

# 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*

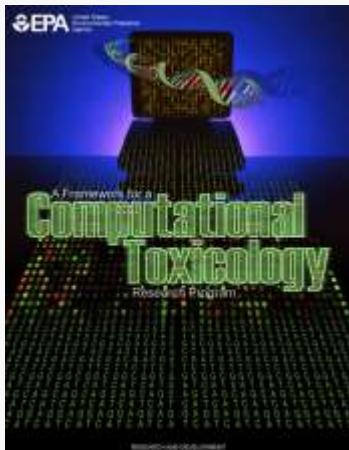
*The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA*

*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.

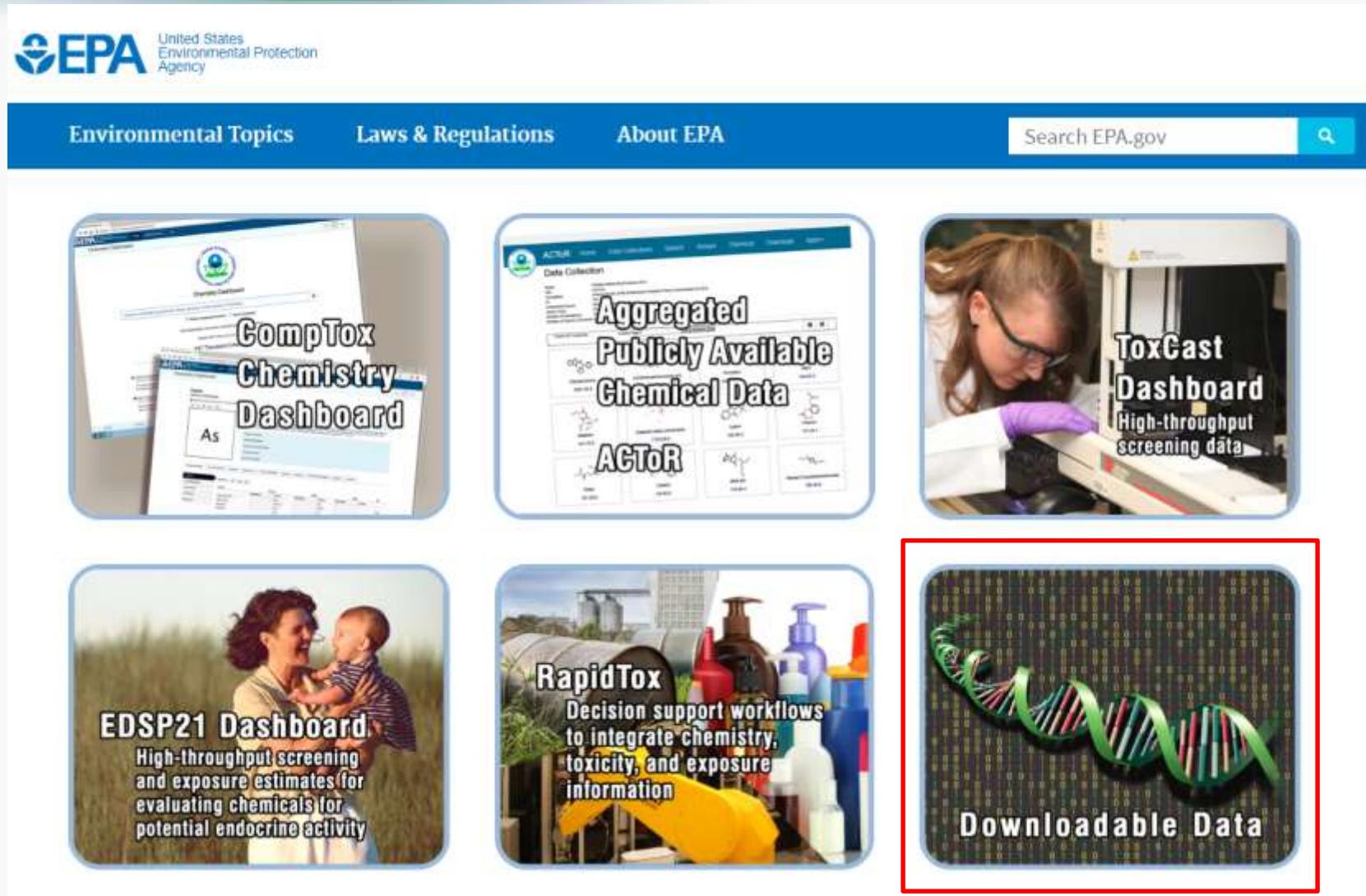
# National Center for Computational Toxicology



- 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/>



The screenshot displays the CompTox Portal homepage with a navigation bar at the top featuring the EPA logo, Environmental Topics, Laws & Regulations, About EPA, and a search bar. Below the navigation bar are six promotional boxes:

- Comptox Chemistry Dashboard**: A screenshot of a dashboard interface showing chemical structures and data for arsenic (As).
- Aggregated Publicly Available Chemical Data**: A screenshot of the ACToR Data Collection interface.
- ToxCast Dashboard**: An image of a scientist in a lab coat and gloves using a high-throughput screening machine, with text overlay: "ToxCast Dashboard High-throughput screening data".
- EDSP21 Dashboard**: An image of a woman holding a baby, with text overlay: "EDSP21 Dashboard High-throughput screening and exposure estimates for evaluating chemicals for potential endocrine activity".
- RapidTox**: An image of various personal care products like bottles and tubes, with text overlay: "RapidTox Decision support workflows to integrate chemistry, toxicity, and exposure information".
- Downloadable Data**: An image of a DNA double helix on a binary code background, with text overlay: "Downloadable Data". This box is highlighted with a red border.

# Downloadable CompTox Data

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



## 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

## 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 research 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 (GIDS) 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	<a href="#">ToxCast Chemicals, Data Management and Quality Considerations Overview</a> <a href="#">Download ToxCast Chemical Information</a> <a href="#">Download ReadMe</a>
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	<a href="#">Assay Annotation User Guide</a> <a href="#">Download Assay Information</a>

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

# Open Data means Reuse For Science

**Chemical  
Research in  
Toxicology**

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

Hao Zhu,<sup>\*,†,‡</sup> Jun Zhang,<sup>†,‡</sup> Marlene T. Kim,<sup>†,‡</sup> Abena Boison,<sup>†</sup> Alexander Sedykh,<sup>‡</sup> and Kimberlee Moran<sup>§</sup>

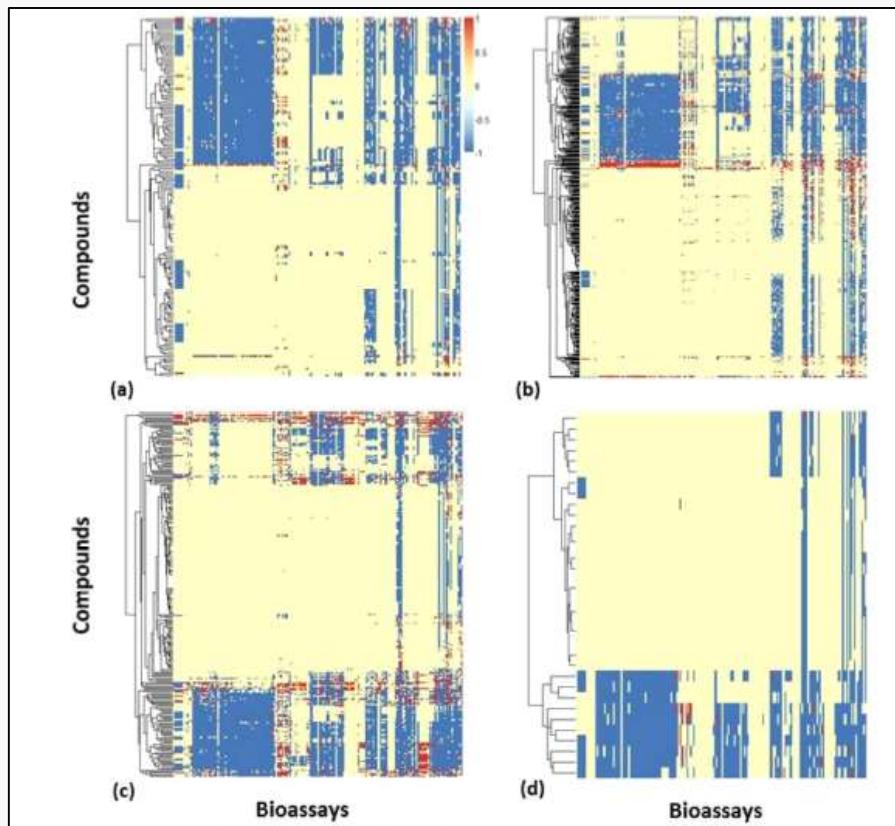


Figure 3. Response spaces of different ToxCast compound categories represented by the data obtained from 193 PubChem bioassays: (a) 171 consumer use chemicals (not including pharmaceuticals or pesticides), (b) 470 pesticides, (c) 245 pharmaceuticals, and (d) 34 phthalate plasticizers and alternatives.

# Open Data means Reuse For Software

# CompTox Dashboard

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

A screenshot of the CompTox Dashboard homepage. At the top, there's a navigation bar with links for Home, Advanced Search, Batch Search, Lists, Predictions, and Downloads. To the right of the navigation is a "Share" button. The main header area features the EPA logo and the text "762 Thousand Chemicals". Below this, there's a search bar with options for "Chemicals", "Product/Use Categories", and "Assay/Gene". The search bar also includes fields for "Search for chemical by systematic name, synonym, CAS number, IUPAC and INN" and "Identifier substring search". A note below the search bar says "See what people are saying, read the dashboard comments!" and "Cite the Dashboard Publication click here".

762 Thousand Chemicals

Chemicals Product/Use Categories Assay/Gene

Search for chemical by systematic name, synonym, CAS number, IUPAC and INN

Identifier substring search

See what people are saying, read the dashboard comments!

Cite the Dashboard Publication click here

### Latest News

Read more news

**YouTube video regarding using the Dashboard for Non-Targeted Analysis**

Ar  
Mar

ch 7th, 2018 at 9:43:36 AM

YouTube video discussing the application of the CompTox Chemistry Dashboard to support non-targeted analysis by mass spectrometry is available. This short video summarizes the advantages the dashboard in terms of data quality and focused data set for environmental non-targeted analysis. View it here on YouTube.

Discover.

About/Disclaimer  
Accessibility  
Privacy

Connect.

ACToR  
DSSTox  
Downloads

Ask.

Contact  
Help

# CompTox Dashboard Chemicals

United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads Share ▾

762 Thousand Chemicals

 [Chemicals](#) [Product/Use Categories](#) [Assay/Gene](#)

- Bisphenol A propoxylate
- Bisphenol A bis(2-hydroxyethyl ether) diacrylate DTXS104066997
- Bisphenol A bis(2-hydroxyethyl ether) dimethacrylate DTXS104066992
- Bisphenol A bis(2-hydroxypropyl) ether DTXS104067152
- Bisphenol A carbonate polymer DTXS1040677646
- Bisphenol A diglycidyl ether DTXS104067624
- Bisphenol A glycidyl methacrylate DTXS107044847
- Bisphenol A propoxylate diglycidyl ether DTXS104069088
- Bisphenol A propoxylate glycerobite diacrylate DTXS104060728

comptox-prod.epa.gov/dashboard

# CompTox Dashboard Products and Use Categories



The screenshot shows the CompTox Dashboard interface. At the top, there is a navigation bar with the EPA logo, followed by links for Home, Advanced Search, Batch Search, Lists, Predictions, and Downloads. A "Share" button is also present. The main content area displays a search bar with the query "hair color" and a results list titled "762 Thousand Chemicals". The results list includes several entries related to hair color products, such as:

- CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as permanent
- CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as for professional use
- CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as temporary
- CPDat PRODUCT category: personal care hair color  
hair coloring products not otherwise categorized
- CPDat PRODUCT category: personal care hair color activator  
chemicals/activators for hair coloring products
- CPDat PRODUCT category: personal care hair color developer  
chemical developers for hair coloring products
- CPDat PRODUCT category: personal care hair color toner  
chemical toners for hair coloring products

At the bottom of the page, there is a footer section with the EPA logo, links for Discover (About/Disclaimer, Accessibility, Privacy), Connect (ACToR, DSSTox, Downloads), and Ask (Contact, Help).

# CompTox Dashboard Assays and Genes



The screenshot shows the CompTox Dashboard interface. At the top, there's a navigation bar with the EPA logo, followed by links for Home, Advanced Search, Batch Search, Lists, Predictions, and Downloads. A "Share" button is also present. The main title "762 Thousand Chemicals" is centered above a search bar. The search bar has a placeholder "Q: estrogen" and a dropdown menu showing search results for various estrogen-related genes. The results include:

- GENE: ESR1  
estrogen receptor 1
- GENE: ESR2  
estrogen receptor 2 (ER beta)
- GENE: ESRRα  
estrogen-related receptor alpha
- GENE: ESRRβ  
estrogen-related receptor beta
- GENE: ESRRγ  
estrogen-related receptor gamma

Below the search results, a message reads: "and curating data, major updates to the batch searching functionality and access to real time predictions for both physiochemical and toxicity endpoints. A list of release notes is available for your review. We look forward to your feedback." At the bottom, there are footer sections for "Discover", "Connect", and "Ask", each with links to About/Disclaimer, Accessibility, Privacy, ACToR, DSSTox, Downloads, Contact, and Help. The footer also features the EPA logo.

# Downloadable data in useful formats

 United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads 

**Chemistry Dashboard** 

**DSSTox SDF File** Posted: 12/14/2016  
This zip file contains the entire chemical structure collection of over 700,000 chemicals from the DSSTox database contained in one large SDF file. The file contains the structure, The DSSTox Structure Identifier (DTXCID), The DSSTOX Substance Identifier (DTXSID listed as PubChem External Data Source), the associated Dashboard URL, associated synonyms and Quality Control Level details. In order to view an SDF file you will need to have access to the appropriate piece of software to open an SDF files. Examples include ChemAxon JChem, ACD/ChemFolder or ChemDraw.

**PHYSPROP Analysis File** Posted: 12/14/2016  
The data associated with the publication "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling" represents the curated data associated with the OPERA models used to predicted properties for the CompTox Chemistry Data. The data include the training and test data sets as well as the KNIME workflows used to perform the curation of the data. For a full understanding of the data and workflows we recommend accessing the publication also.

**DSSTox Mapping File** Posted: 12/14/2016  
The DSSTOX mapping file contains mappings between the DSSTox substance identifier (DTXSID) and the associated InChI String and InChI Key. The file is made available as a Tab Separated Value (TSV) file with each entry represented as shown:  
DTXSID7020001 InChI=1S/C11H9N3/c12-10-6-5-8-7-3-1-2-4-9(7)13-11(8)14-10/h1-6H,(H3,12,13,14) FJTNLJLPLJDTRM-UHFFFAOYSA-N

**DSSTox Predicted Property Data** Posted: 12/14/2016  
A number of property prediction models were developed using curated data as described in the publication "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling". These property prediction models include logP, water solubility, bioconcentration factor and many others. The files include DTXIDs, names and the predicted properties where possible. The models cannot predict properties for all chemicals contained in the database (for example, inorganics, organometallics and elements cannot be handled).

**DSSTox Synonyms File** Posted: 12/14/2016

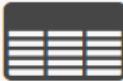
# EPA FigShare Page

<https://epa.figshare.com/>



A screenshot of the EPA FigShare page. The background is a collage of various environmental images, including cherry blossoms, a bridge over water, a sunset, a forest, and a large tree. At the top, there is a navigation bar with the FigShare logo, a search bar containing "search on figshare", and buttons for "Browse", "Upload", "Sign up", and "Log in". In the center, there is a white rectangular overlay containing the EPA logo and the text "Discover research from United States Environmental Protection Agency". Below this, there is a "Follow" button with a plus sign. The overall layout is clean and modern, designed to showcase scientific research and environmental data.

# Datasets with versioning

 DATASET	 DATASET	 DATASET	 DATASET
<a href="#">ReadMe for Animal Toxicity Study ToxRefDB files</a> EPA's National Center f... 29/03/2018	<a href="#">Collaborative Estrogen Receptor Activity Prediction Project (CERAP... EPA's National Center f...</a> 29/03/2018	<a href="#">ToxCast Database (invitroDB) EPA's National Center f...</a> 29/03/2018	<a href="#">Mapping file of InChIStrings, InChIKeys and DTXSIDs for the EP... Antony Williams</a> 12/08/2016
 DATASET	 DATASET	 DATASET	 DATASET
<a href="#">Chemistry Dashboard Data: Physprop Analysis Readme EPA's National Center f...</a> 29/03/2018	<a href="#">Chemistry Dashboard Data: Physprop Analysis EPA's National Center f...</a> 10/05/2018	<a href="#">Standard Laboratory Protocol for Tox21 Assays EPA's National Center f...</a> 29/03/2018	<a href="#">ToxCast &amp; Tox21 high-throughput assay information EPA's National Center f...</a> 29/03/2018
 DATASET	 DATASET	 DATASET	 DATASET
<a href="#">ReadMe for ToxCast and Tox21 high-throughput assay information EPA's National Center f...</a> 29/03/2018	<a href="#">Collaborative Estrogen Receptor Activity Prediction Project (CERAP... EPA's National Center f...</a> 29/03/2018	<a href="#">Chemistry Dashboard Data: Abstract Sifter User Guide EPA's National Center f...</a> 29/03/2018	<a href="#">ToxCast Database (invitroDB) for Mac Users EPA's National Center f...</a> 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. Judson

*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

Dimensions

 e.g. plastic AND instrument

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, toxicology, ecotoxicity, and toxicity bioassay data, informed through an integration hub with tools for additional EPA data and public datasets.

more

## Publication metrics

### Dimensions Badge



”



8

8

Total citations

Recent citations



n/a

n/a

Field Citation Ratio

Relative Citation Ratio

### Altmetric



24



Twitter (40)



Reddit (1)



Mendeley (20)



CiteULike (2)

Field Citation Ratio  
Relative Citation Ratio



# The impact of our paper versus the impact of our data...187 downloads

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

Dataset posted on 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.

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>

797  
views

187  
downloads

2  
citations



### CATEGORIES

- [Cheminformatics](#)
- [Computational Chemistry](#)
- [Environmental Chemistry](#)

### KEYWORD(S)

- [EPA CompTox Dashboard](#) [DTXSID](#)
- [InChIKeys](#) [Environmental Chemistry](#)
- [EPA](#)

### LICENCE



CC0

### EXPORT

- [RefWorks](#)  
[BibTeX](#)

F

indable



A

ccessible



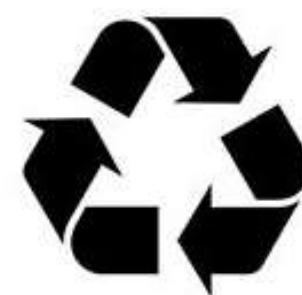
I

nteroperable



R

eusable



FAIRsharing.org standards, databases, policies

Search all of FAIRsharing Standards Databases Policies Collections Add/Claim Content Stats Log In or Register

databases > doi:10.25504/fairsharing.tfj7gt Suggest an edit/Questions?

## R EPA Comptox Chemistry Dashboard

### General Information

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

Homepage <https://comptox.epa.gov>  
Developed in United States  
Created in 2016

### Scope and data types

Bioactivity Chemistry Environmental Science Environmental Contaminant Exposure Physical Properties Spectroscopy Toxicology

# Examples of our transparency

## 1. OPERA Prediction Models



[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

32 Shares    1.6k Downloads    2 Citations



# What are OPERA Models?

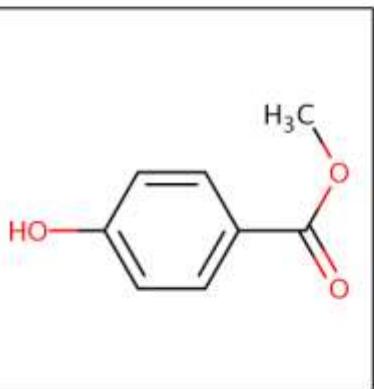
United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads Share Search all data

Save PDF

## OPERA Models: LogP: Octanol-Water

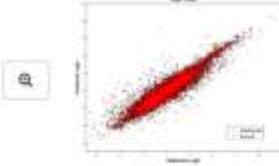
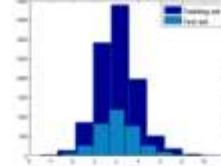
Methylparaben  
99-76-3 | DTXSID4022529



**Model Results**

Predicted value: 1.91  
Global applicability domain: Inside  
Local applicability domain index: 0.91  
Confidence level: 0.87

**Model Performance**

Weighted KNN model

QMRF

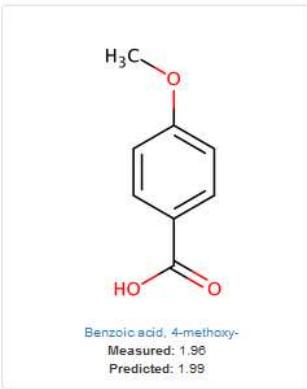
5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.86	0.67	0.86	0.78

Nearest Neighbors from the Training Set

  
Methylparaben  
Measured: 1.96  
Predicted: 1.91

  
Methyl 3-hydroxybenzoate  
Measured: 1.89  
Predicted: 1.91

  
Benzoin  
Measured: 2.02  
Predicted: 2.00

  
Benzeneacetic acid, 4-methoxy-, methyl ester  
Measured: 1.96  
Predicted: 1.99

  
Benzeneacetic acid, 4-hydroxy-, methyl ester  
Measured: 1.63  
Predicted: 1.60

# What are OPERA Models? Detailed QMRF reports

	<b><i>QMRF identifier (JRC Inventory):Q17-16-0016</i></b>
	<b><i>QMRF Title:</i></b> OPERA-model for Octanol-water partition coefficient
	<b><i>Printing Date:</i></b> 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

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

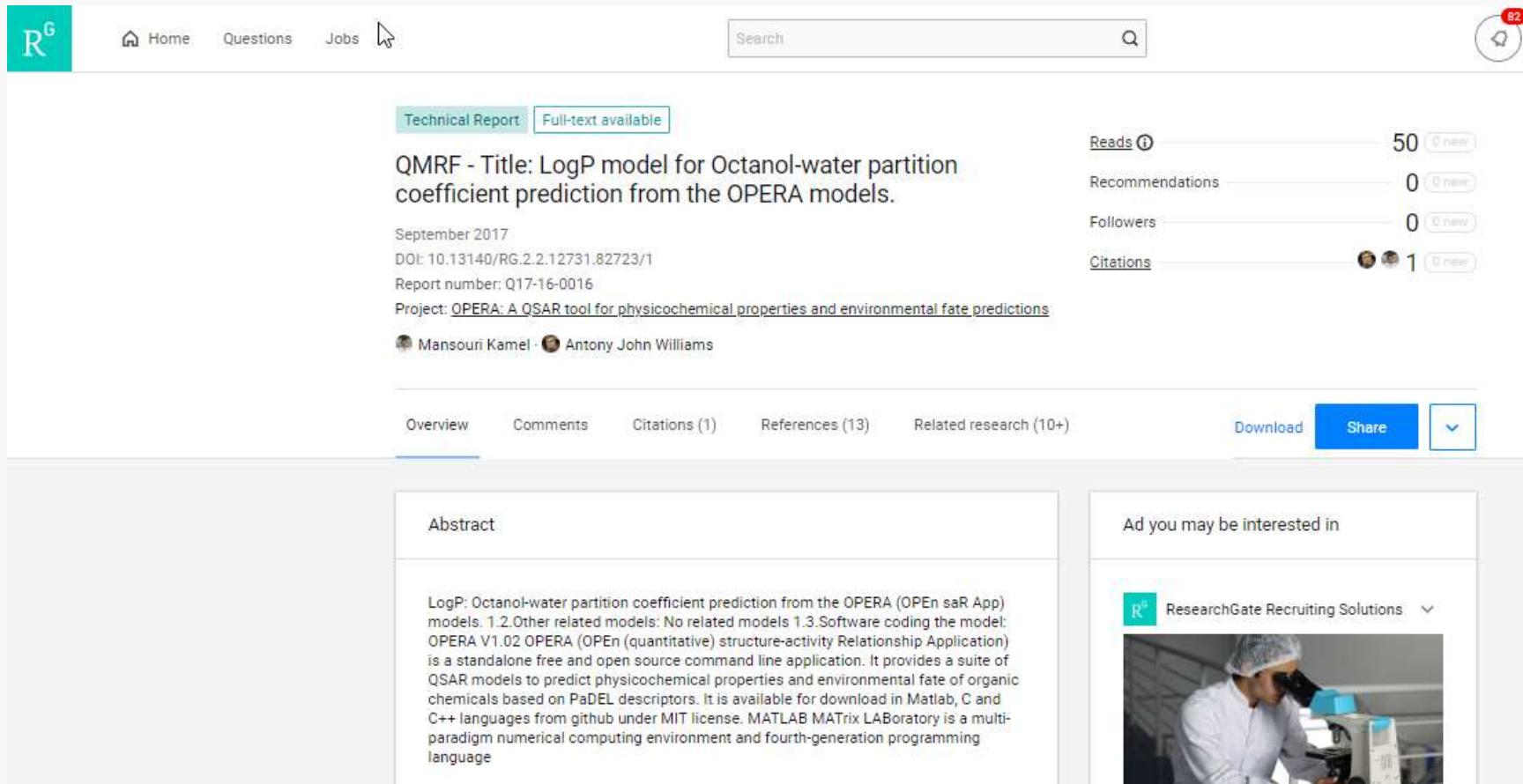
# How Transparent with the Code?

<https://github.com/kmansouri/OPERA>



 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?



The screenshot shows a ResearchGate profile page for a technical report titled "QMRF - Title: LogP model for Octanol-water partition coefficient prediction from the OPERA models". The page includes a search bar, navigation links (Home, Questions, Jobs), and social metrics (Reads: 50, Recommendations: 0, Followers: 0, Citations: 1). The abstract section describes the LogP model as part of the OPERA (OPEN saR App) and provides details about its software coding and availability. A sidebar on the right suggests related research.

R<sup>G</sup>

Home Questions Jobs

Search

QMRF - Title: LogP model for Octanol-water partition coefficient prediction from the OPERA models.

September 2017

DOI: 10.13140/RG.2.2.12731.82723/1

Report number: Q17-16-0016

Project: [OPERA: A QSAR tool for physicochemical properties and environmental fate predictions](#)

Mansouri Kamel · Antony John Williams

Overview Comments Citations (1) References (13) Related research (10+) Download Share

Abstract

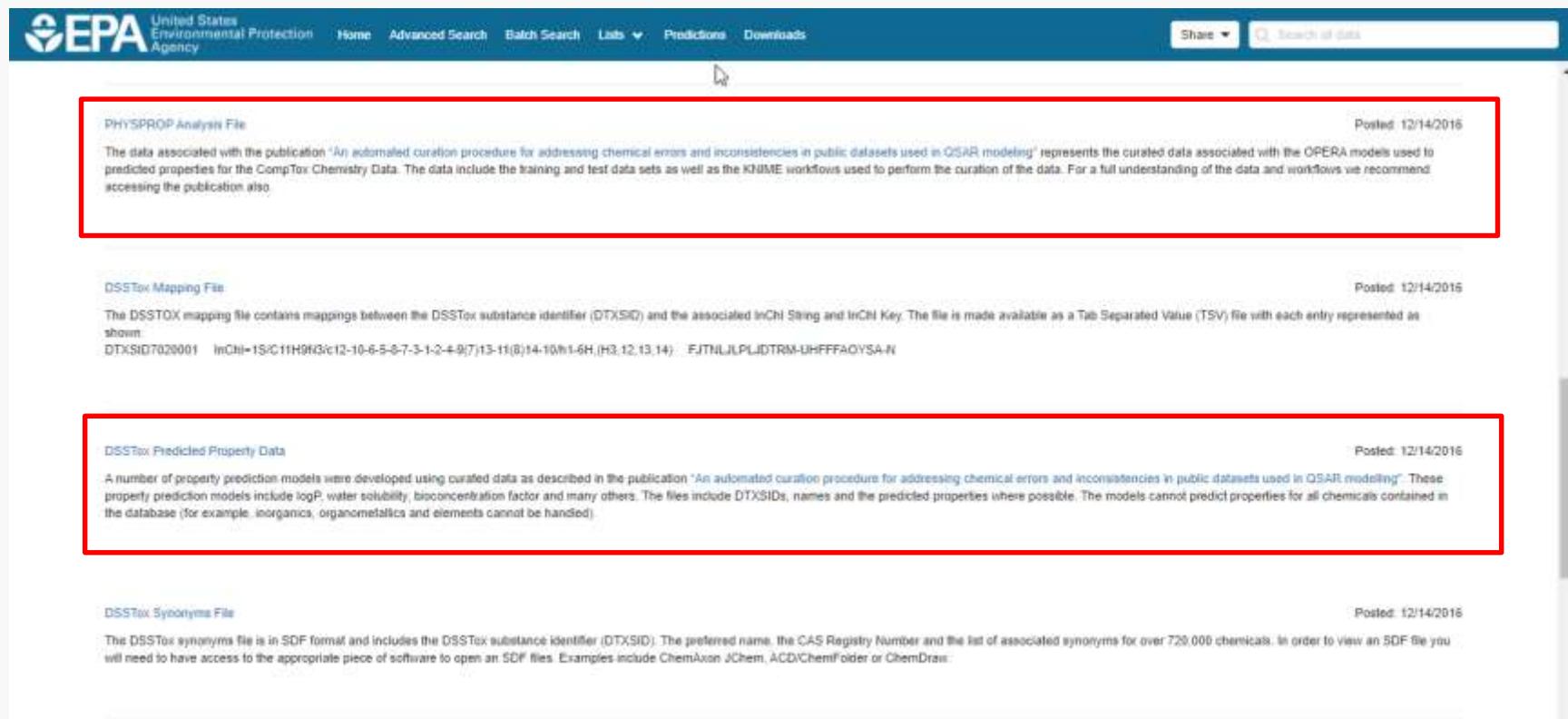
LogP: Octanol-water partition coefficient prediction from the OPERA (OPEN saR App) models. 1.2. Other related models: No related models 1.3. Software coding the model: OPERA V1.02 OPERA (OPEN (quantitative) structure-activity Relationship Application) is a standalone free and open source command line 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. MATLAB MATrix LABoratory is a multi-paradigm numerical computing environment and fourth-generation programming language

Ad you may be interested in

R<sup>G</sup> ResearchGate Recruiting Solutions



# How widely do we share?



The screenshot shows a web page from the United States Environmental Protection Agency (EPA) featuring four data sharing entries:

- PHYSPROP Analysis File** (Posted: 12/14/2016)  
The data associated with the publication "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling" represents the curated data associated with the OPERA models used to predict properties for the CompTox Chemistry Data. The data include the training and test data sets as well as the KNIME workflows used to perform the curation of the data. For a full understanding of the data and workflows we recommend accessing the publication also.
- DSSTox Mapping File** (Posted: 12/14/2016)  
The DSSTox mapping file contains mappings between the DSSTox substance identifier (DTXSID) and the associated InChI String and InChI Key. The file is made available as a Tab Separated Value (TSV) file with each entry represented as shown:  
DTXSID7028001 InChI=1S/C11H963/c12-10-6-5-8-7-3-1-2-4-9(7)13-11(8)14-10/h1-6H,(H3,12,13,14) . FJTNLJLPLIDTRM-UHFFFAOYSA-N
- DSSTox Predicted Property Data** (Posted: 12/14/2016)  
A number of property prediction models were developed using curated data as described in the publication "An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling". These property prediction models include logP, water solubility, bioconcentration factor and many others. The files include DTXSIDs, names and the predicted properties (where possible). The models cannot predict properties for all chemicals contained in the database (for example, inorganics, organometallics and elements cannot be handled).
- DSSTox Synonyms File** (Posted: 12/14/2016)  
The DSSTox synonyms file is in SDF format and includes the DSSTox substance identifier (DTXSID). The preferred name, the CAS Registry Number and the list of associated synonyms for over 720,000 chemicals. In order to view an SDF file you will need to have access to the appropriate piece of software to open an SDF file. Examples include ChemAxon JChem, ACD/Chem3D or ChemDraw.

# Journals for Data Articles



# Data Journals Hold Promise

Helping you publish, discover,  
and reuse research data



## Credit

Credit, through a citable publication, for depositing & sharing your data



## Reuse

Complete, curated & standardized descriptions enable the reuse of your data



## Quality

Rigorous community based peer review



## Discovery

Find datasets relevant to your research



## Open

Promotes & endorses open science principles & available to all through a Creative Commons license



## Service

In-house curation, rapid peer review & publication of your data descriptions

# Examples of our transparency

## 2. The Chemical and Products Database



The screenshot shows a web page for a scientific article. At the top left is a "MENU" button with a hand cursor icon. To its right is the title "SCIENTIFIC DATA" in large, white, sans-serif capital letters. Below the title is a small graphic of binary code: "10110", "0111001", "11011110", and "01101101". Underneath the title, there are three links: "Data Descriptor" (gray), "OPEN" (red), and "Published: 10 July 2018" (blue). The main content area features a large, dark gray text block describing the database. Below this, author information is listed in blue text, followed by a "Download Citation" link at the bottom.

MENU

# SCIENTIFIC DATA

10110  
0111001  
11011110  
01101101

Data Descriptor | OPEN | Published: 10 July 2018

## The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products

Kathie L. Dionisio, Katherine Phillips, Paul S. Price, Christopher M. Grulke, Antony Williams, Derya Biryol, Tao Hong & Kristin K. Isaacs

*Scientific Data* **5**, Article number: 180125 (2018) | Download Citation

# Examples of our transparency

## 2. The Chemical and Products Database

United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads Share ▾

762 Thousand Chemicals

Chemicals Product Use Categories Assay/Gene

Q: hair color

CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as permanent

CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as for professional use

CPDat PRODUCT category: personal care hair color  
hair colors and dyes characterized as temporary

CPDat PRODUCT category: personal care hair color  
hair coloring products not otherwise categorized

CPDat PRODUCT category: personal care hair color activator  
chemicals/activators for hair coloring products

CPDat PRODUCT category: personal care hair color developer  
chemical developers for hair coloring products

CPDat PRODUCT category: personal care hair color toner  
chemical toners for hair coloring products

Discover About/Disclaimer Accessibility Privacy Connect ACToR DSSTox Downloads Ask Contact Help



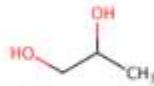
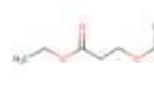
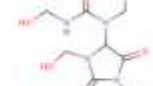
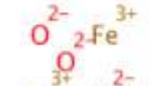
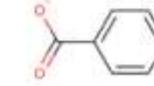
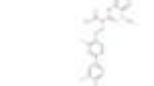
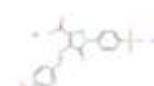
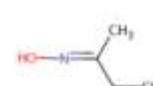
# What chemicals are in Arts and Crafts Paint?

United States Environmental Protection Agency Home Advanced Search Batch Search Lists Predictions Downloads Share Search all data

Searched by Product & Use Categories  
Results for CPDat Product Category: Arts And Crafts: Arts And Crafts Paint

84 chemicals

Show info: DTXSID CASRN TOXCAST Select all Filter by: Name or CASRN Hide

 <p>1,2-Propanediol DTXSID: DTXSID00021208 CASRN: 57-95-5 TOXCAST: 1/539</p>	 <p>Eisan DTXSID: DTXSID00045234 CASRN: 17372-67-1 TOXCAST: 45/302</p>	<p>1 related chemical structure with this substance</p> <p>Poly(vinylpyridine) DTXSID: DTXSID0025941 CASRN: 6003-39-8 TOXCAST: 0</p>	 <p>Ethyl 3-ethoxypropanoate DTXSID: DTXSID0027108 CASRN: 793-95-8 TOXCAST: 0</p>	<p>2 related chemical structures with this substance</p> <p>Dipropylene glycol monomethyl ether DTXSID: DTXSID0027963 CASRN: 24580-94-8 TOXCAST: 8/615</p>	 <p>Diisobutyl urea DTXSID: DTXSID0028569 CASRN: 73491-02-8 TOXCAST: 8/295</p>
 <p>Iron(II) oxide DTXSID: DTXSID0029632 CASRN: 1309-37-1 TOXCAST: 0</p>	 <p>CI Pigment Red 5 DTXSID: DTXSID0026394 CASRN: 8410-41-9 TOXCAST: 0</p>	 <p>Sodium benzoate DTXSID: DTXSID1020140 CASRN: 532-32-1 TOXCAST: 1/540</p>	 <p>CI Pigment Yellow 55 DTXSID: DTXSID1021483 CASRN: 9987-15-7 TOXCAST: 0</p>	 <p>FD&amp;C Yellow 5 DTXSID: DTXSID1021459 CASRN: 1934-21-0 TOXCAST: 17/572</p>	 <p>2-Butanone oxime DTXSID: DTXSID1021221 CASRN: 66-29-7 TOXCAST: 4/500</p>

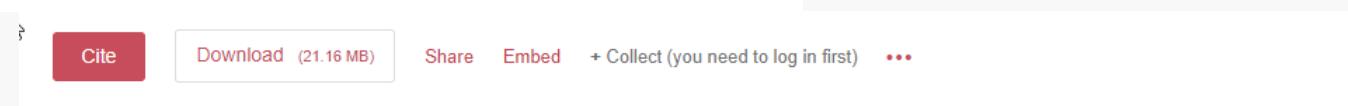
# Examples of our transparency

## 2. The Chemical and Products Database

### 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

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...



#### CATEGORIES

- Toxicology

#### KEYWORD(S)

- Computational Toxicology
- CPDat
- Chemical and Products Database
- MySQL

#### LICENCE



# 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

*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

2018

# Use APPROPRIATE formats

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



**High-throughput, computer assisted,  
specific MetID. A revolution for drug  
discovery**

# Chemical structures as text

**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(OC)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(OC)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(OC)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)(CCCN)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( 0)c(c1) 0

c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC

# How do I extract structures?

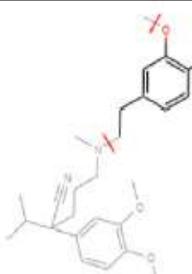
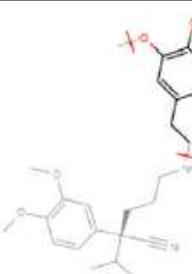
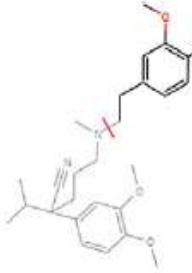
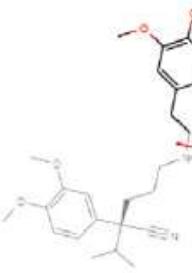
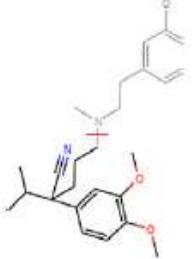
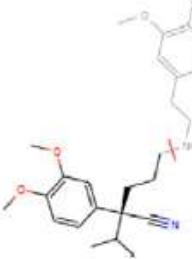
- Copy and paste into Excel as a start point
- Assume no loss of formatting!
- Convert SMILES to structures
- Copy-Paste doesn't work

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

c1(CCNC)ccc( 0)c(c1) 0  
c1cc(CC NCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c10C

# Structure Drawings Are Worse

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	$\Delta$	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**

<u>Name of Compound</u>	<u>CAS No.</u>	<u>Quantitation Ion</u>	<u>Qualifier Ions</u>
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...

<u>Name of Compound</u>	<u>CAS No.</u>
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

Integrated Biological Pathway model for the Estrogen Receptor

Search ERMODEL Chemicals

List Details

Description: Dataset associated with "Integrated Model of Chemical Perturbations of a Biological Pathway Using 16 In Vitro High-Throughput Screening Assays for the Estrogen Receptor" by Judson et al. ([LINK](#)). A computational network model that integrates 16 in vitro, high-throughput screening assays measuring estrogen receptor (ER) binding, dimerization, chromatin binding, transcriptional activation, and ER-dependent cell proliferation. The network model uses activity patterns across the in vitro assays to predict whether a chemical is an ER agonist or antagonist, or is otherwise influencing the assays through a manner dependent on the physics and chemistry of the technology platform ("assay interference"). The method is applied to a library of 1812 commercial and environmental chemicals, including 45 ER positive and negative reference chemicals.

Number of Chemicals: 1812

1812 chemicals

Download / Send

Show info: DTXSID CASRN TOXCAST Select all

Sort by: DTXSID

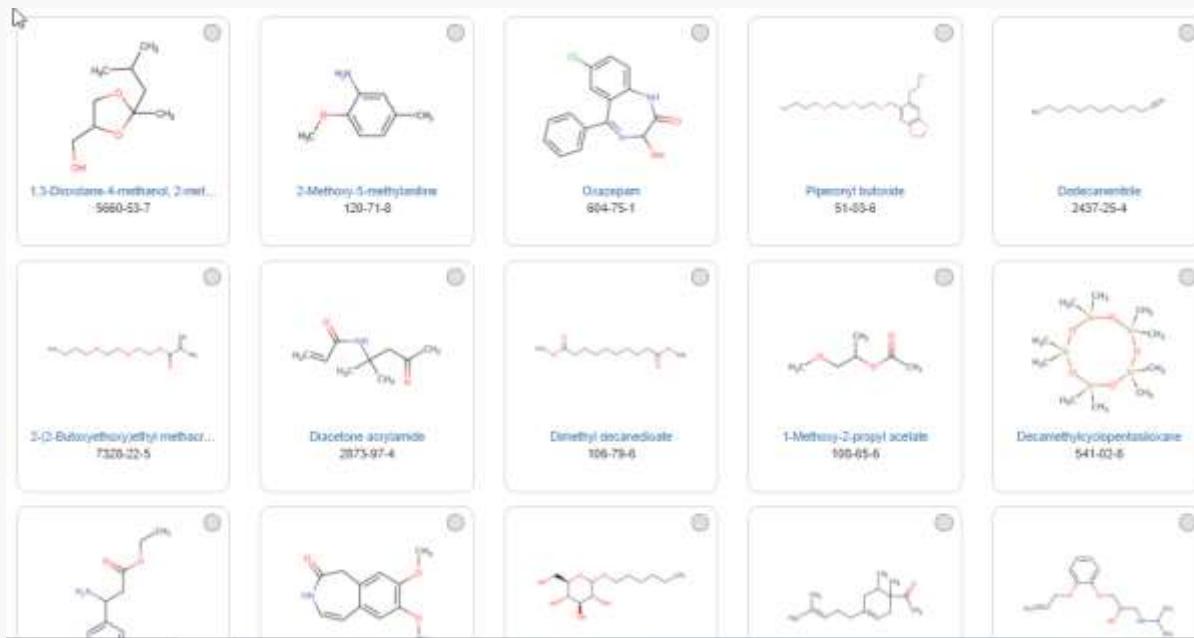
Filter by: Name or CASRN Hide



The screenshot shows a list of 1812 chemicals. At the top, there are buttons for 'Download / Send' and 'Select all'. Below these are 'Show info' buttons for DTXSID, CASRN, and TOXCAST, along with a 'More' button. There are also 'Sort by' and 'Filter by' dropdown menus. The main area displays four chemical structures: 1) A complex polychlorinated biphenyl-like molecule; 2) 4-Aminopyrimidine; 3) 2-hydroxy-3-hydroxymethyl-4-oxo-5-oxazolylcyclopentanecarboxylic acid; 4) A cyclic amide derivative of cyclohexene.

# 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

Integrated Biologic

Search ERMODEL Chemicals

Substring search

Download as

List

TSV

Design

Excel

SDF

Number

Send to

Batch search

Download / Send

# 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

 A circular icon containing the number '10' with a red border, indicating the altmetric score of the publication.

CATEGORIES:

- Cheminformatics
- Computational Chemistry
- Environmental Chemistry

KEYWORD(S):

- EPA CompTox Dashboard
- DTSID
- InChIKeys
- Environmental Chemistry
- EPA

# PERSONAL affection for Open Peer Review

**F1000Research**  
Open for Science

Search  SUBMIT YOUR RESEARCH

BROWSE GATEWAYS HOW TO PUBLISH ABOUT BLOG MY RESEARCH SIGN IN

SOFTWARE TOOL ARTICLE

**Abstract Sifter: a comprehensive front-end system to PubMed [version 1; referees: 2 approved]**

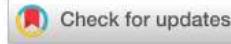
✉ Nancy Baker <sup>1</sup>, Thomas Knudsen<sup>2</sup>, Antony Williams <sup>2</sup>

[Author details.](#)

 This article is included in the [Chemical Information Science gateway](#).

**Abstract**

The Abstract Sifter is a Microsoft Excel based application that enhances existing search capabilities of PubMed. The Abstract Sifter assists researchers to search effectively, triage results, and keep track of articles of interest. The tool implements an innovative "sifter" functionality for relevance

**Check for updates** 

**METRICS**

640
VIEWS
119
DOWNLOADS

**Open Peer Review**

Referee Status:  

Version(s)	1	2
Version 1 published 21 Dec 2017	 <a href="#">read report</a>	 <a href="#">read report</a>

1 Pauliina Damdimopoulou  , Karolinska Institutet, Sweden  
Astrud Tuck, Karolinska Institutet, Sweden

2 Qingliang (Leon) Li  , National Institutes of Health, USA

[All reports \(2\)](#)


# 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...

## ▼ What if you run out of funding? What happens to my data?



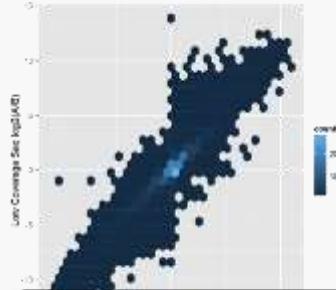
COS established a \$250,000 preservation fund for hosted data in the event that COS had to curtail or close its offices. If activated, the preservation fund will preserve and maintain read access to hosted data. This fund is sufficient for 50+ years of read access hosting at present costs. COS will incorporate growth of the preservation fund as part of its funding model as data storage scales. For information about OSF backups and technical preservation details, see the [OSF Backup and Preservation Policy](#).

# 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

- The last public release of ToxCast data (invitroDB\_v2) was in 3<sup>rd</sup> 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

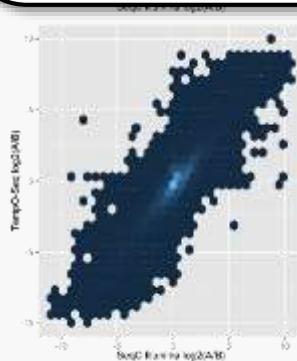


TruSeq  
 $r^2$  0.74

## 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...

## **Antony Williams**

US EPA Office of Research and Development

National Center for Computational Toxicology (NCCT)

[Williams.Antony@epa.gov](mailto:Williams.Antony@epa.gov)

ORCID: <https://orcid.org/0000-0002-2668-4821>