

An Online Prediction Platform to Support the Environmental Sciences

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

*National Center for Computational Toxicology,
U.S. Environmental Protection Agency, Research Triangle Park, NC*

This work was reviewed by the U.S. EPA and approved for presentation but does not necessarily reflect official Agency policy.

*March, 13-17, 2016
ACS Spring Meeting, San Diego, CA*

EPA's National Center for Computational Toxicology (NCCT)



- ToxCast (EPA) & Tox21 (Multi-Agency)
 - screening >3800 (ToxCast) to >10K (Tox21) environmentally relevant chemicals across 10's to 100's of HTS assays
- ToxRef DB – guideline animal toxicity study reference DB
- ExpoCast (DBs, models), Rapid Toxicity Assessments
 - rapid exposure estimates, integration of hazard and exposure (IVIVE) estimates for RA
- Public-facing, web-dashboards
 - facilitate access to & utility of EPA data with web tools

*Chemical
databases*

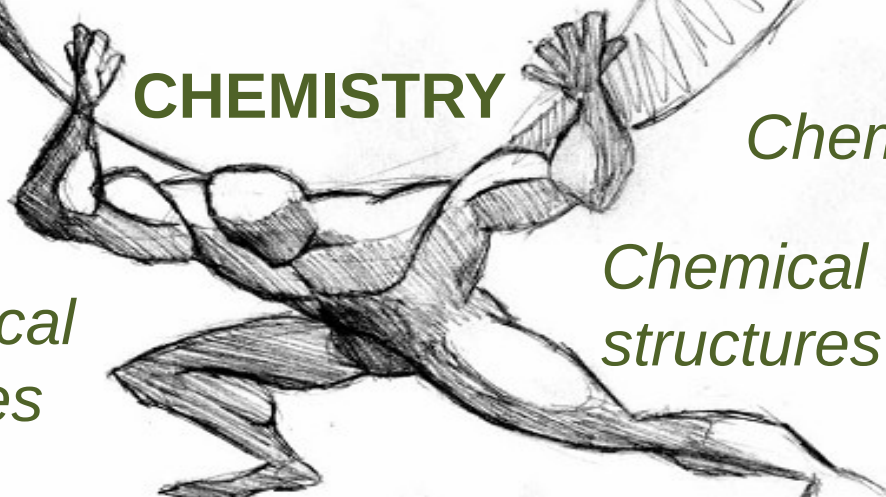
CHEMISTRY

Cheminformatics

*Chemical
linkages*

*Chemical
structures*

*SAR/QSAR
models*



NCCT (iCSS) Chemistry Dashboard



- Why build yet another public chemistry tool?
- Who is it being built for?
- How is it being built?
- What functionality & data will be included?

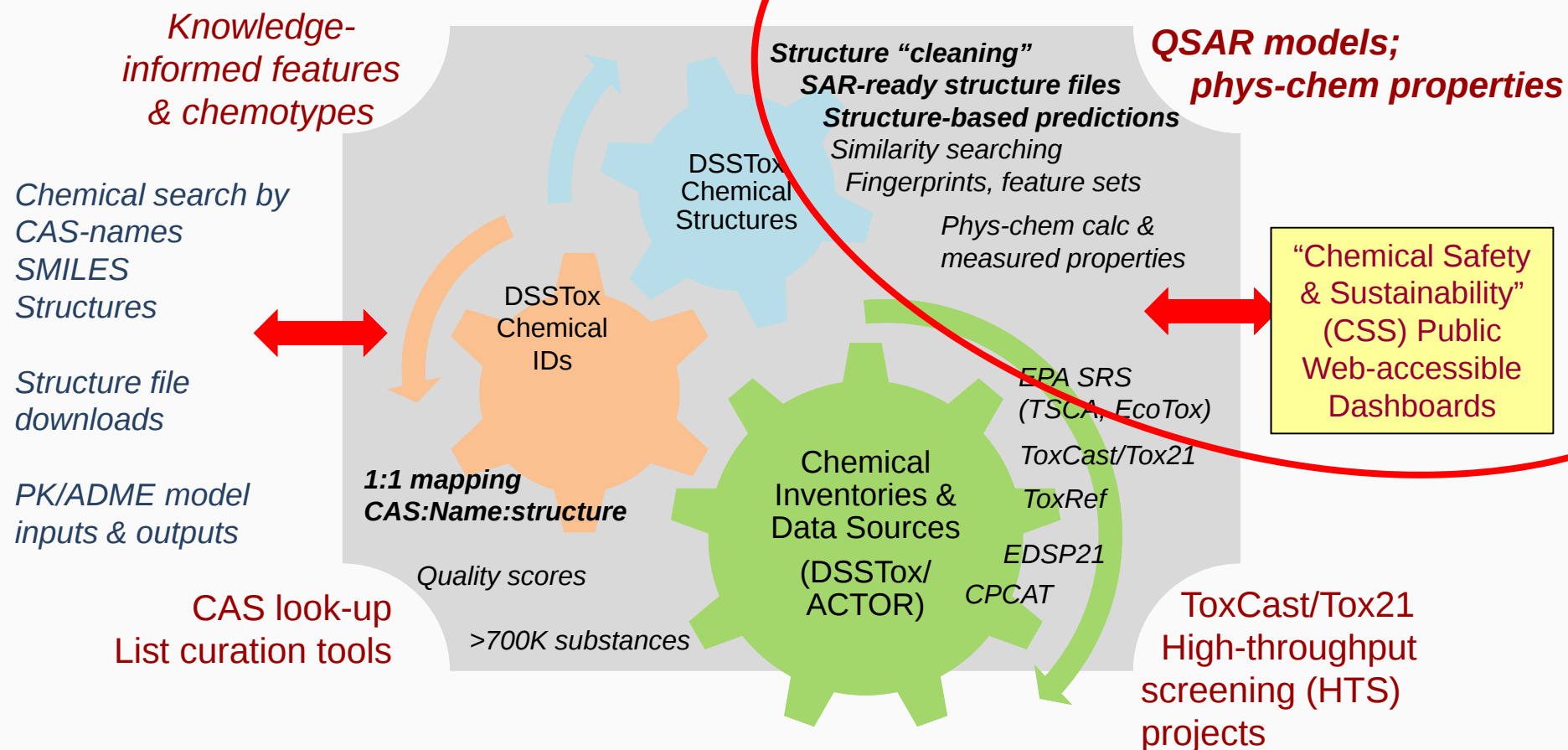
Using Chemistry to improve linkages across EPA Resources



- NCCT ToxCast/EDSP Dashboards
- NCCT ExpoCast, CPCat, ACToR
- EPA Chemical Safety & Sustainability (CSS) research
- EPA Substance Registry System (SRS)
- EPA Integrated Risk Information System (IRIS)
- EPA Office of Pesticides (Inerts, Active Ingredients)
- EPA Endocrine Disruption Screening Program (EDSP)
- EPA Office of Water's Chemical Contaminants List
- EPA Office of Pollution Prevention & Toxics ChemView

Just lists (pdfs, tables, names, CAS, etc.)!!

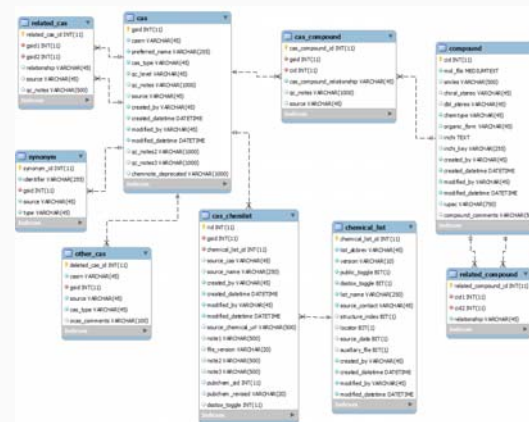
Chemistry Foundation to Support Multiple NCCT & EPA Projects



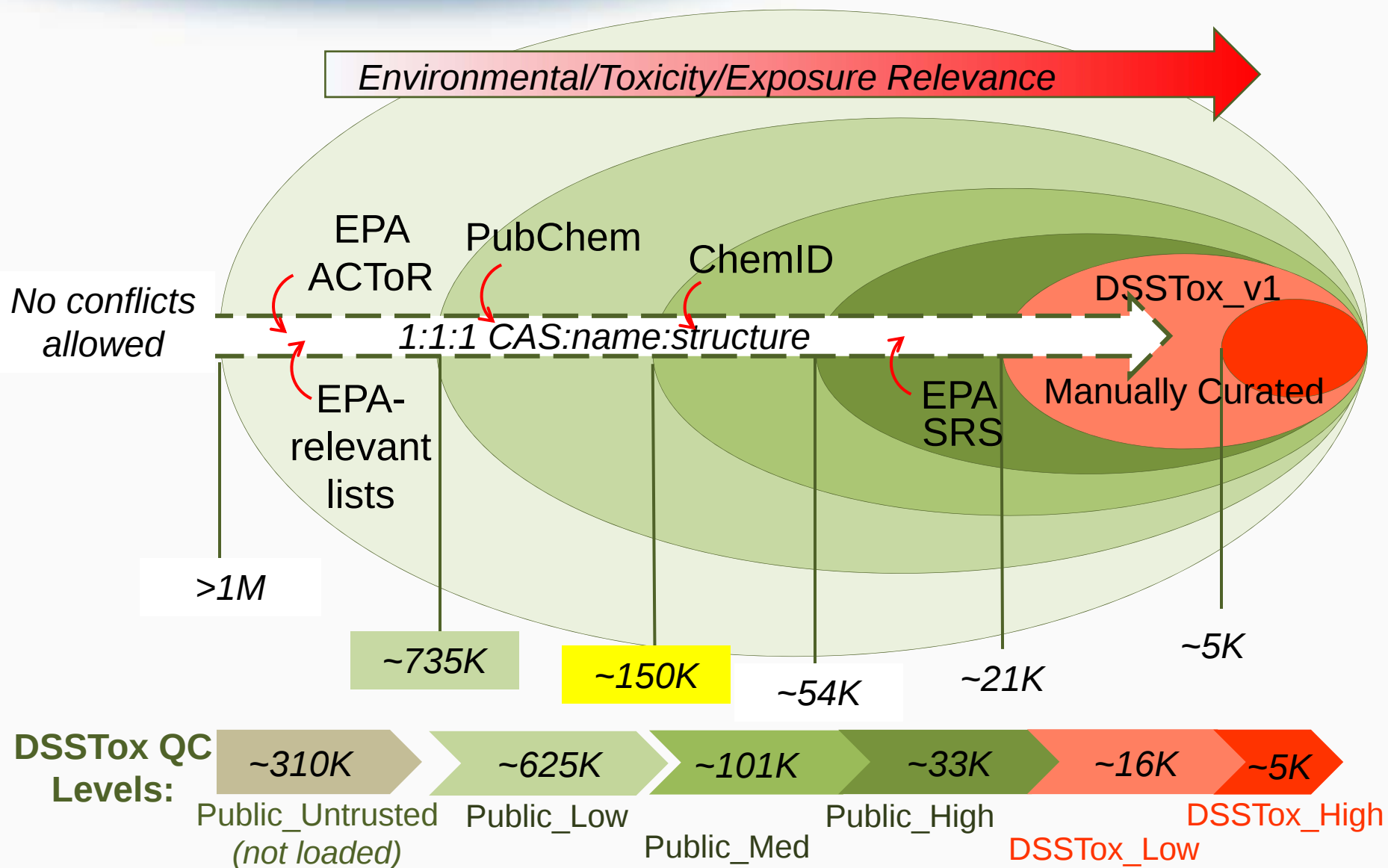
United States
Environmental Protection
Agency

- Convert DSSTox tables to MySQL
- Develop curation interface
- Implement cheminformatics workflow
- Expand chemical content
- Expand data inventories
- Web-services & Dashboard access

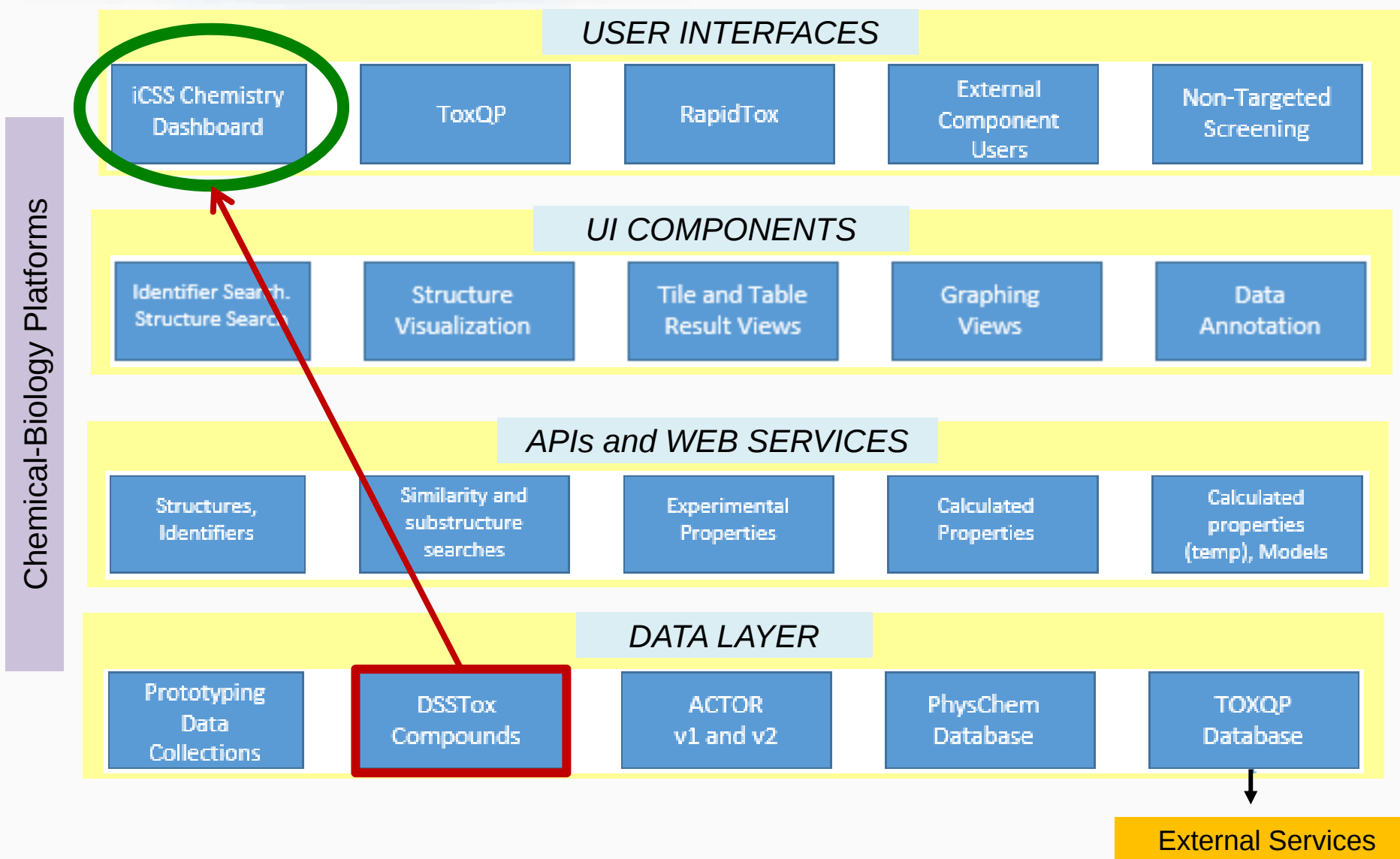
- manually curated 25K substance records
- EPA-focus, environmental tox datasets
- Emphasis on accurate CAS-name-structure annotations
- Public resource for high-quality structure-data files (SDF)



EPA's Distributed Structure-Searchable Toxicity Database (DSSTox)



Building the Chemistry Software Architecture



iCSS Chemistry Dashboard



D
E
V
E
L
O
P
M
E
N
T

- Web interface supporting EPA's Chemical Safety & Sustainability research (iCSS)
- To be released April 2016, currently in beta-testing
- Will provide public access to DSSTox db
- UI in development only <4mo, building on substantial in-house data resources
- Initially name/ID-based searching only ...



Chemistry Dashboard

Search a chemical by systematic name, synonym, CAS number, or InChIKey



☐ Single component search ☐ Ignore isotopes

Need more? Use [advanced search](#).

734 Thousand Chemicals

Componentizing API for Resolving Chemical Names



Search a chemical by systematic name, synonym, CAS number, or InChIKey



- Look familiar?
 - *ChemSpider, PubChem, ChEMBL, NCI Resolver ...*
- ONE reusable EPA component API that searches for Names and CAS Registry Numbers...
- ... and SMILES, and InChIKeys, and recognizes invalid CAS Numbers, and converts SMILES when they aren't in DB, etc. etc.

iCSS Chemistry Dashboard Releasing in April 2016



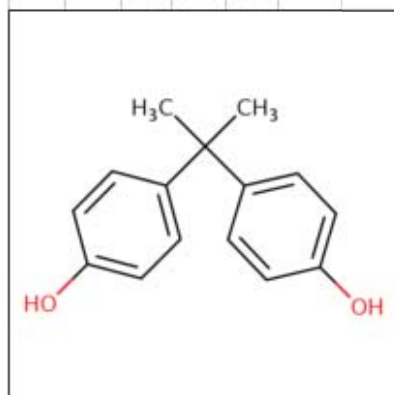
Bisphenol A

80-05-7 **DTXSID7020182**

Formerly DSSTox GSID, new unique public substance ID, essential for RDF/semantic web applications

? Searched by Synonym: Found 1 result for 'bisphenol A'.

2D 3D



Intrinsic Properties

Molecular Formula: C₁₅H₁₆O₂

Average Mass: 228.291 g/mol



Monoisotopic Mass: 228.11503 g/mol



Structural Identifiers

Citation

Chemical
Properties

External Links

Synonyms

PubChem
Biological Activities

PubChem Articles

PubChem Patents

Comments

CSV Excel

Property

Average (Exp.)

Range (Exp.)

Average (Pred.)

Range (Pred.)

[Solubility](#)

0.001 (1)

0.0005257 to 0.0005257

0.38 (2)

0.003675 to 0.7565

[Boiling Point](#)

200.0 (1)

200.0 to 200.0

348.95 (2)

334.4 to 363.5

[LogP](#)

3.357 (3)

3.32 to 3.431

3.524 (3)

3.205 to 3.727

[Atmospheric Hydroxylation Rate](#)

N/A

N/A

0.0 (1)

4.237e-11 to 4.237e-11

→ Melting Point

About

Contact

ACToR



DSSTox

Privacy

Accessibility

Help

iCSS Chemistry Dashboard

Releasing in April 2016



Chemical Properties: Melting Point

	Average	Range
Experimental	154.929 (7)	153.0 to 158.0
Predicted	144.033 (3)	131.8 to 158.0

CSV Excel

Property	Raw Result	Mean Result	Minimum Result	Maximum Result	Result Unit	Result Type	Source
Estimated MP (oC)	131.76	131.8	131.8	131.8	°C	predicted	EPI SUITE
Melting Point	153-158 °C	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	154-157 °C	155.5	154.0	157.0	°C	experimental	Merck Millipore
Melting Point	153-158 °C	155.5	153.0	158.0	°C	experimental	Alfa Aesar
Melting Point	155-158 °C	156.5	155.0	158.0	°C	predicted	J and K Scientific
Melting Point	156 °C	156.0	156.0	156.0	°C	experimental	TCI
Melting	153 °C	153.0	153.0	153.0	°C	experimental	Jean-Claude Br

- Original raw source result view
- Users can submit comments
- Source details provided
- Range of results possible for single chemical

iCSS Chemistry Dashboard Releasing in April 2016



- Links external resources: EPA, NIH, property predictors

The dashboard features a top navigation bar with tabs: Chemical Properties, External Links, Synonyms, PubChem Biological Activities, PubChem Articles, and PubChem Patents. Below this, there are five main categories: General, Toxicology, Publications, Biochemistry, and Prediction. Each category contains a list of external resources. A red box highlights the 'Chemicalize' link in the Prediction category. Another red box highlights the 'PubChem' link in the General category, with an arrow pointing to a detailed bioactivity summary table. A green callout bubble points to the 'Chemicalize' link, and another points to the bioactivity summary table.

General

- ToxCast Dashboard 2
- EPA Substance Registry Service
- NIST Chemistry Webbook
- eChemPortal
- Household Products Database
- HSDB
- PubChem**
- ChempSpider
- EDSP Dashboard
- CPCat
- DrugBank
- HMDB

Toxicology

- ACTOR
- DrugPortal

Publications

- Toxline
- Environmental Health Perspectives
- NIEHS
- National Toxicology Program
- Google Books

Biochemistry

- CCRIS

Prediction

- Chemicalize**

Bioactivity Summary Table:

Structure	Substance	Source	Type	Value (pTS)	Compound Name	Bioactive Name
	10011000	Source	Patent	62.000	Dependent A	pTS score for small molecule agonists of the p53 signaling pathway (pTS 0.000, Type: Confirmed)
	10011000	Source	Patent		Dependent A	pTS score for small molecule agonists that induce genotoxicity in human embryonic kidney cells expressing hTERT (pTS 0.000, Type: Confirmed)
	10011000	Source	Patent	55.000	Dependent A	pTS score for small molecule agonists of the p53 signaling pathway - cell viability (pTS 0.000, Type: Confirmed)
	10011000	Source	Patent		Dependent A	pTS score for small molecule agonists that induce genotoxicity in human embryonic kidney cells expressing hTERT (pTS 0.000, Type: Confirmed)
	10011000	Source	Patent		Dependent A	pTS of p53 Response Regulator Agonist: pTS (pTS 0.000, Type: Screening)

e.g., ChemAxon's
"Chemicalize" web-service

e.g., PubChem's
bioactivity summary

Linkage to External Predictor, e.g. Chemicalize



Chemicalize.org (powered by ChemAxon)

← → ↺ ⬆ www.chemicalize.org/structure/#!m

chemicalize.org
by ChemAxon

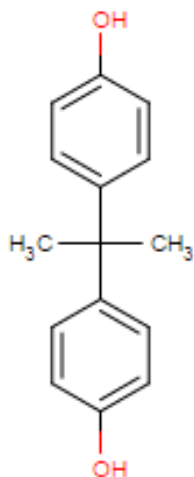
Properties Viewer

Webpage Viewer Chem Search Web Search Doc

CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1

Molecule

Molecule



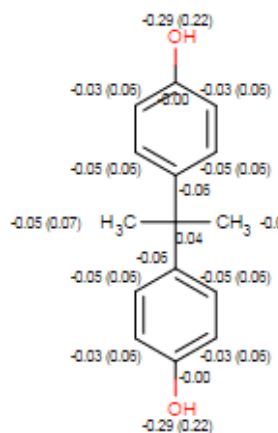
Names and identifiers

Common names: 4,4'-bisphenol A, 4,4'-bisphenol A, diphenylolpropyl bisphenol A, bisphenol-A, bisphenol A, 4,4'-isopropylidenediphenol, IUPAC: 4-[2-(4-hydroxyphenyl)propan-2-yl]phenol

Smiles: CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1
InChI: 1S/C15H16O2/c1-15(2,11-3-7-13(16)8-4-11)12-5-9-14(17)10-6
InChI key: IISBACLAFFKSPIT-UHFFFAOYSA-N
CAS: 27100-33-0, 80-05-7, 137885-53-1, 27360-89-0, 28108-82-3, ...

Charge

Atomic
Charges



Elemental Analysis

Formula: C₁₅H₁₆O₂
Isotope formula: C₁₅H₁₆O₂
Composition: C (78.92%), H (7.06%), O (14.02%)
Isotope composition: C (78.92%), H (7.06%), O (14.02%)
Mass: 228.2883
Exact mass: 228.115029758

Webpages

Sulfone

29 Apr 2015 - original page
wn.com/Sulfone

Microsoft Word - NLPFIN New.doc

27 Mar 2014 - original page
reach-clp-helpdesk.de/de/Downloads/NLP-Liste.pdf?__blob=pul

No title

17 May 2014

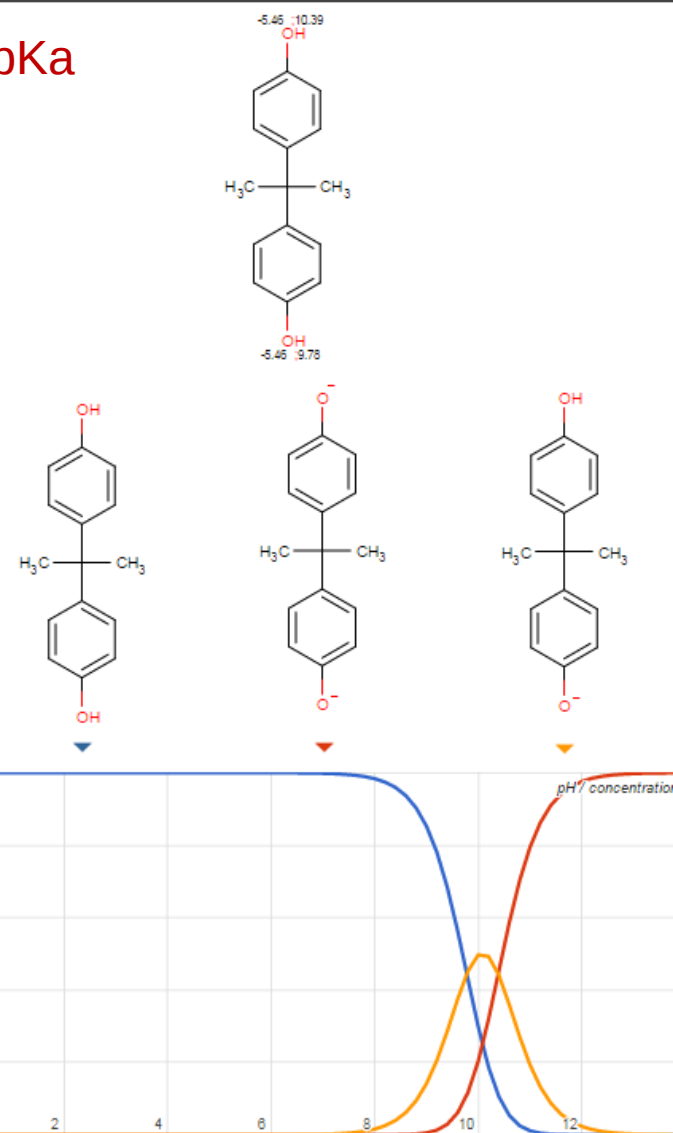
No title

17 Oct 2012 - original page
patbase.com/getimg/prt_text.asp?id=1551799&pn=EP1248780E

Breast Neoplasms - Genes | CTD

20 Nov 2015 - original page
ctdbase.org/detail.go?sessionId=E301927D5D648246D5BAA72

pKa



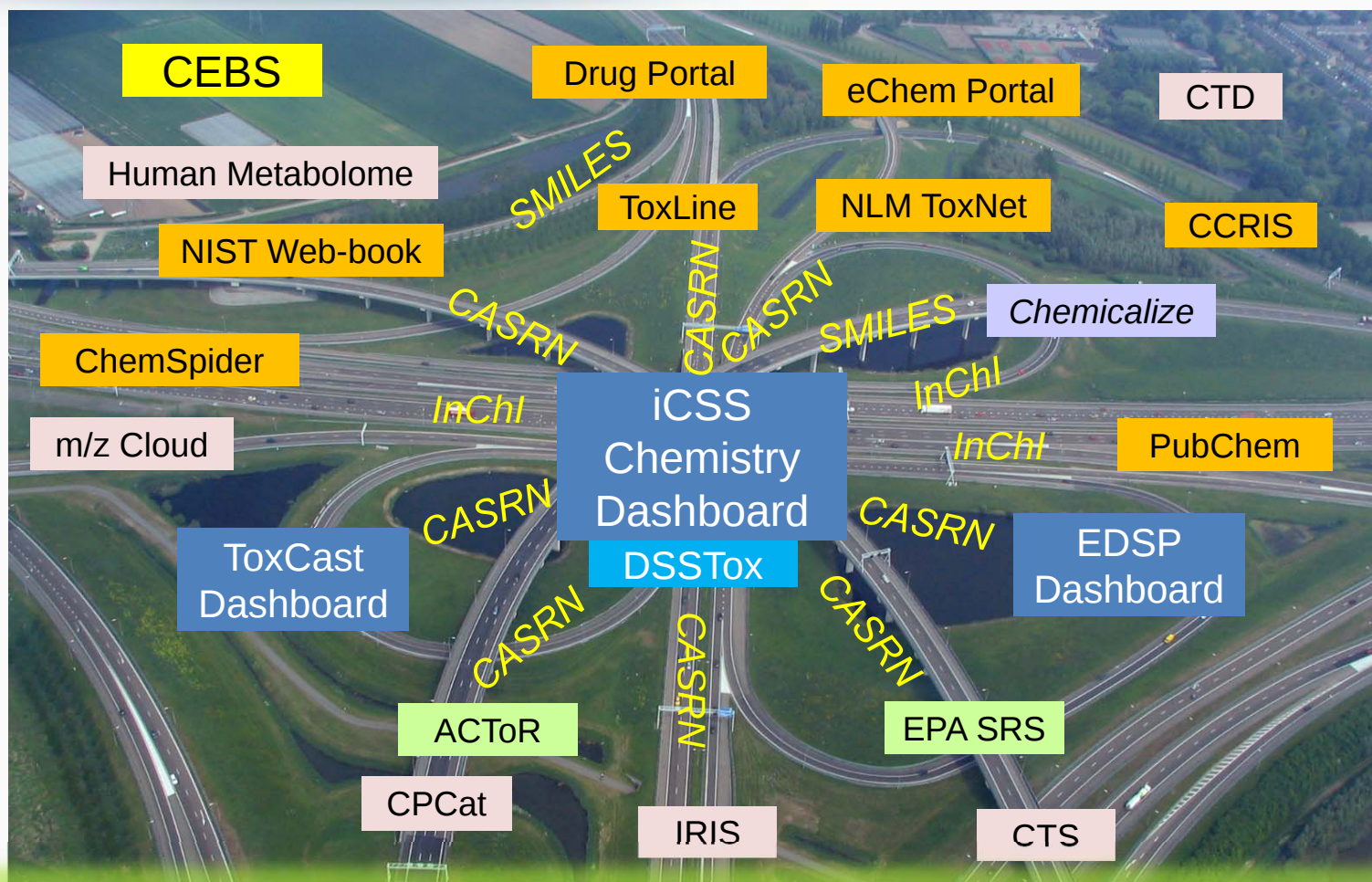
How are linkages created? e.g., CASRN



- ACTOR <http://actor.epa.gov/actor/GenericChemical?casrn=80-05-7>
- Toxcast Dashboard <http://actor.epa.gov/dashboard2/#chemical/80-05-7>
- NIST Webbook <http://webbook.nist.gov/cgi/cbook.cgi?ID=80-05-7>
- TOXNET <http://toxnet.nlm.nih.gov/cgi-bin/sis/search2/r?dbs+toxline:@term+@rn+80-05-7>
- EPA SRS http://iaspub.epa.gov/sor_internet/registry/substreg/searchandretrieve/advancedsearch/externalSearch.do?p_type=CASNO&p_value=80-05-7
- Some links pass InChIKeys for searching
- Some links pass SMILES for searching/prediction
- Some links pass a list of rank-ordered identifiers for searching (think Google Books, Pubmed, Google Scholar)

Requires linked website be “chemically indexed”

Creating Linkages to Data Sources ...



Creating linkages to indexed content is the easy part,
... creating accurate chemical-data listings is hard!

Diverse quality in databases



- Challenges in assembling a database
 - Sourcing high quality data – sorting wheat from chaff
 - How to mesh data together – based on structure? On name? On CAS number? Other identifier?
 - Checking self-consistency of data?
 - Structure validation versus property value validation – very different challenges

Existing Online Property Prediction Platforms



- ACD/Interactive Laboratory (Ilab)
- Igor Tetko's OCHEM
- The OECD Toolbox
- etc...

Learn from others' examples,
... make open, transparent, reusable,
... tailor to EPA's needs!

iCSS Chemistry Dashboard

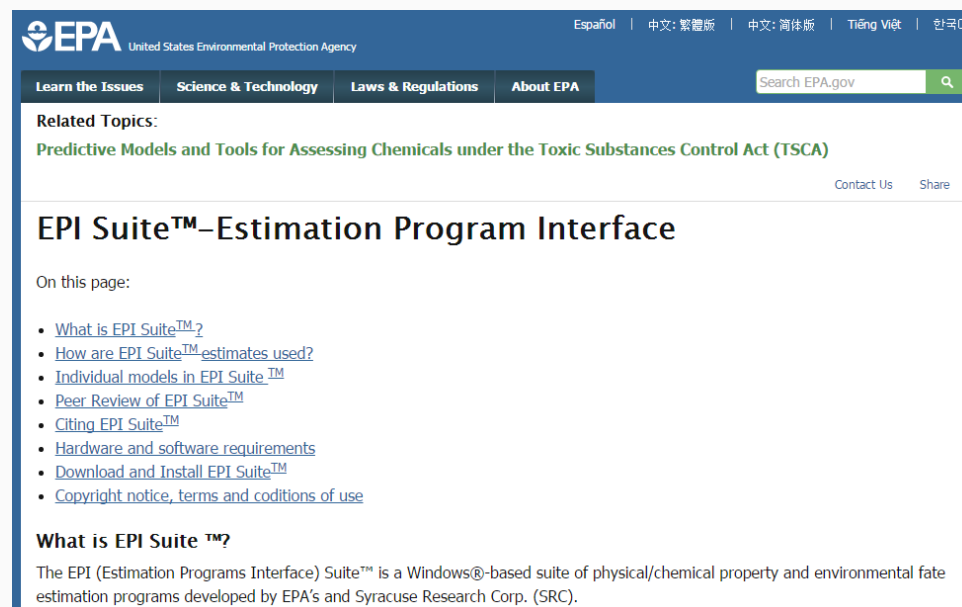


- Deliver web-based platform hosting chemical property data of interest to “environmental toxicology & exposure” scientists (powered by DSSTox) 
- Provide access to experimental data sets 
 - e.g. EPISuite’s PHYSPROP: logP, BP, MP, WSol, etc.

but...

“data quality” is a major concern & challenge

a LOT of time was spent cleaning & curating data!



The screenshot shows the EPA website's "Science & Technology" section. The main heading is "EPI Suite™—Estimation Program Interface". Below this, a list of links provides information about the suite, including its purpose, usage, models, peer reviews, citation guidelines, hardware requirements, installation instructions, and copyright notice. The page also includes a search bar, navigation tabs, and a footer with contact information.

EPA United States Environmental Protection Agency

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

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

Search EPA.gov

Related Topics:
Predictive Models and Tools for Assessing Chemicals under the Toxic Substances Control Act (TSCA)

Contact Us | Share

EPI Suite™—Estimation Program Interface

On this page:

- [What is EPI Suite™?](#)
- [How are EPI Suite™ estimates used?](#)
- [Individual models in EPI Suite™](#)
- [Peer Review of EPI Suite™](#)
- [Citing EPI Suite™](#)
- [Hardware and software requirements](#)
- [Download and Install EPI Suite™](#)
- [Copyright notice, terms and conditions of use](#)

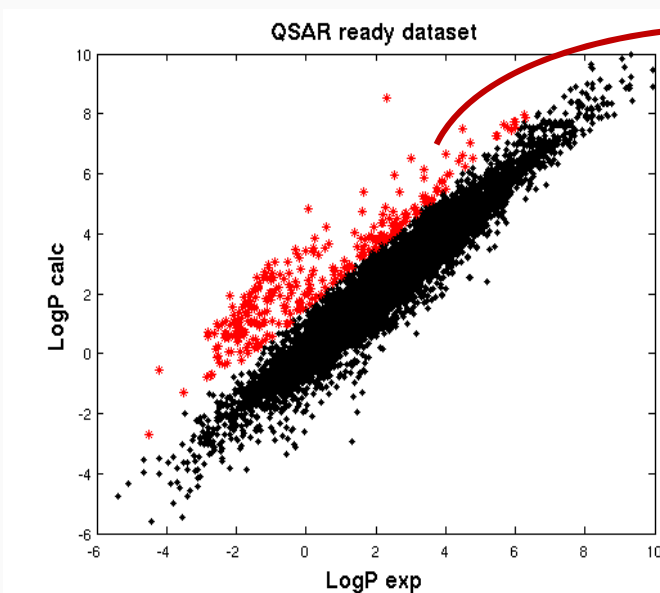
What is EPI Suite™?

The EPI (Estimation Programs Interface) Suite™ is a Windows®-based suite of physical/chemical property and environmental fate estimation programs developed by EPA's and Syracuse Research Corp. (SRC).

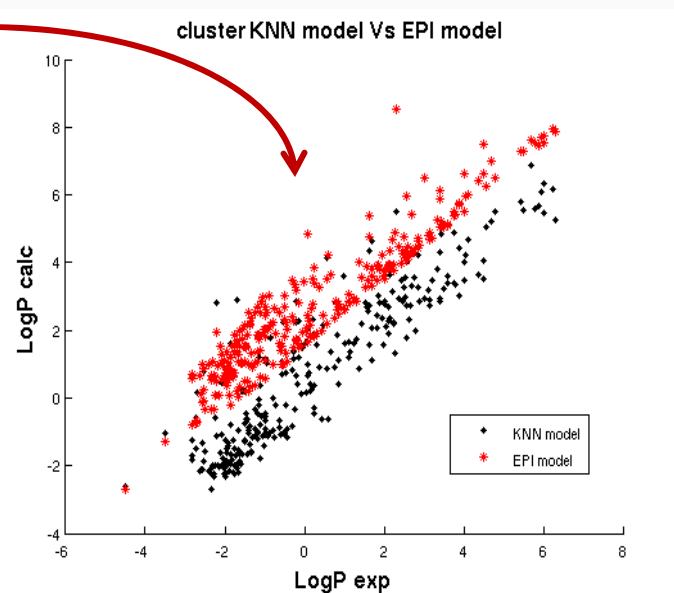
Global vs Local Models

Applicability Domain (AD) of original EPI Suite LogP Model
(accuracy of the global model is an issue)

EPI Suite Predicted vs. Experimental



EPI Suite vs kNN Comparison of Cluster



* *280 cmpd cluster outside AD of EPI Suite* → Compared to new model predictions

Building “NCCT Models”



- Collect “Open Data” for various endpoints of interest to EPA, from e.g.
 - *PubChem*
 - *eChemPortal*
 - *Open Data Sources*
- & Build QSAR prediction models
 - ✓ Using modern machine-learning approaches
 - ✓ Ensuring high quality data as inputs
 - ✓ Populating database with measured & predicted values
 - ✓ Making experimental data training sets available
 - ✓ Allowing real-time predictions (*in the future*)

Open Data Set - example



- 200,000 melting points and 13,000 pyrolysis data points

Research article

Open Access

The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS

Igor V. Tetko^{1,2*}, Daniel M. Lowe³ and Antony J. Williams⁴

[Journal of Cheminformatics](#)

December 2016, 8:2

- Modeling this much data is a challenge!
- Accessibility to data?
 - Figshare.com
 - DataDryad.com
 - Institutional Repositories
 - Publishers?

The screenshot shows the Figshare website interface. At the top, there is a navigation bar with the Figshare logo, a search bar, and user options. Below the navigation bar, there are tabs for 'My data', 'Projects', and 'Activity'. A table of datasets is displayed, with the first row highlighted in red. The highlighted row is titled 'Melting Point & Pyrolysis Point Dataset' and contains the following information:

	STATUS	TYPE	CREATED	SIZE
Melting Point and Pyrolysis Point Data for Tens of Thousands of Chemicals	●	FILESET	9.12.2015 13:35	645.36 MB
Serving the medicinal chemistry community with Royal Society of Chemistry cheminformatics platforms	●	PRESENTATION	1.5.2015 23:46	10.47 MB
The Benefits of Participation in the Social Web of Science	●	PRESENTATION	1.5.2015 23:37	6.58 MB
Accessing Royal Society of Chemistry resources and making chemistry mobile	●	PRESENTATION	22.3.2014 04:55	13.46 MB

Integrating other EPA predictors



- ExpoCast
 - near & far-field exposure models
- Environmental Fate Simulator (EFS)
 - air/soil/water distribution, biotransformation
- CERAPP
 - estrogen receptor activity QSAR model
- T.E.S.T (*in progress*)
 - phys-chem properties & toxicity endpoints

e.g., T.E.S.T.



<http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test>



United States Environmental Protection Agency

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

Learn the Issues

Science & Technology

Laws & Regulations

About EPA

Search EPA.gov



Related Topics: [Safer Chemicals Research](#)

[Contact Us](#)

[Share](#)

Toxicity Estimation Software Tool (TEST)

On this page:

- [QSAR Methodologies](#)
- [What's New in Version 4.1?](#)
- [Prior Version History](#)
- [System Requirements](#)
- [Installation Instructions](#)
- [Publications](#)
- [Get Email Alerts](#)

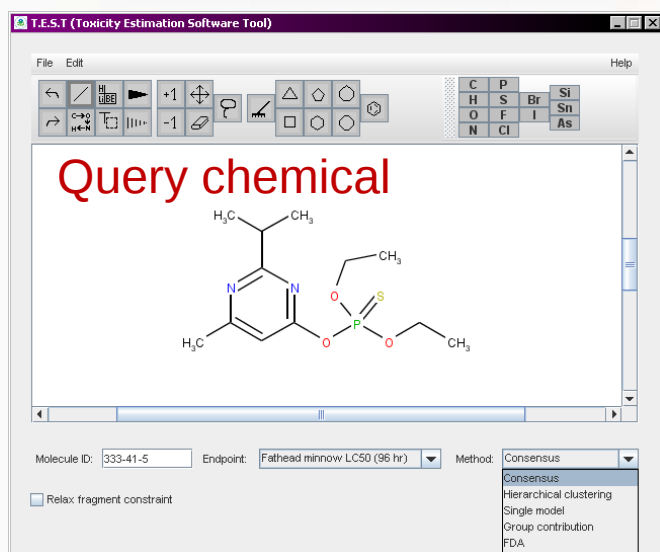
- Downloadable Windows desktop app
- Several phys-chem and tox endpoints predicted
- Multiple QSAR modeling methods employed
- Multiple views of data

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

Option

Fathead minnow LC50 (96 hr)
Daphnia magna LC50 (48 hr)
T. pyriformis IGC50 (48 hr)
Oral rat LD50
Bioaccumulation factor
Developmental Toxicity
Mutagenicity
Normal boiling point
Vapor pressure at 25°C
Melting point
Flash point
Density
Surface tension at 25°C
Thermal conductivity at 25°C
Viscosity at 25°C
Water solubility at 25°C
Molecular Descriptors

T.E.S.T. QSAR Model Prediction Views



Model # 1296

Toxicity prediction results for 333-41-5 for Hierarchical clustering method

Prediction results			
Endpoint	Experimental value CAS: 333-41-5 Source: ECOTOX	Predicted value ^a	Prediction interval
Fathead minnow LC ₅₀ (96 hr) -Log(mol/L)	4.81	5.39	4.54 ≤ Tox ≤ 6.24
Fathead minnow LC ₅₀ (96 hr) mg/L	4.70	1.23	0.17 ≤ Tox ≤ 8.71

^aNote: the test chemical was present in the external test set.

Prediction results

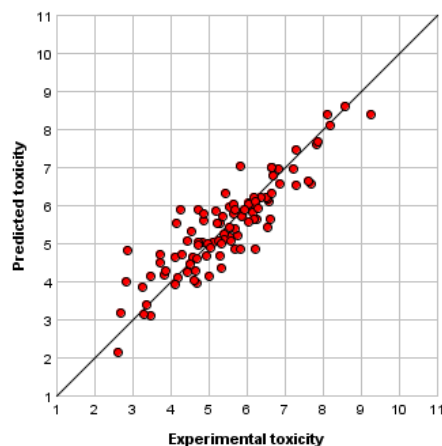
Cluster model predictions and statistics

Cluster model	Test chemical descriptor values	Prediction interval -Log(mol/L)	r ²	q ²	#chemicals
1296	Descriptors	6.010 ± 1.136	0.793	0.733	101
1300	Descriptors	5.458 ± 1.312	0.729	0.645	111
1301	Descriptors	5.136 ± 1.169	0.747	0.718	294
1302	Descriptors	4.922 ± 1.182	0.774	0.751	641

Model statistics

Parameter	Value
Endpoint	Fathead minnow LC ₅₀ (96 hr)
r ²	0.793
q ²	0.733
#chemicals	101
Model	Model # 1296

Model fit results



Cluster models with violated constraints

Cluster Model	r ²	q ²	# chemicals	Message
1121	0.810	0.576	10	Rmax constraint not met
1209	0.799	0.574	11	Fragment constraint not met
1247	0.919	0.647	20	Fragment constraint not met
1264	0.869	0.781	22	Fragment constraint not met
1268	0.675	0.553	24	Fragment constraint not met

[Descriptor values for test chemical](#)

A Platform for Predictions



- Developing a platform for hosting prediction models
- Coordinating efforts with other EPA prediction platforms and programs

The collage features three overlapping screenshots of EPA web pages:

- Top Left:** "Exposure Assessment Tools and Models" page. It includes the EPA logo, a search bar with "All EPA" and "This Area" radio buttons, and a breadcrumb trail: "You are here: EPA Home » Chemical Safety and Pollution Prevention » Prevention, Pesticides & Toxicology". Below the header is the "Estimation Program Interface (EPI) Suite".
- Top Right:** "Risk Management Sustainable Technology" page. It features the EPA logo, navigation links, and a breadcrumb trail: "You are here: EPA Home » Research » Risk Management Research » Sustainable Technology » Quantitative Structure Activity Relationship". The main heading is "Quantitative Structure Activity Relationship" with an "Introduction" link.
- Bottom:** "Ecosystems Research" page. It includes the EPA logo, navigation links, and a breadcrumb trail: "You are here: EPA Home » Exposure Research » Ecosystems Research » Environmental Fate Simulator". The main heading is "Research in Action" with a sub-heading "Environmental Fate Simulator: Forecasting how chemicals move in the environment".

A small image of a green frog is visible in the bottom left corner of the collage.

Conclusions



- iCSS Chemistry Dashboard will provide public access to data & services focused on chemicals of interest to EPA & environmental tox communities
- Initially serve up results for measured & pre-predicted data, *on-the-fly predictions in the future*
- Data quality impacts models - *public domain chemical-data linkages require curation/validation*
- All data and models will be available as **OPEN DATA and OPEN CODE**

Stay tuned!!

Acknowledgements



NCCT Contributors to the iCSS Chemistry Dashboard

- Jeff Edwards
- Jeremy Fitzpatrick
- Jordan Foster
- Jason Harris
- Dave Lyons
- Kamel Mansouri
- Aria Smith
- Jennifer Smith
- Indira Thillainadarajah
- Chris Grulke
- Antony Williams



Want to learn more?



1. Expansion of DSSTox: Leveraging public data to create a semantic cheminformatics resource with quality annotations for support of U.S. EPA applications (CINF 130)

Division of Chemical Information PAPER ID: 2385741

SESSION: Chemistry, Data & the Semantic Web: An Important Triple to Advance Science, 8:15 AM - 11:55 AM

Wednesday, March, 16, 2016 from 10:40 AM - 11:05 AM

Room 25B - San Diego Convention Center

2. Influence of data curation on QSAR Modeling – examining issues of quality versus quantity of data (MPPG 124)

Multidisciplinary Program Planning Group PAPER ID: 2394881

SESSION: Big Data Science, 1:30 PM - 4:30 PM

Thursday, March, 17, 2016 from 1:30 PM - 2:00 PM

ROOM & LOCATION: Room 3 - San Diego Convention Center