2015	DSSTox_20151019	ToxCast Chemicals: Data Management and Quality Considerations Overview Download ToxCast Chemical Information Download ReadMe
ToxCast & Tox21 high-throughput assay information	ToxCast high-throughput assay information including assay annotation user guide, assay target information, study design information and quality statistics on the assays.	October 2015	invitrodb_v2	Assay Annotation User Guide Download Assay Information

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

Open Data means Reuse For Science

EPA United States Environmental Protection Agency

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[§]

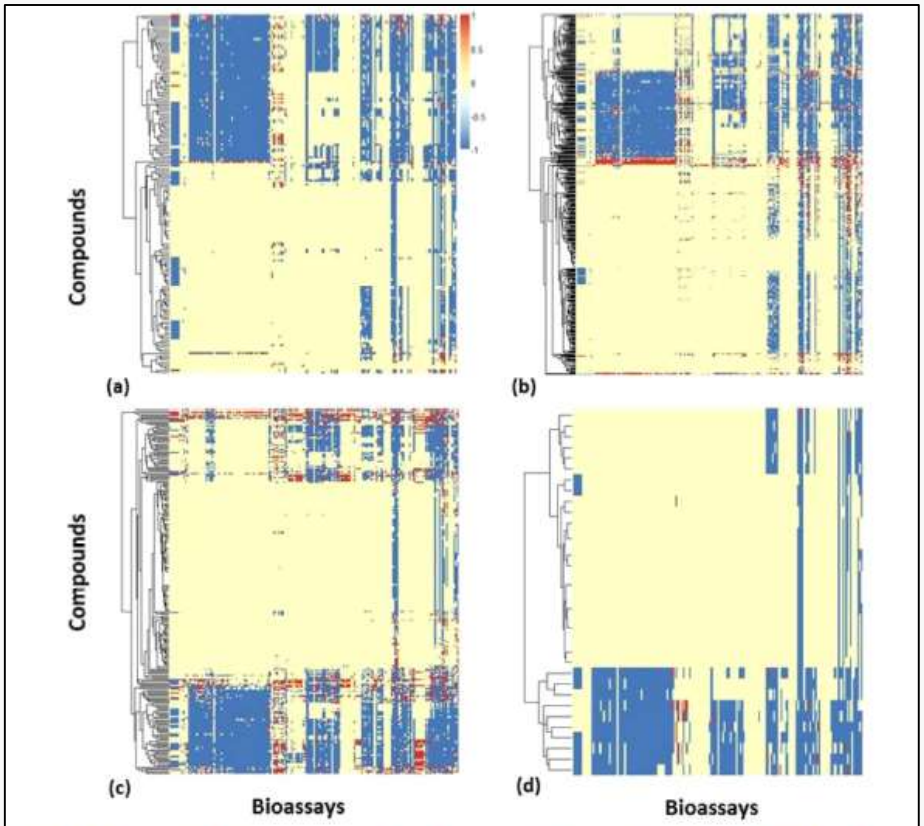
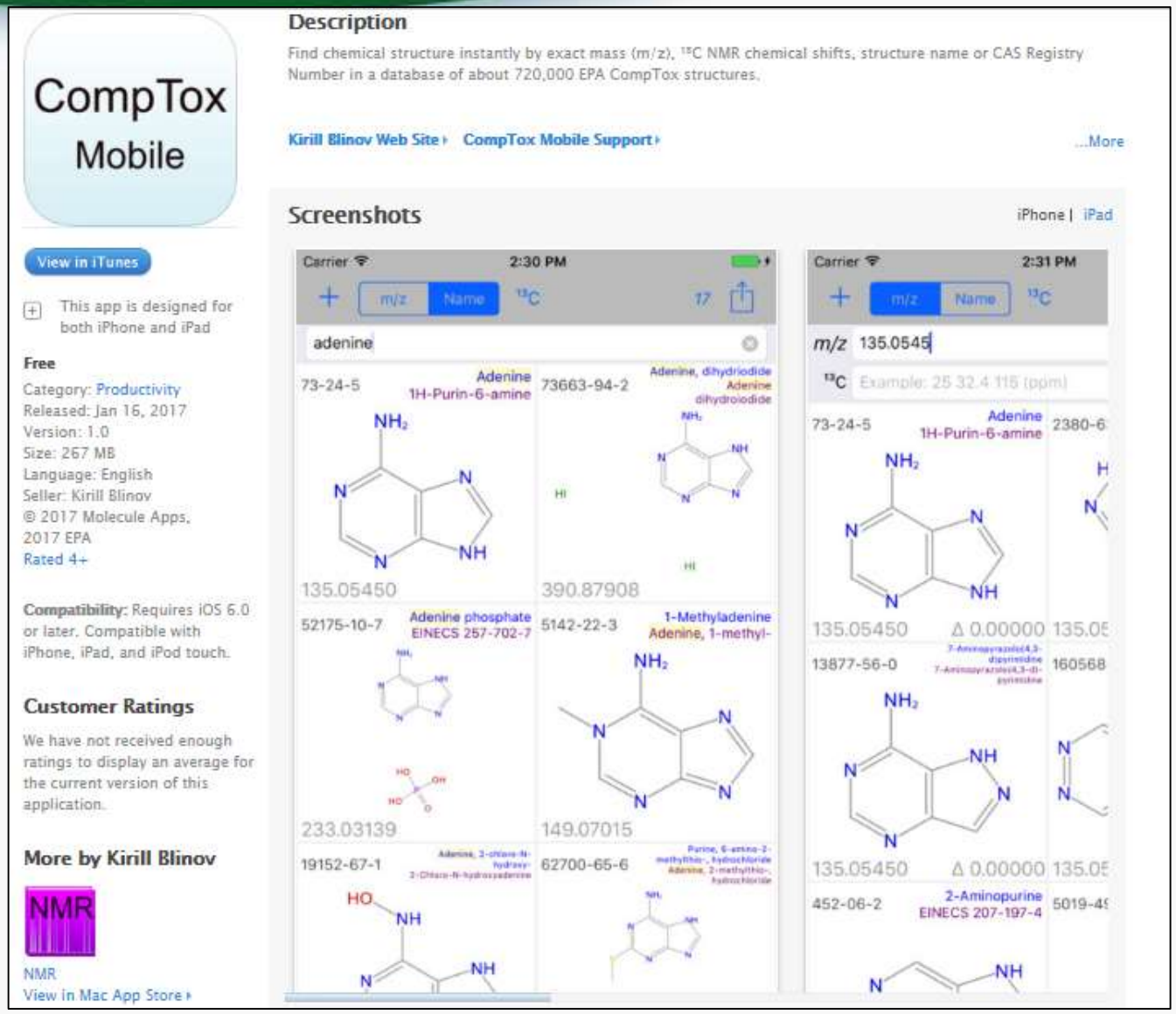


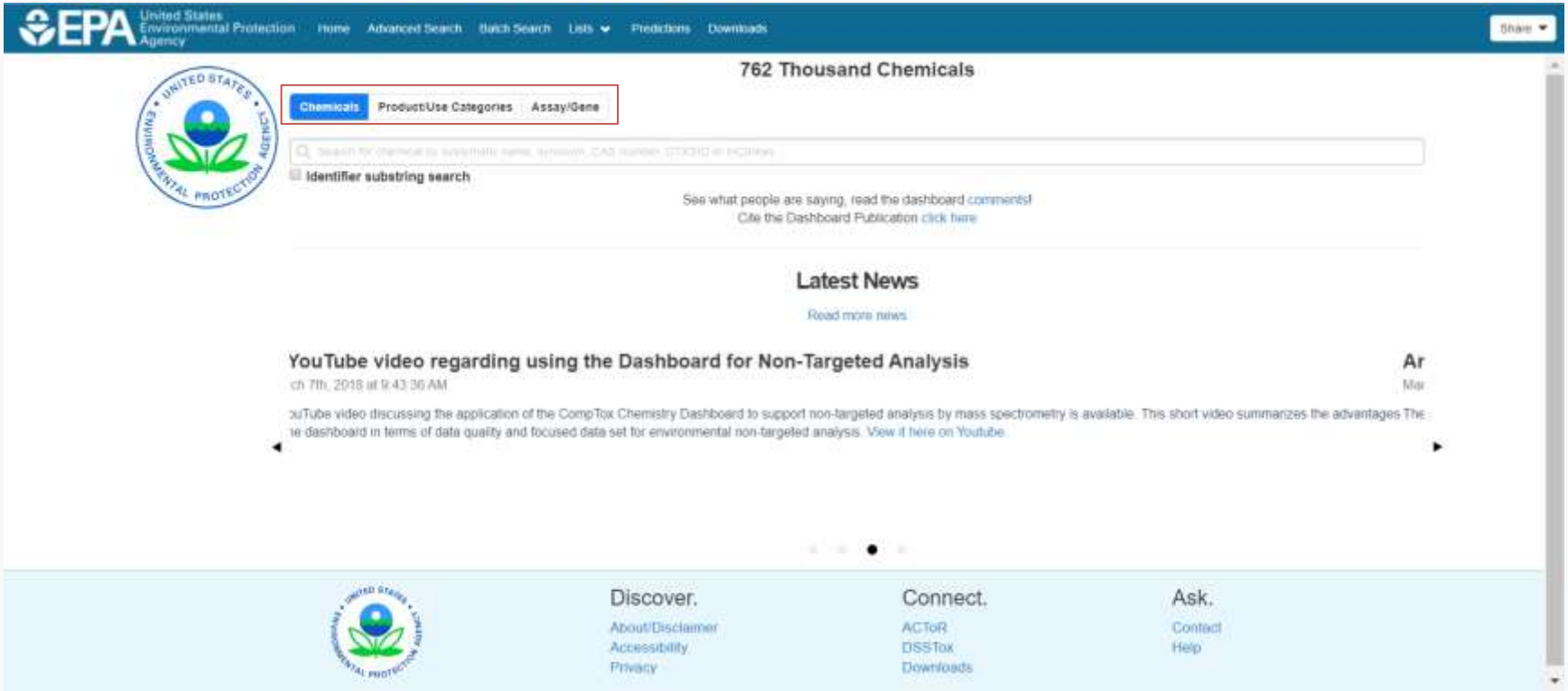
Figure 3. Response spaces of different ToxCast compound categories represented by the data obtained from 193 PubChem bioassays: (a) 171 consumer use chemicals (not including pharmaceuticals or pesticides), (b) 470 pesticides, (c) 245 pharmaceuticals, and (d) 34 phthalate plasticizers and alternatives.

Open Data means Reuse For Software

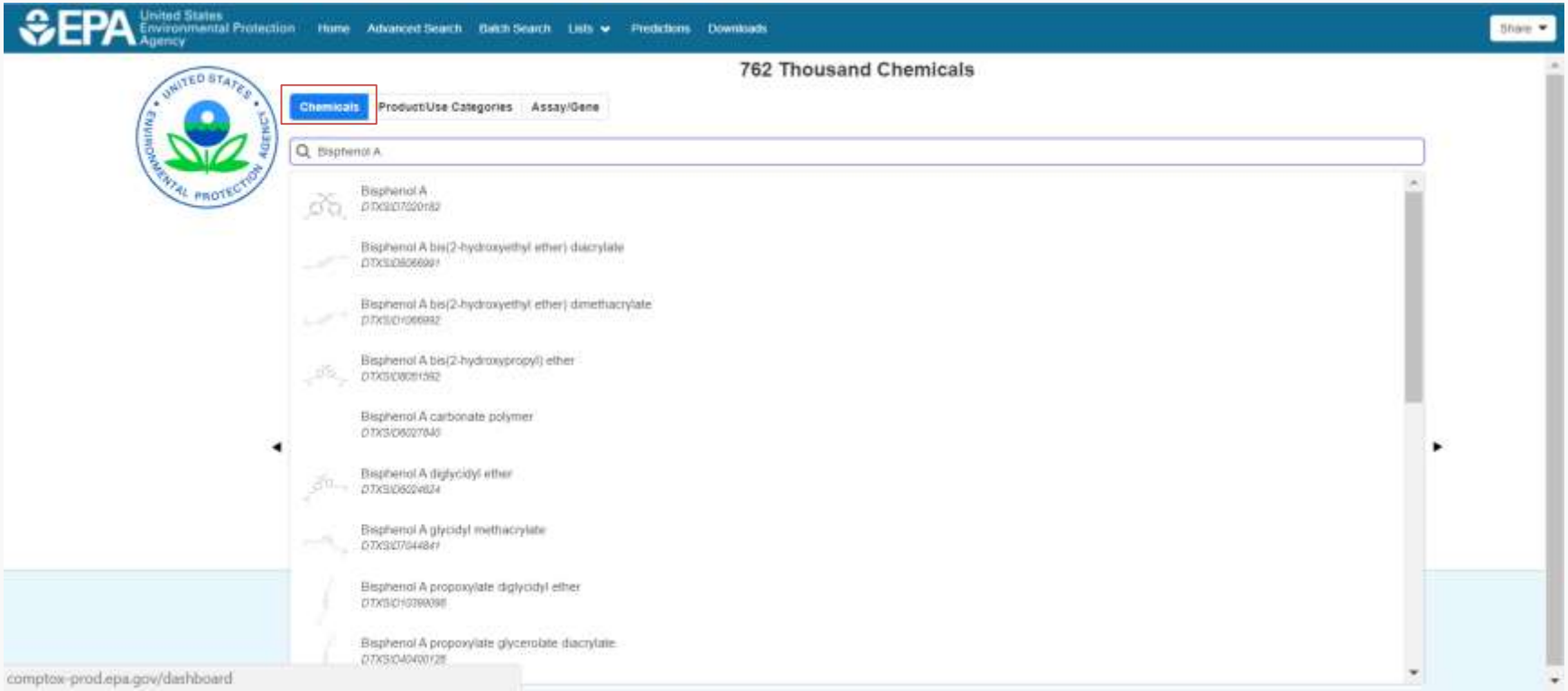


CompTox Dashboard

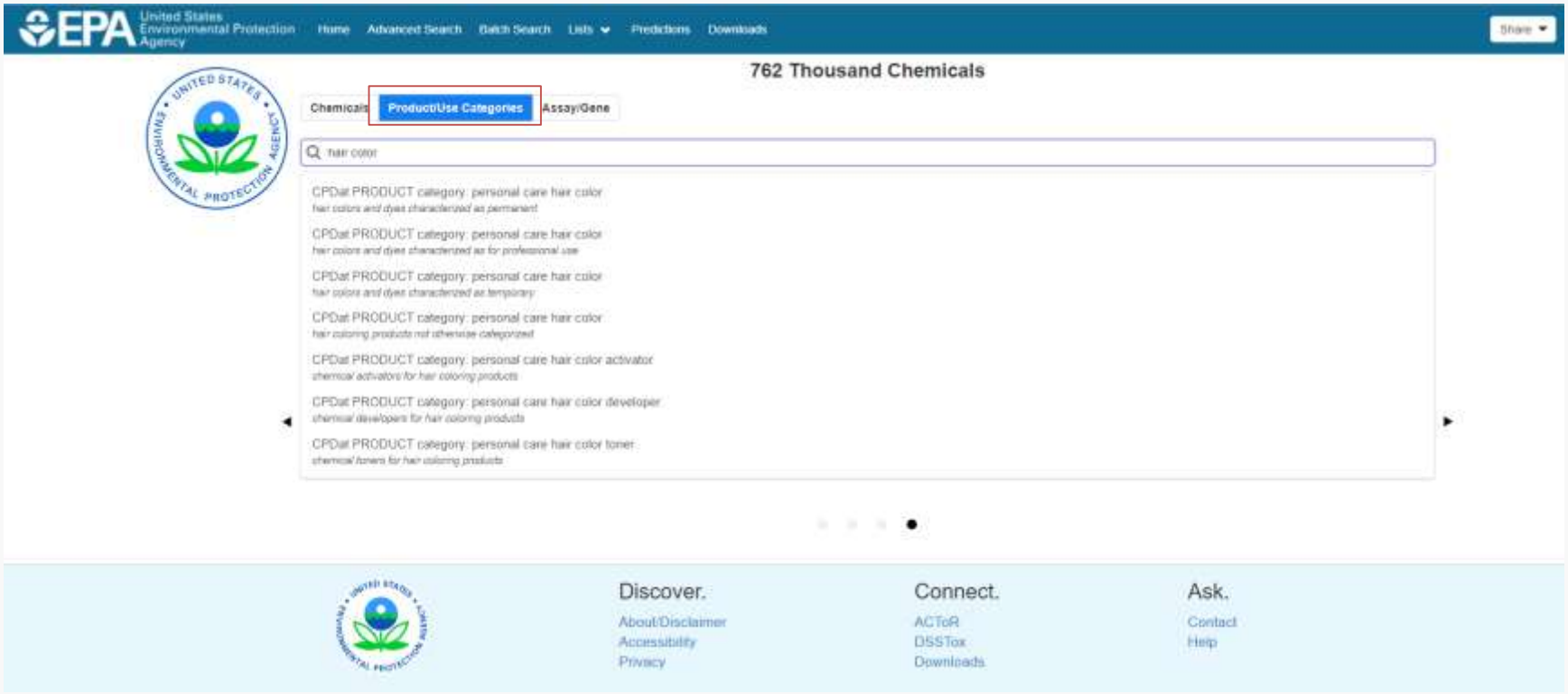
https://comptox.epa.gov/dashboard



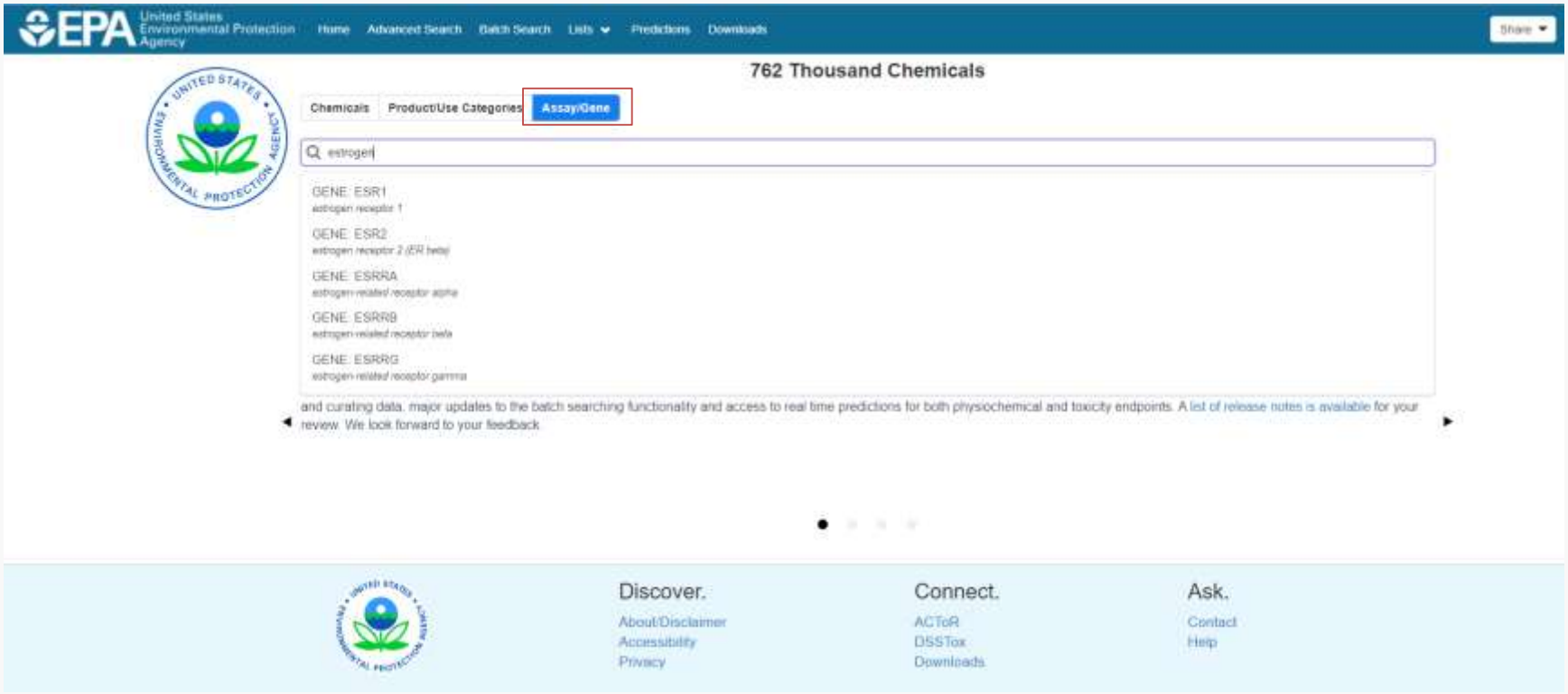
CompTox Dashboard Chemicals



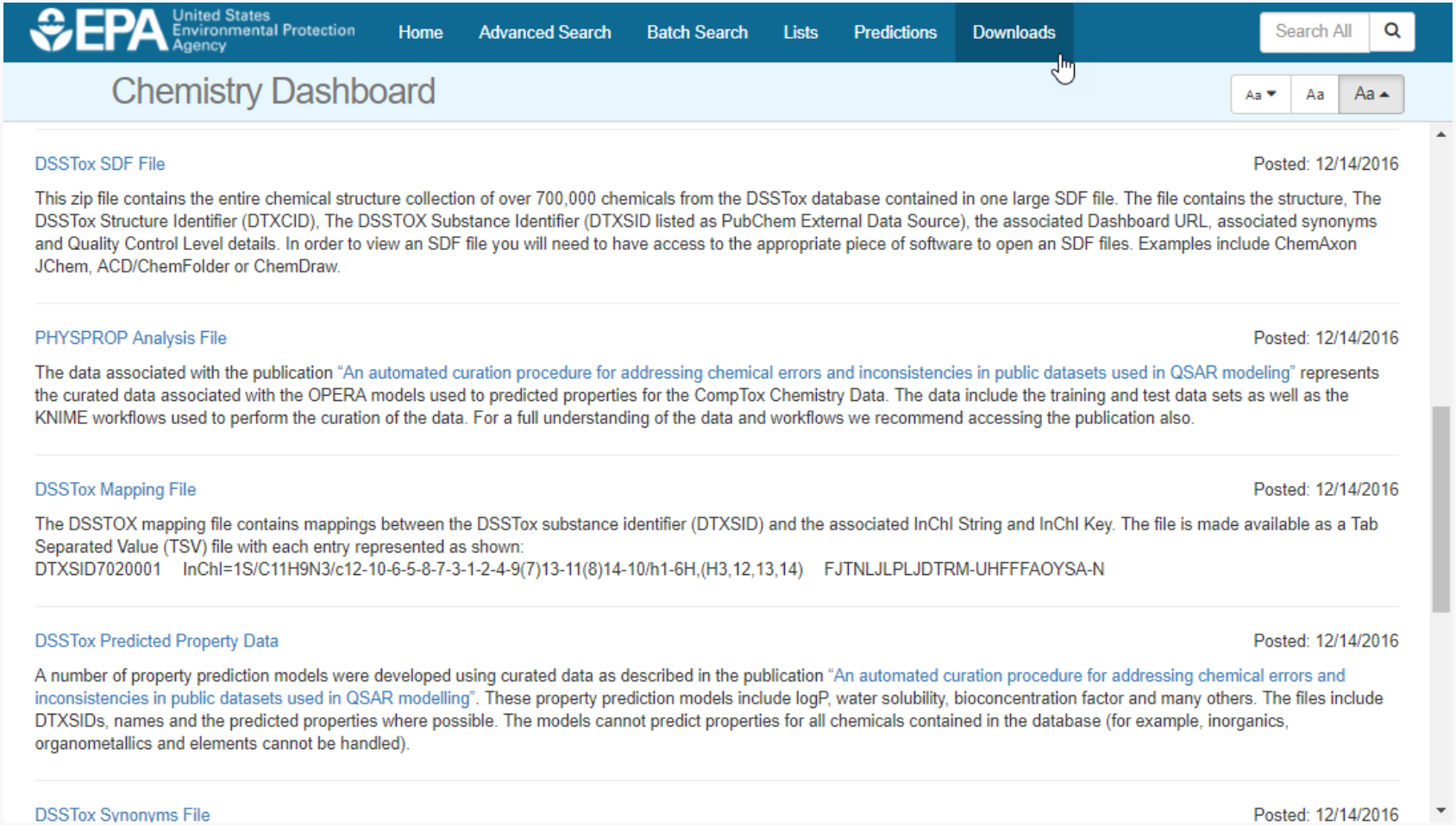
CompTox Dashboard
Products and Use Categories



CompTox Dashboard
Assays and Genes

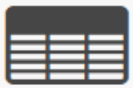


Downloadable data in useful formats

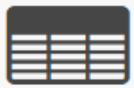


EPA FigShare Page
https://epa.figshare.com/

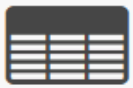
Datasets with versioning



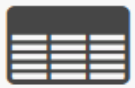
ReadMe for Animal Toxicity Study ToxRefDB files
EPA's National Center f... 29/03/2018



Collaborative Estrogen Receptor Activity Prediction Project (CERAP...
EPA's National Center f... 29/03/2018

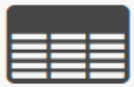


ToxCast Database (invitroDB)
EPA's National Center f... 29/03/2018



Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EP...
Antony Williams 12/08/2016

Chemistry Dashboard Data: Physprop Analysis Readme
EPA's National Center f... 29/03/2018



Chemistry Dashboard Data: Physprop Analysis
EPA's National Center f... 10/05/2018

Standard Laboratory Protocol for Tox21 Assays
EPA's National Center f... 29/03/2018

ToxCast & Tox21 high-throughput assay information
EPA's National Center f... 29/03/2018

ReadMe for ToxCast and Tox21 high-throughput assay information
EPA's National Center f... 29/03/2018

Collaborative Estrogen Receptor Activity Prediction Project (CERAP...
EPA's National Center f... 29/03/2018

Chemistry Dashboard Data: Abstract Sifter User Guide
EPA's National Center f... 29/03/2018

ToxCast Database (invitroDB) for Mac Users
EPA's National Center f... 22/05/2018

Measuring Our IMPACT

Database | Open Access

The CompTox Chemistry Dashboard: a community data resource for environmental chem

Antony J. Williams, Christopher M. Grulke, Jeff Edwards, Andrew D. McEachran
Nancy C. Baker, Grace Patlewicz, Imran Shah, John F. Wambaugh, Richard S. Judso

Journal of Cheminformatics 2017 9:61
https://doi.org/10.1186/s13321-017-0247-6 | © The Author(s) 2017
Received: 30 September 2017 | **Accepted:** 18 November 2017 | **Published:** 28 Nov

Publication metrics About

Dimensions Badge

8

8 Total citations
8 Recent citations
n/a Field Citation Ratio
n/a Relative Citation Ratio

Altmetric

24

Twitter (40)
Reddit (1)
Mendeley (20)
CiteULike (2)

Dimensions

Publication - Article

The CompTox Chemistry Dashboard: a community data resource for environmental chemistry

Journal of Cheminformatics, 9(1), 61, 2017
https://doi.org/10.1186/s13321-017-0247-6

Authors

Antony J. Williams - National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA
Christopher M. Grulke - National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA
Jeff Edwards - National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA
8 more

Abstract

Despite an abundance of online databases providing access to chemical data, there is increasing demand for high-quality, structure-curated, open data to meet the various needs of the environmental sciences and computational toxicology communities. The U.S. Environmental Protection Agency's (EPA) web-based CompTox Chemistry Dashboard is addressing these needs by integrating diverse types of relevant domain data through a cheminformatics layer, built upon a database of curated substances linked to chemical structures. These data include physicochemical, environmental fate and transport,

more

The impact of our paper versus the impact of our data…187 downloads

EPA United States Environmental Protection Agency

Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EPA CompTox Dashboard

Dataset posted on 12.08.2016, 18:38 by Antony Williams

797 views | 187 downloads | 2 citations

10

The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physical chemical properties. This information is used by scientists to predict the potential health risks of chemicals.

The iCSS CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. The dashboard provides access to a variety of information on over 700,000 chemicals currently in use.

Within the dashboard, users can access chemical structures, experimental and predicted physicochemical and toxicity data, and additional links to relevant websites and applications. It maps curated physicochemical property data associated with chemical substances to their corresponding chemical structures.

This data are compiled from sources including the EPA's computational toxicology research databases, and public domain databases such as the National Center for Biotechnology Information's PubChem database.

This dataset is a mapping file between the dashboard chemicals and the associated InChIStrings and InChIKeys.

This file is the version produced by an export of the database on July 1st 2016

REFERENCES

- https://www.epa.gov/chemical-research/icss-chemistry-dashboard

CATEGORIES

- Cheminformatics
- Computational Chemistry
- Environmental Chemistry

KEYWORD(S)

EPA CompTox Dashboard | DTXSID | InChIKeys | Environmental Chemistry | EPA

LICENCE

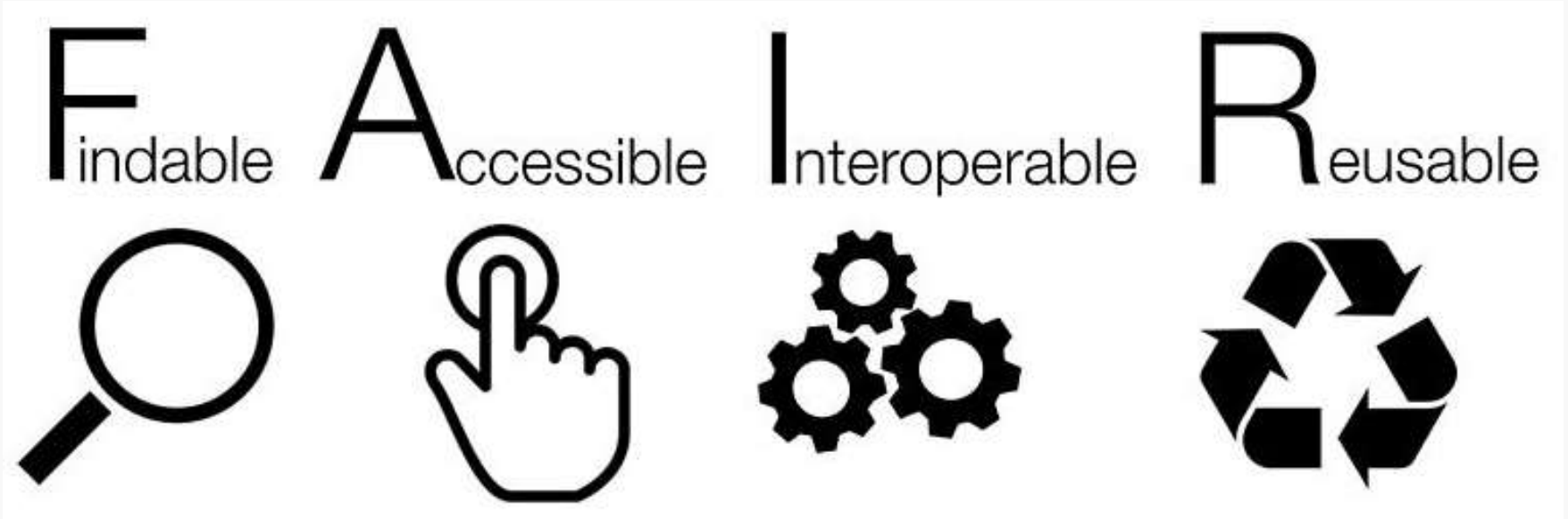
CC0

EXPORT

RefWorks

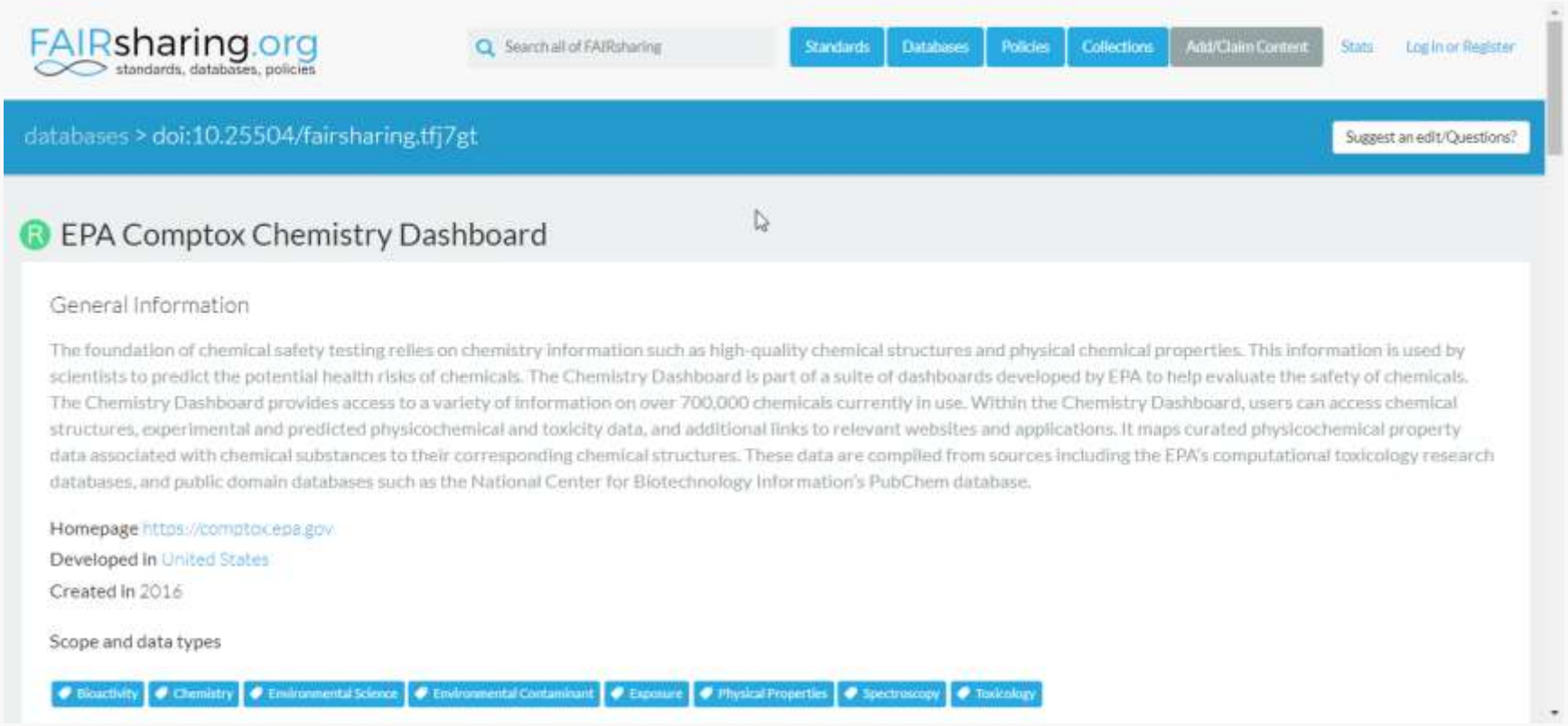
BibTeX

FAIR Data



FAIRsharing.org

EPA United States Environmental Protection Agency



Examples of our transparency
1. OPERA Prediction Models

EPA United States Environmental Protection Agency

Springer Link

Journal of Cheminformatics
December 2018, 10:10 | Cite as

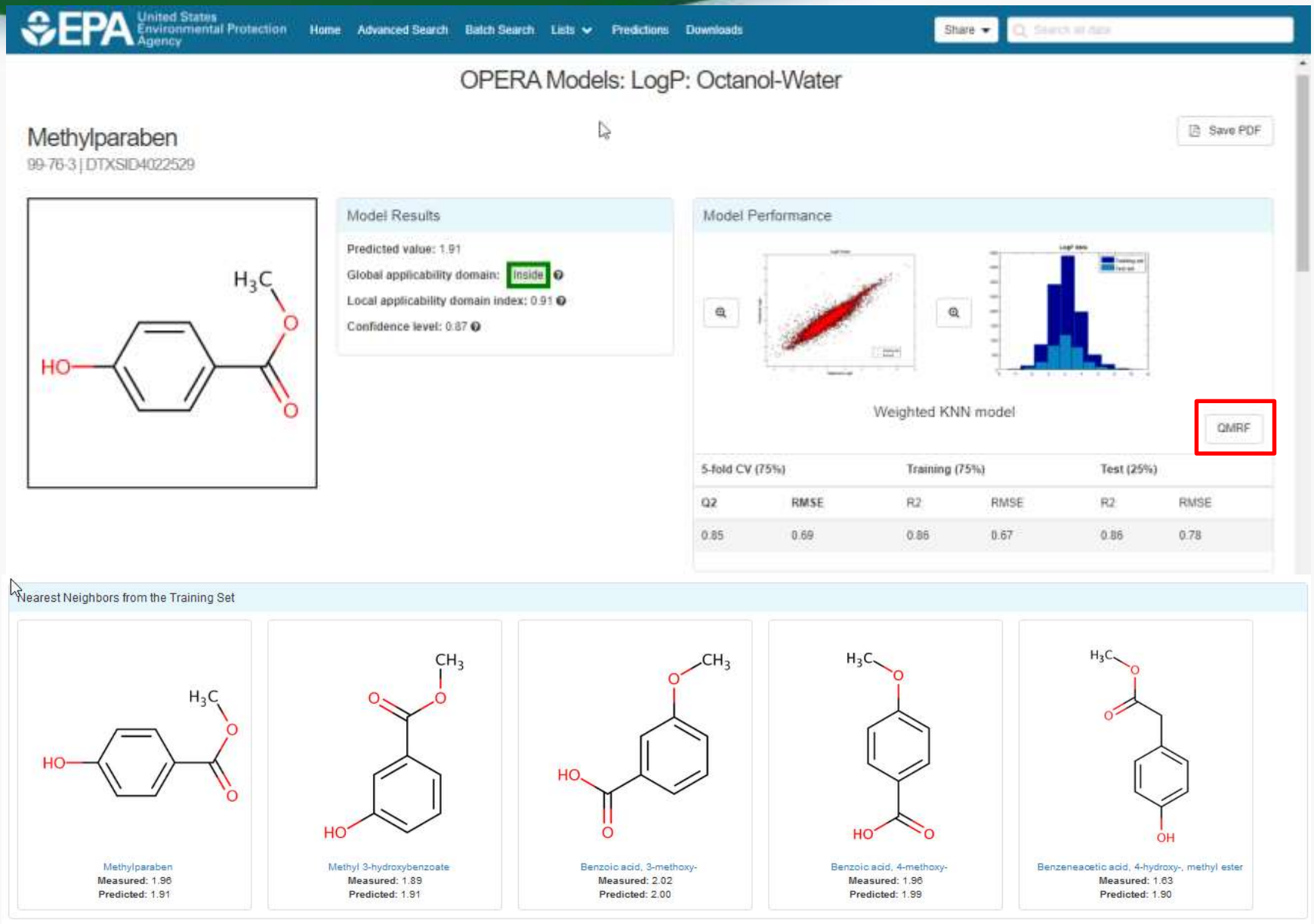
OPERA models for predicting physicochemical properties and environmental fate endpoints

Authors | Authors and affiliations

Kamel Mansouri, Chris M. Grulke, Richard S. Judson, Antony J. Williams

Open Access | Research article
First Online: 08 March 2018

What are OPERA Models?



What are OPERA Models?
Detailed QMRF reports

	***QMRF identifier (JRC Inventory):*** **Q17-16-0016**
	QMRF Title: OPERA-model for Octanol-water partition coefficient
	Printing Date: **Oct 17, 2017**

1.QSAR identifier

1.1.QSAR identifier (title):

OPERA-model for Octanol-water partition coefficient

1.2.Other related models:

No related models

1.3.Software coding the model:

OPERA V1.5

OPERA (OPEn (quantitative) structure-activity Relationship Application) is a standalone free and open source command line application. It provides a suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals based on PaDEL descriptors. It is available for download in Matlab, C and C++ languages from github under MIT license.

How Transparent in the Publication?

Supplementary material

13321_2018_263_MOESM1_ESM.zip (14.6 mb)

Additional file 1: S1. Training and test sets of the models with the corresponding JRC validated QMRFs.

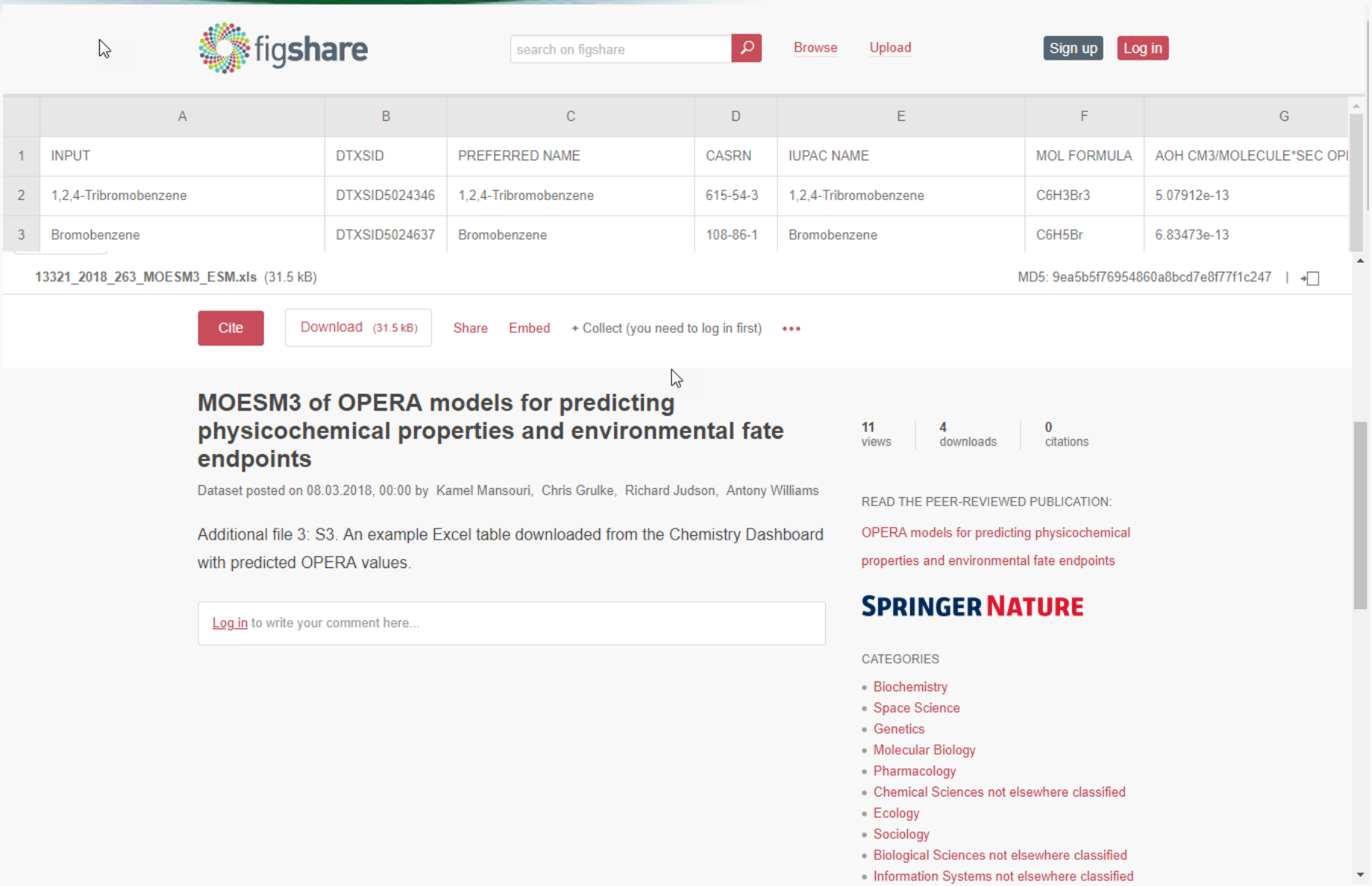
13321_2018_263_MOESM2_ESM.txt (4 kb)

Additional file 2: S2. OPERA command line help file.

13321_2018_263_MOESM3_ESM.xls (32 kb)

Additional file 3: S3. An example Excel table downloaded from the Chemistry Dashboard with predicted OPERA values.

On FigShare



figshare

search on figshare

Browse Upload Sign up Log in

	A	B	C	D	E	F	G
1	INPUT	DTXSID	PREFERRED NAME	CASRN	IUPAC NAME	MOL FORMULA	AOH CM3/MOLECULE*SEC OPI
2	1,2,4-Tribromobenzene	DTXSID5024346	1,2,4-Tribromobenzene	615-54-3	1,2,4-Tribromobenzene	C6H3Br3	5.07912e-13
3	Bromobenzene	DTXSID5024637	Bromobenzene	108-86-1	Bromobenzene	C6H5Br	6.83473e-13

13321_2018_263_MOESM3_ESM.xls (31.5 kB) MD5: 9ea5b5f76954860a8bcd7e8f77f1c247 |

Cite Download (31.5 kB) Share Embed + Collect (you need to log in first) •••

MOESM3 of OPERA models for predicting physicochemical properties and environmental fate endpoints

Dataset posted on 08.03.2018, 00:00 by Kamel Mansouri, Chris Grulke, Richard Judson, Antony Williams

Additional file 3: S3. An example Excel table downloaded from the Chemistry Dashboard with predicted OPERA values.

Log in to write your comment here...

11 views | 4 downloads | 0 citations

READ THE PEER-REVIEWED PUBLICATION:

OPERA models for predicting physicochemical properties and environmental fate endpoints

SPRINGER NATURE

CATEGORIES

- Biochemistry
- Space Science
- Genetics
- Molecular Biology
- Pharmacology
- Chemical Sciences not elsewhere classified
- Ecology
- Sociology
- Biological Sciences not elsewhere classified
- Information Systems not elsewhere classified

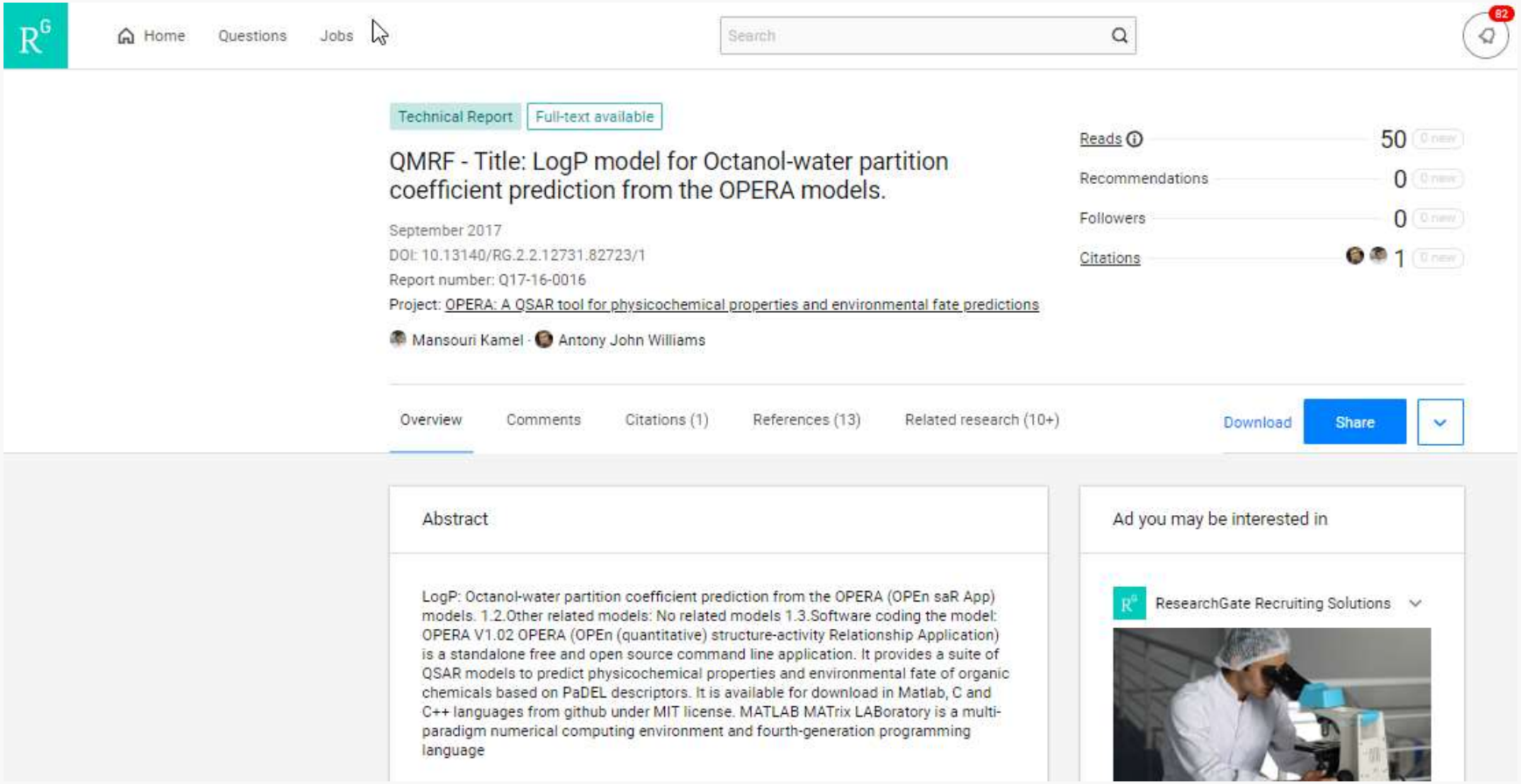
How Transparent with the Code?

https://github.com/kmansouri/OPERA

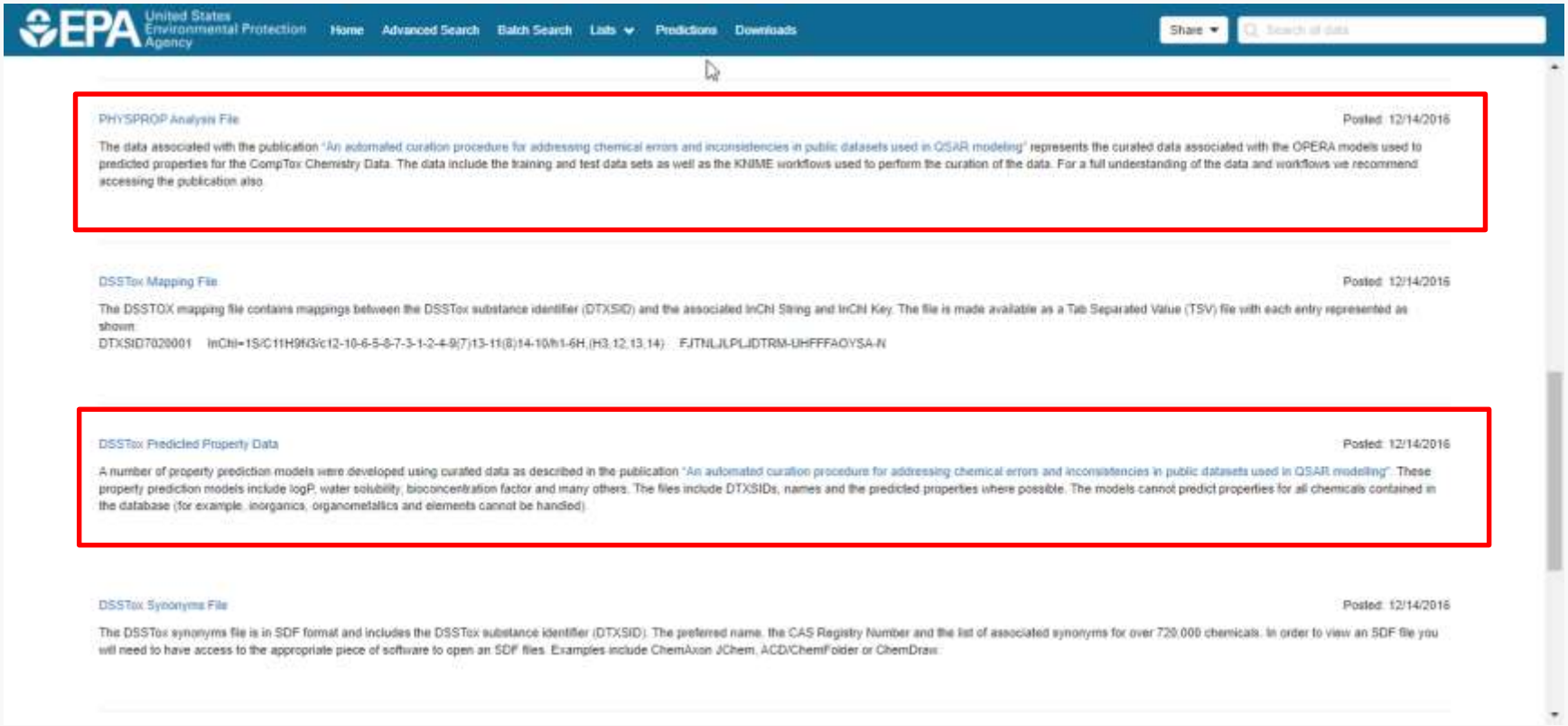
EPA United States Environmental Protection Agency

kmansouri Update README.md		Latest commit 4c7313b 3 days ago
Icon.png	OPERA 1.2 icon	a year ago
LICENSE	Initial commit	2 years ago
Logo.png	Added logo and icon	2 years ago
Matlab_Source_code.zip	OPERA 1.5 MATLAB source code	8 months ago
OPERA_CLi_Linux.tar.gz	OPERA 1.5 Linux	8 months ago
OPERA_CPP_library.tar.gz	OPERA 1.5 C++ Library	8 months ago
OPERA_C_library.tar.gz	OPERA 1.5 C Library	8 months ago
OPERA_Data_SDF.zip	OPERA 1.5 Datasets	10 months ago
OPERA_py.zip	OPERA 1.5 Python Library	8 months ago
OPERA_win.zip	OPERA 1.5 Windows	8 months ago
PaDEL-Descriptor.zip	PaDEL-Descriptors	a year ago
PaDEL_Descriptors.tar.gz	Added code and libraries	2 years ago
QMRFs.zip	QMRF reports of the QSAR models	10 months ago
README.md	Update README.md	3 days ago
icons.zip	OPERA 1.2 icons different sizes	a year ago

How widely do we share?



How widely do we share?



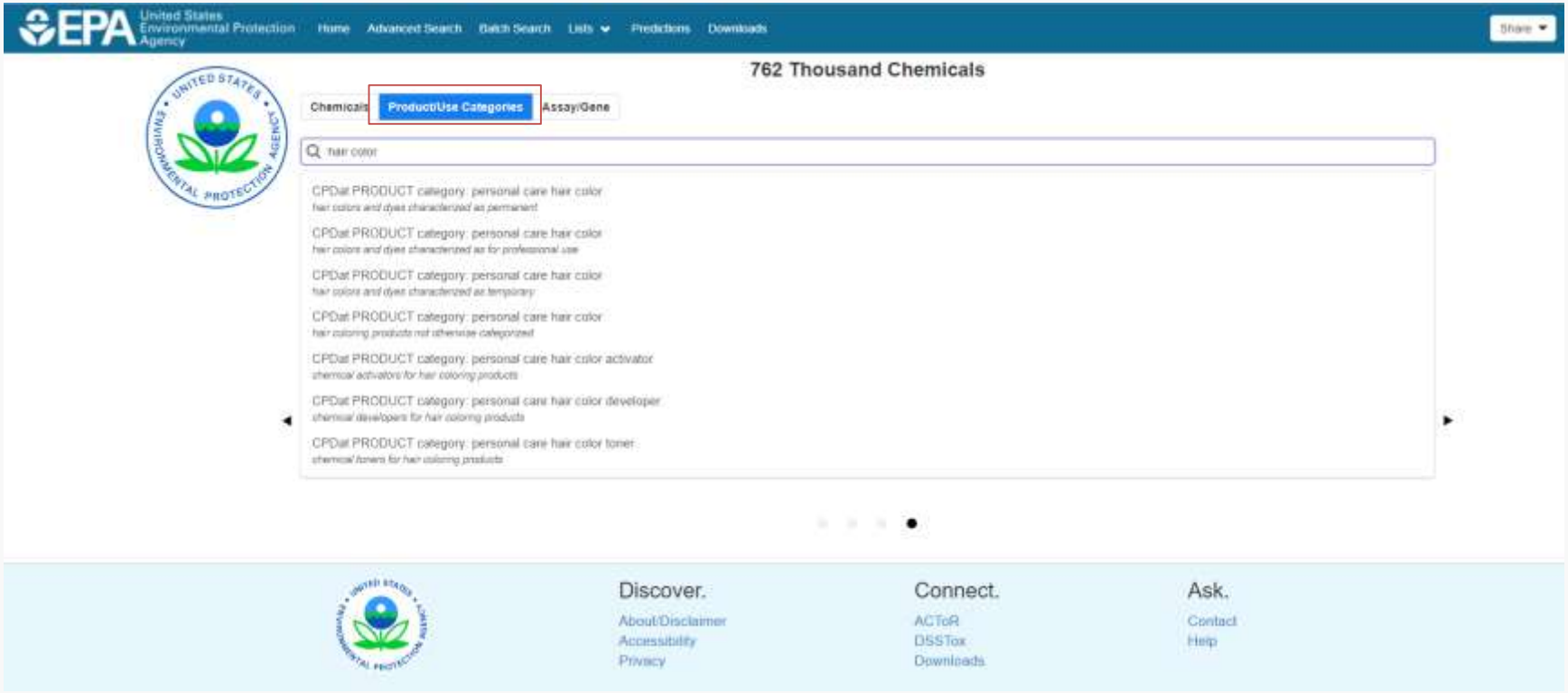
Journals for Data Articles

EPA United States Environmental Protection Agency

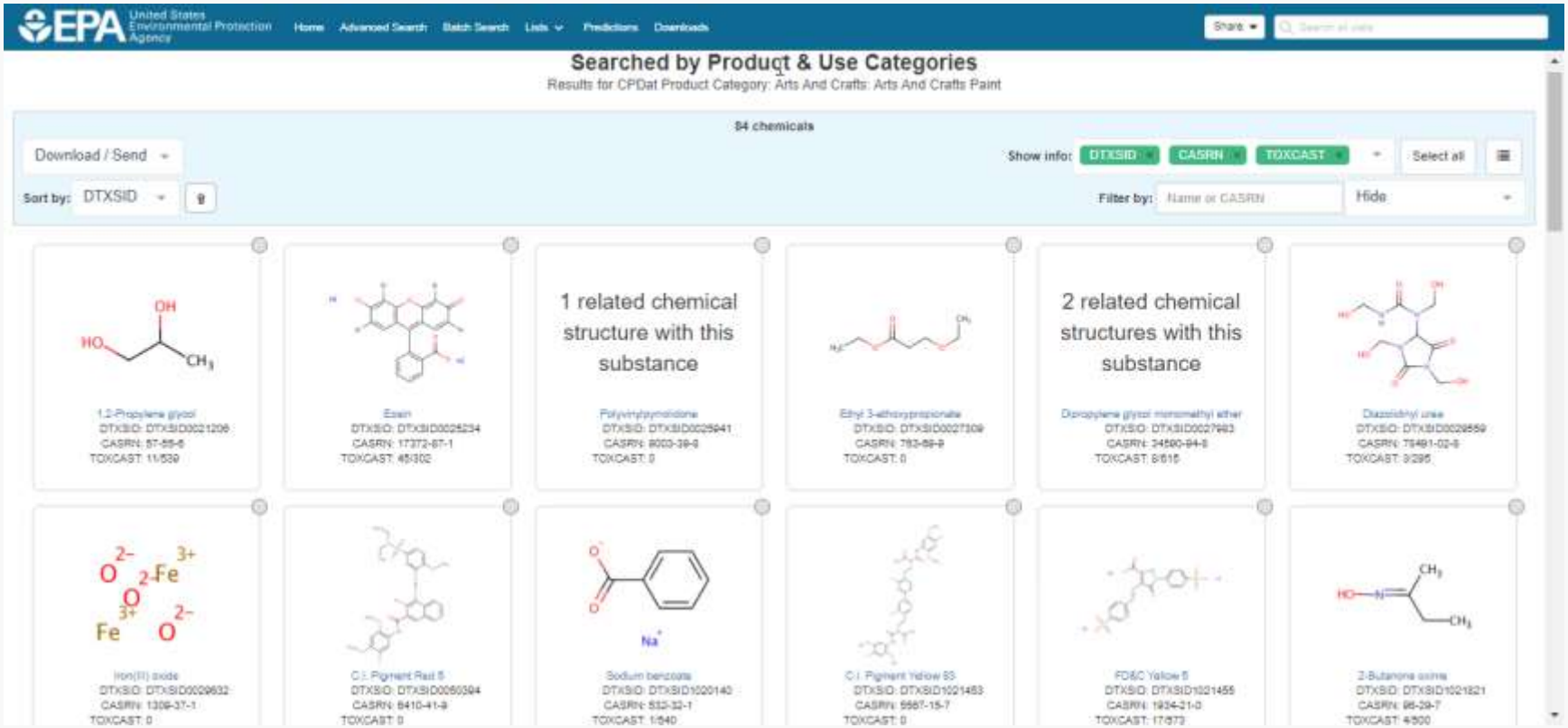
Data Journals Hold Promise

Examples of our transparency
2. The Chemical and Products Database

Examples of our transparency
2. The Chemical and Products Database



What chemicals are in Arts and Crafts Paint?



Examples of our transparency
2. The Chemical and Products Database

EPA United States Environmental Protection Agency

Data Citations

1. Williams, A. *Figshare*
http://dx.doi.org/10.23645/epacomptox.5352997 (2017)

Cite | Download (21.16 MB) | Share | Embed | + Collect (you need to log in first) | •••

The Chemical and Products Database (CPDat) MySQL Data File

349 views | 39 downloads | 0 citations

Version 2 | Fileset posted on 13.10.2017, 14:07 by Antony Williams

Quantitative data on product chemical composition is a necessary parameter for characterizing near-field exposure. This data set comprises reported and predicted information on >75,000 chemicals contained in >15,000 consumer products. The data's primary intended use is for exposure, risk, and safety assessments. The data set includes specific products with quantitative or qualitative ingredient information, which has been publicly disclosed through material safety data sheets (MSDS) and ingredient lists. A single product category from a refined and harmonized set of categories has been assigned to each product. The data set also contains information on the functional role of chemicals in products, which can inform predictions of the concentrations in which they occur. These data will be useful to exposure and risk assessors evaluating chemical and product safety.

The data set presented here is in the form of a MySQL relational database, which mimics CPDat data available under the 'Exposure' tab of the CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard) as of August 2017.

Log in to write your comment here...

EPA United States Environmental Protection Agency

CATEGORIES
- Toxicology

KEYWORD(S)
Computational Toxicology | CPDat | Chemical and Products Database | MySQL

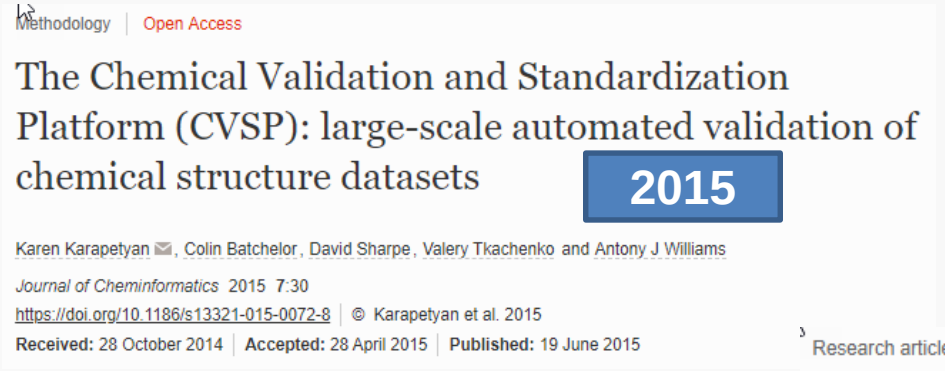
LICENCE
CC0

Are we oversharing?

- We share our datasets
- We share our models
- We share our code
- We share our database schemas
- We share our database dumps
- We are not yet sharing all code under an application like the CompTox Dashboard..

A Recurring Plea for Formats

- Chemical data exchange formats are critical
- We all know it's imperfect, that there are efforts afoot, and it's taking time



Methodology | Open Access

The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets

2015

Karen Karapetyan, Colin Batchelor, David Sharpe, Valery Tkachenko and Antony J Williams

Journal of Cheminformatics 2015 **7**:30

https://doi.org/10.1186/s13321-015-0072-8 | © Karapetyan et al. 2015

Received: 28 October 2014 | **Accepted:** 28 April 2015 | **Published:** 19 June 2015

Research article | Open Access

PubChem chemical structure standardization

Volker D. Hähnke, Sunghwan Kim and Evan E. Bolton

2018

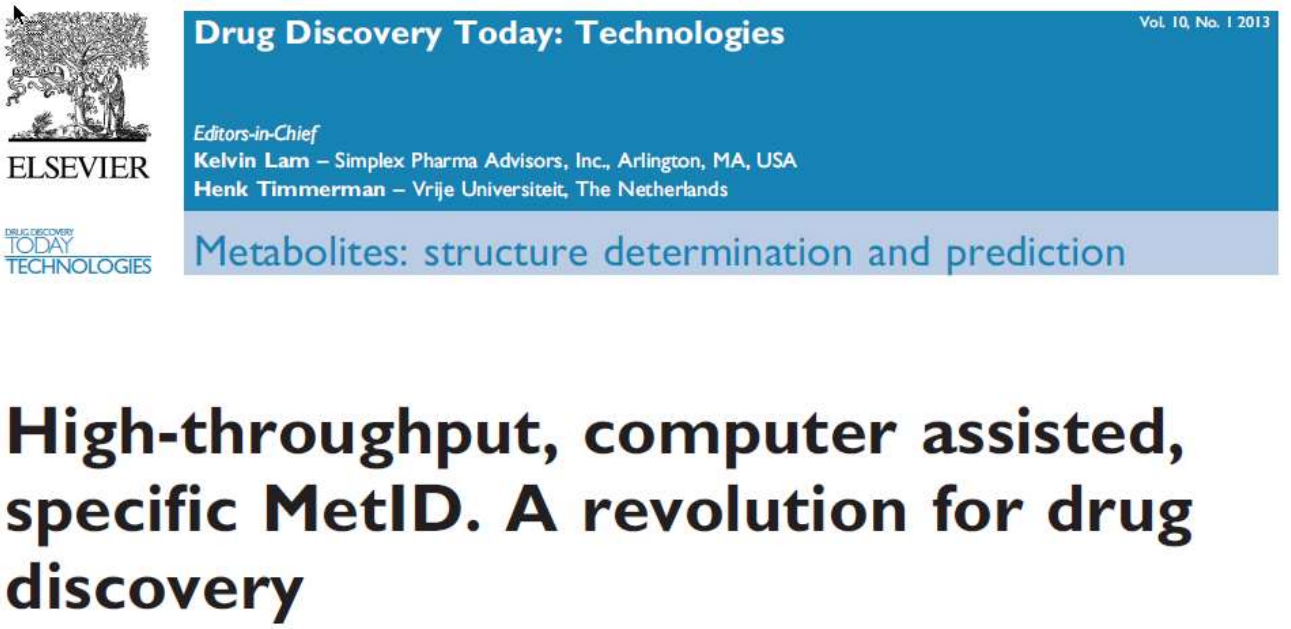
Journal of Cheminformatics 2018 **10**:36

https://doi.org/10.1186/s13321-018-0293-8 | © The Author(s) 2018

Received: 18 April 2018 | **Accepted:** 1 August 2018 | **Published:** 10 August 2018

Use APPROPRIATE formats

- Our databases use chemical structures
- Many journal articles deliver poor data



Chemical structures as text

EPA United States Environmental Protection Agency

Table 1. Metabolites found in the different incubations tested

Name	RT	m/z	Formula	m/z Diff (ppm)	Mass score	SMILES
Parent	3.13	455.2926	C27H38N2O4	−3.48		N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2
M6 −164	2.26	291.2077	C17H26N2O2	−1.37	429	N(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1
M16 −14	3.06	441.2743	C26H36N2O4	2.41	534	c1cc(CCNCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
M14 +16	2.92	471.2866	C27H38N2O5	−1.59	476	N#CC(CCCN(C)CC(O)c1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2
M9 −14	2.78	441.2761	C26H36N2O4	−1.7	590	C(#N)C(CCCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(O)c(OC)c2
M11 −14	2.84	441.2742	C26H36N2O4	2.57	473	Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC
M12 +2	2.87	457.2707	C26H36N2O5	−0.93	570	O(C)c1cc(ccc1OC)C(C#N)(CCCNCC(O)c2ccc(OC)c(OC)c2)C(C)C
M5 −178	2.2	277.1894	C16H24N2O2	7.84	419	C(C)(C)C(C#N)(CCCN)c1ccc(OC)c(OC)c1
M8 +2	2.67	457.2708	C26H36N2O5	−1.18	581	OC(CN(C)CCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1)c2ccc(O)c(OC)c2
M15 −14	2.92	441.2743	C26H36N2O4	2.23	614	Oc1ccc(CCN(C)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC
M2 −259	0.73	196.1326	C11H17NO2	5.91	618	c1(CCNC)ccc(OC)c(c1)OC
M10 −28	2.8	427.2617	C25H34N2O4	−4.79	487	c1(OC)cc(ccc1OC)C(C#N)(CCCNCCc2ccc(O)c(OC)c2)C(C)C
M7 +2	2.46	457.2717	C26H36N2O5	−3.23	492	c1cc(CCN(O)CCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
M17 +16	3.21	471.2853	C27H38N2O5	1.19	534	N#CC(CCCN(C)CCc1ccc(OC)c(OC)c1)(c2ccc(OC)c(OC)c2)C(C)(C)O
M4 −178	1.86	277.1927	C16H24N2O2	−3.8	444	COc1cc(ccc1O)C(C#N)(CCCNC)C(C)C
M1 −289	0.44	166.0858	C9H11NO2	5.93	136	c1(CCNC)ccc(=O)c(c1)=O
M13 −16	2.88	439.2603	C26H34N2O4	−1.43	367	c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC

How do I extract structures?

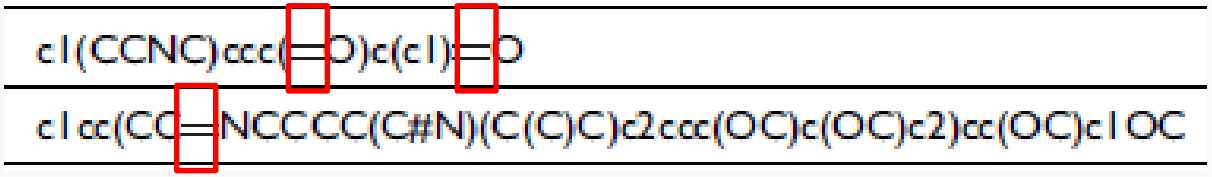
- Copy and paste into Excel as a start point
- Assume no loss of formatting!
- Convert SMILES to structures
- But Copy-Paste doesn't work

c1(CCNC)ccc(=O)c(c1)=O
c1cc(CC=NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC

```
c1(CCNC)ccc( O)c(c1) O
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

How do I extract structures?

- Copy and paste into Excel as a start point
- Assume no loss of formatting!
- Convert SMILES to structures
- Copy-Paste doesn't work

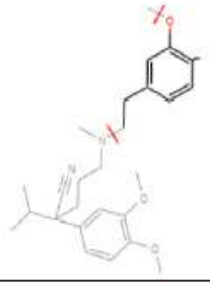
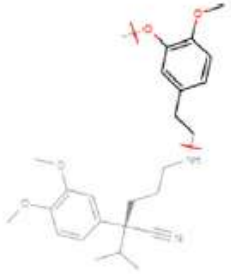
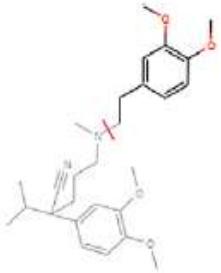
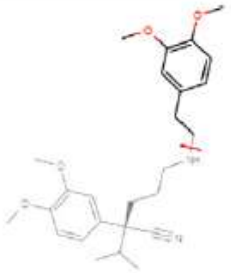
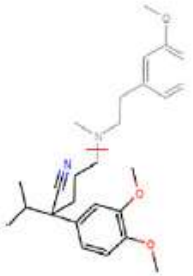
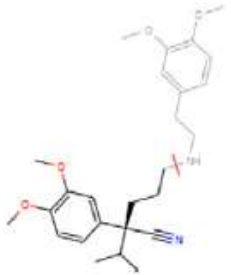



```
c1(CCNC)ccc( O)c(c1) O
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC
```

Structure Drawings Are Worse

EPA United States Environmental Protection Agency

Table 2. Selection of fragments that help in the M16-16 metabolite structure elucidation

Sub. obs. m/z	Sub. cal. m/z	Sub. m/z diff. ppm	Substrate	Metabolite	Δ	Met. obs. m/z	Met. calc. m/z	Met. m/z diff. ppm
150.0664	150.0681	11.42			+0	150.0670	150.0681	7.25
165.0869	165.0916	28.22			+0	165.0892	165.0916	14.30
260.1637	260.1651	5.33			+0	260.1652	260.1651	−0.50

Names and CASRNs are **NOT** structures

- In our domain most chemicals are text – chemical names and CAS Numbers

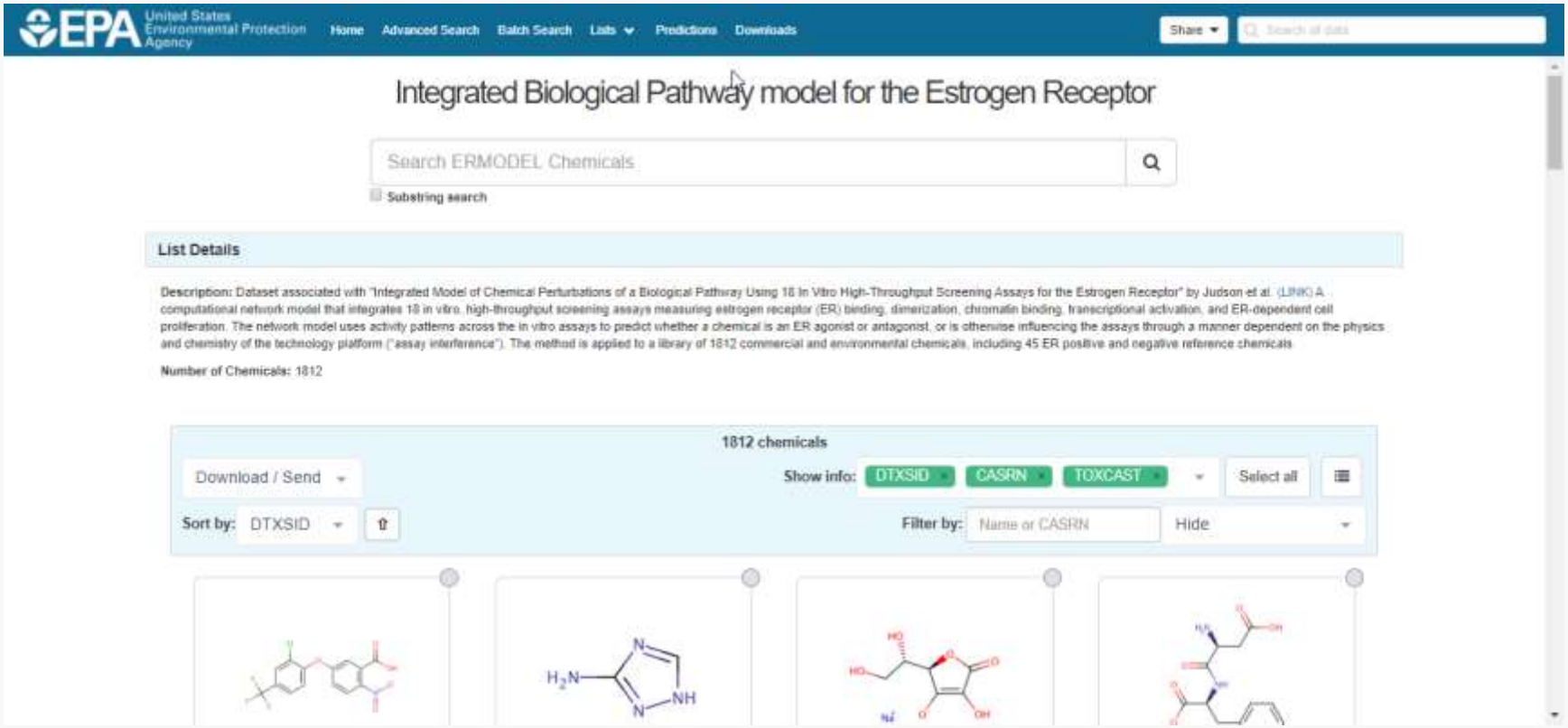
Attachment D (Method 3)
SIM quantitation ions and qualifiers for internal standards, references method analysis, and surrogates

Name of Compound	CAS No.	Quantitation Ion	Qualifier Ions
Phenol-d6 (SS)	13187-88-3	99	71, 42
Phenol	108-95-2	94	66
1,4-Dichlorobenzene	106-46-0	146	111. 75, 50
Acetophenone	98-86-2	105	77, 51, 120
Acenaphthene-d10 (IS)	15067-26-2	162	160, 80
p-Cresol	106-44-5	107	108, 77
Isophorone	78-59-1	82	138, 54
Camphor	76-22-2	95	81, 108, 152
Isoborneol	124-76-5	95	110, 121, 136
Menthol	89, 78, 1	71	81, 123, 138
Naphthalene	91-20-3	128	102, 51
Methyl salicilate	119-36-8	120	92, 152, 65

And generally problematic…

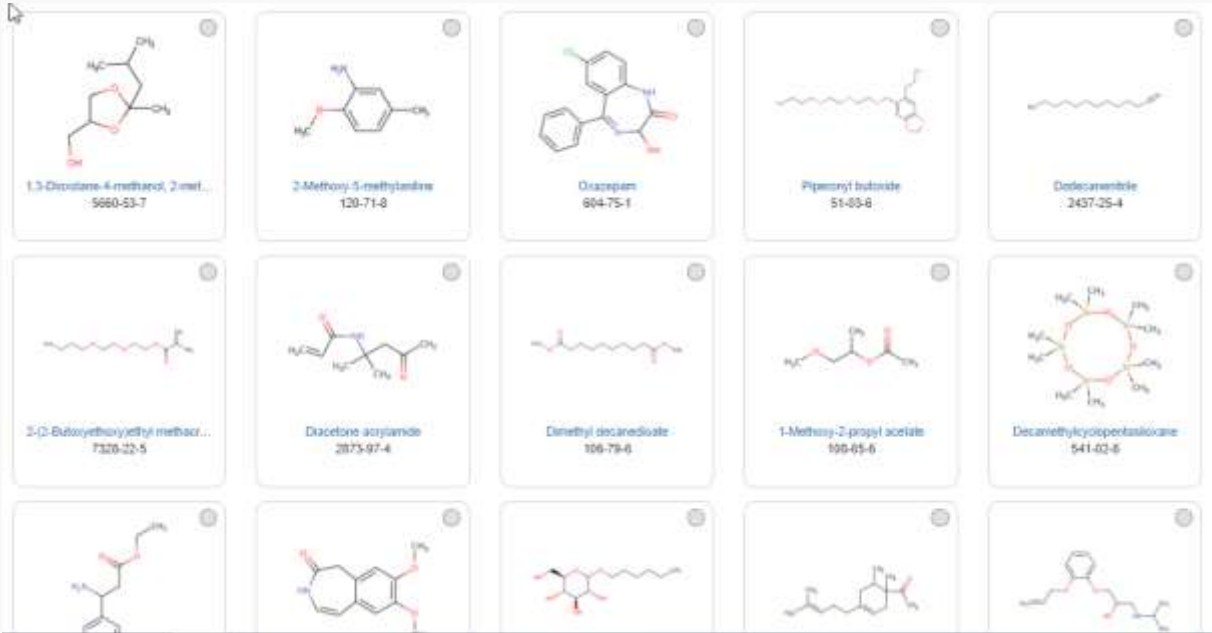
Name of Compound	CAS No.
Phenol-d6 (SS)	13187-88-3
Phenol	108-95-2
1,4-Dichlorobenzene	106-46-0
Acetophenone	98-86-2
Acenaphthene-d10 (IS)	15067-26-2
p-Cresol	106-44-5
Isophorone	78-59-1
Camphor	76-22-2
Isoborneol	124-76-5
Menthol	89, 78, 1
Naphthalene	91-20-3
Methyl salicilate	119-36-8

So this is how we publish our chemical substance data

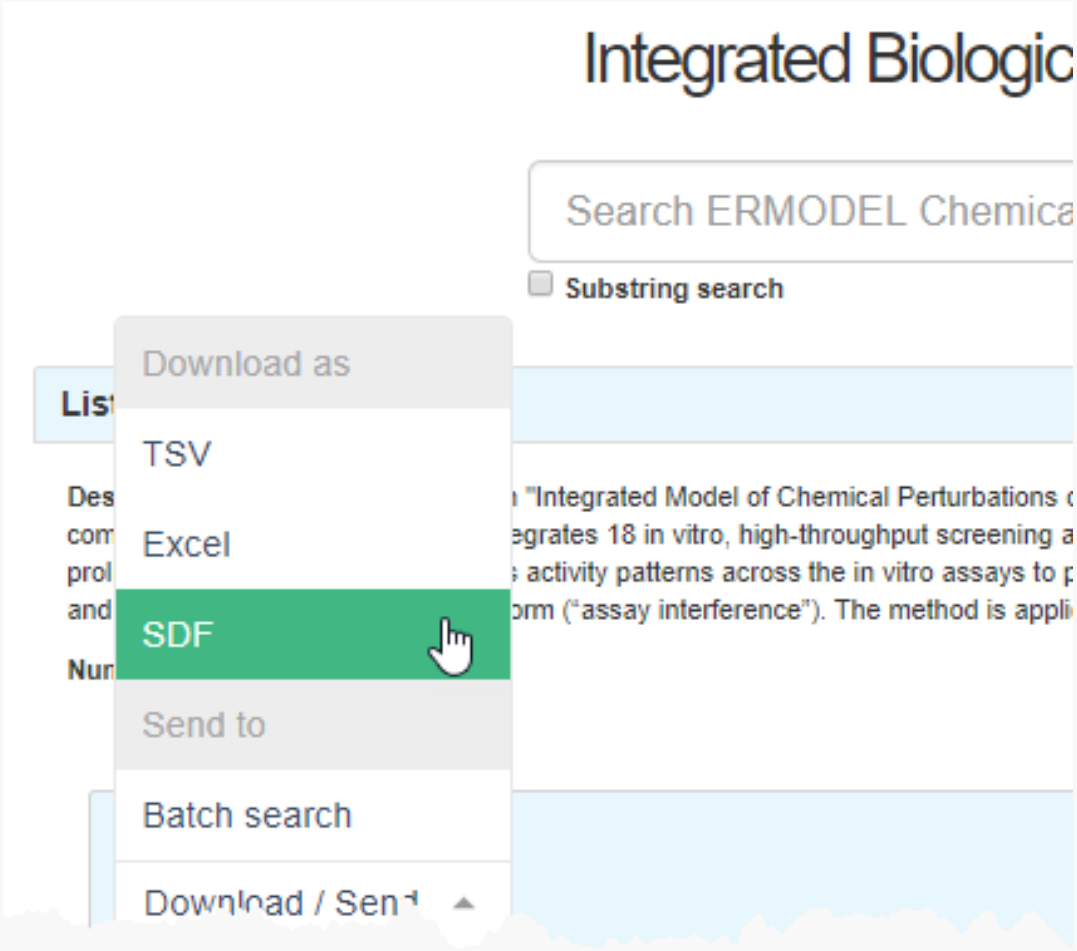


Publish data **for your domain**

- Push data out in immediately useful formats
  - Make multiple formats available as appropriate – SQL database dumps, Excel files, etc.
  - For chemistry - SDF files – includes LAYOUT and data (but requires cheminformatics tools), Excel files generally handled at any desktop.

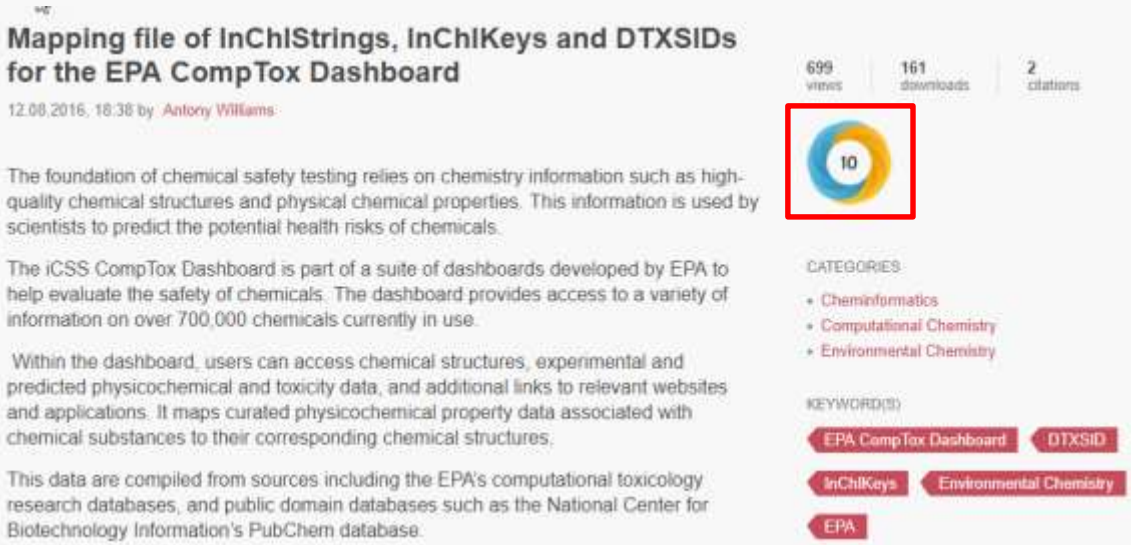


So this is how we publish our chemical substance data



When we publish now…

- Add the data as a “list” to our Lists of Chemicals
- Generally store files on our FTP site PLUS copies in a repository (or two)
- Multiple formats of data as appropriate
  - Can be as supplementary data or DOI’ed data files
- DOI’ed data gives altmetrics also..



Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EPA CompTox Dashboard

12.08.2016, 18:38 by Antony Williams

The foundation of chemical safety testing relies on chemistry information such as high-quality chemical structures and physical chemical properties. This information is used by scientists to predict the potential health risks of chemicals.

The iCSS CompTox Dashboard is part of a suite of dashboards developed by EPA to help evaluate the safety of chemicals. The dashboard provides access to a variety of information on over 700,000 chemicals currently in use.

Within the dashboard, users can access chemical structures, experimental and predicted physicochemical and toxicity data, and additional links to relevant websites and applications. It maps curated physicochemical property data associated with chemical substances to their corresponding chemical structures.

This data are compiled from sources including the EPA's computational toxicology research databases, and public domain databases such as the National Center for Biotechnology Information's PubChem database.

699 views | 161 downloads | 2 citations

10

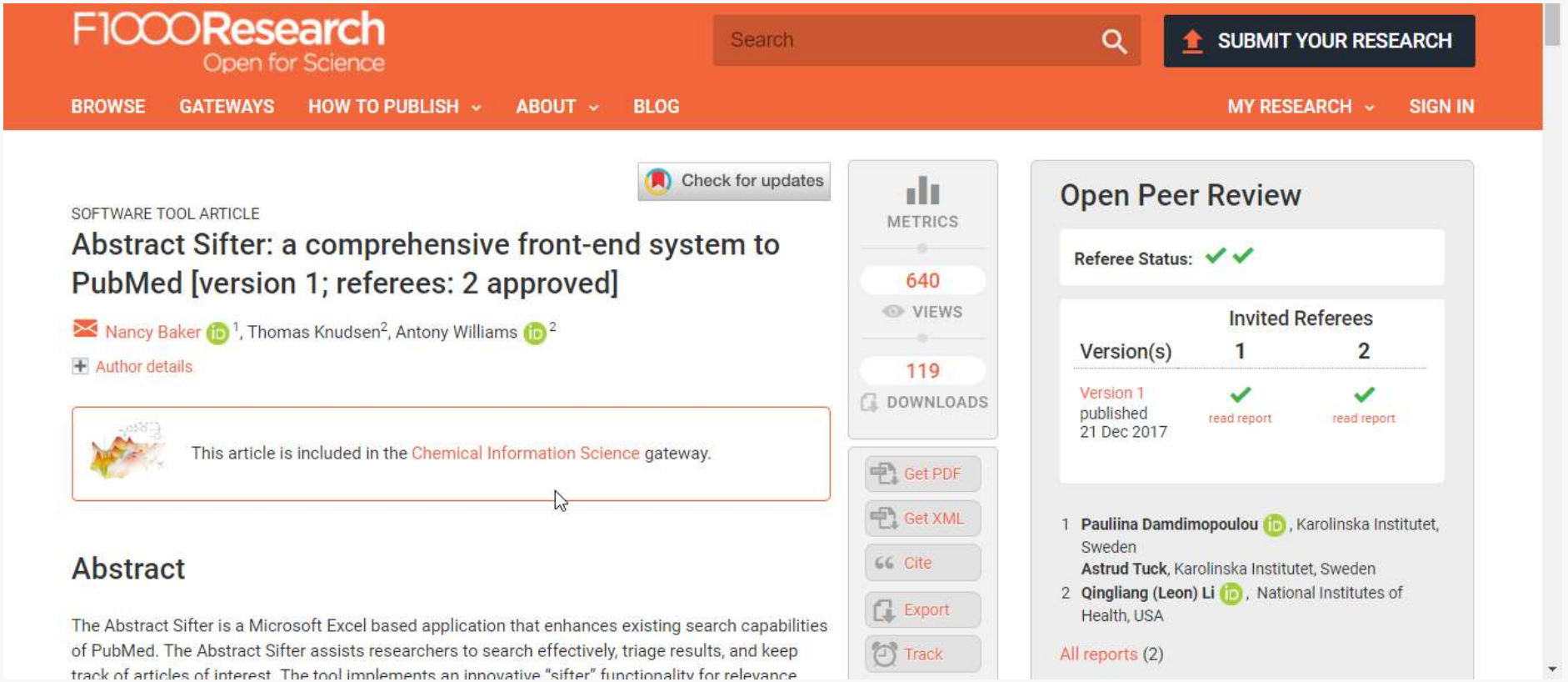
CATEGORIES

- Cheminformatics
- Computational Chemistry
- Environmental Chemistry

KEYWORD(S)

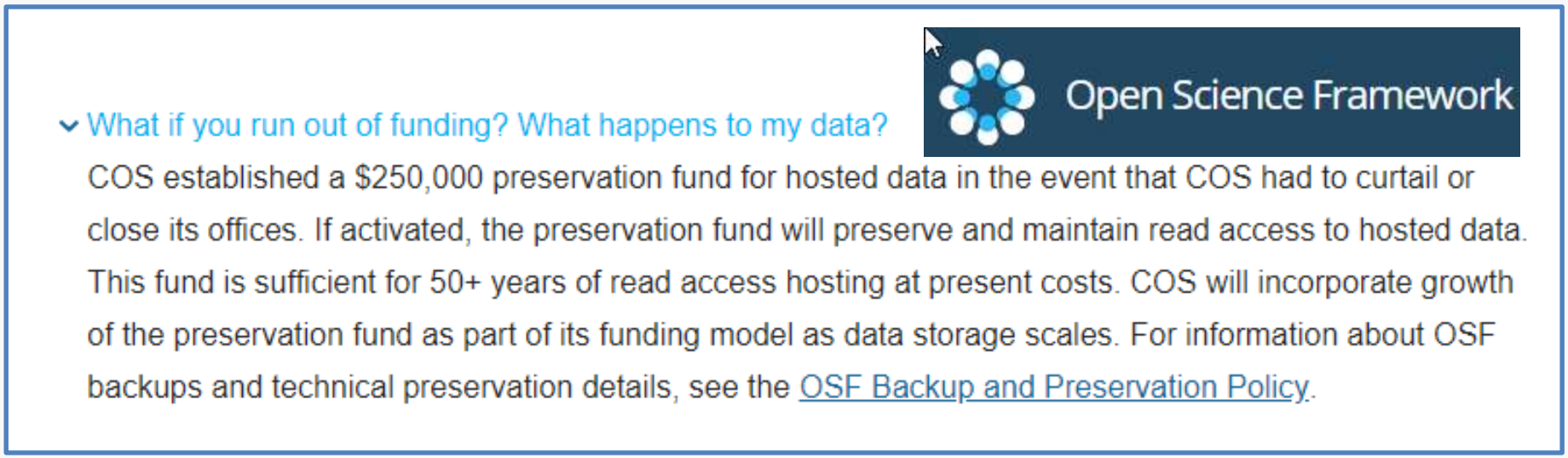
EPA CompTox Dashboard | DTXSID | InChIKeys | Environmental Chemistry | EPA

PERSONAL affection for **Open** Peer Review



Will my data always be available?

- There are no guarantees our data will always be on all sites. They come and go.
- Some organizations are just as concerned as you about your data…



Future Work

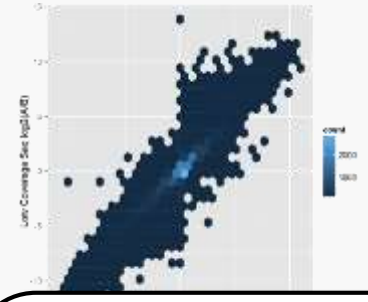
- We have a lot more models to make available – Open *and* free web services
- We are planning for embeddable widgets to access our Open data
- The challenge of versioning – ongoing curation of data requires version tracking

NCCT “InvitroDB_v3”

- The last public release of ToxCast data (invitroDB_v2) was in 3rd Quarter of 2015
- The next release invitroDB_v3 is Fall 2018
- Data includes new assays, new chemicals, new pipelining, results of data curation
- Data will also release via CompTox Dashboard
- Data will be available at https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data

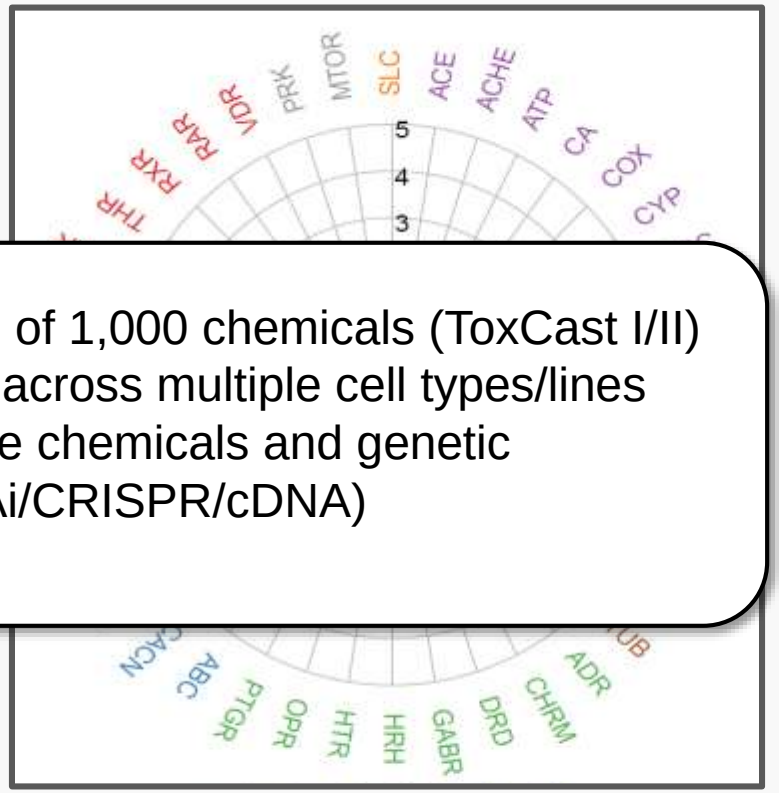
NCCT Transcriptomics Data will Deliver 50 Terabytes of Data for Analysis

EPA United States Environmental Protection Agency

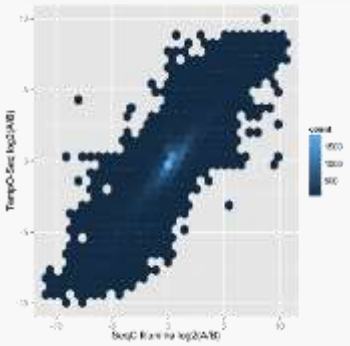


TruSeq
r^2 0.74

MOA Analysis Pipeline



- Large scale screen of 1,000 chemicals (ToxCast I/II) 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

Conclusions

- For publications and for applications deliver your data in “fit-for-purpose” form
- Multiple formats of data are appropriate
- Consider your audience(s) and “how would YOU want the data”???
- Not everything can be put in a supplementary file – use repositories and DOI your data
- Take the benefits of DOIs to measure altmetrics

Conclusion

- NCCT works with transparency in mind – our data is released for community usage – as data and in apps
- The CompTox Dashboard is the new application to surface all data as the architecture expands
- We attempt to deliver data in as transparent a form as possible for our scientific publications
- We consider long-term access and versioning in our releases. Lots of work to do…

…and let other people use it…

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