

OPERA models to support regulatory purposes

Kamel Mansouri
Integrated Laboratory Systems

Disclaimer: ILS staff provide technical support for NICEATM, but do not represent NIEHS, NTP, or the official positions of any federal agency.



QSARs for regulatory purposes

The 5 OECD Principles

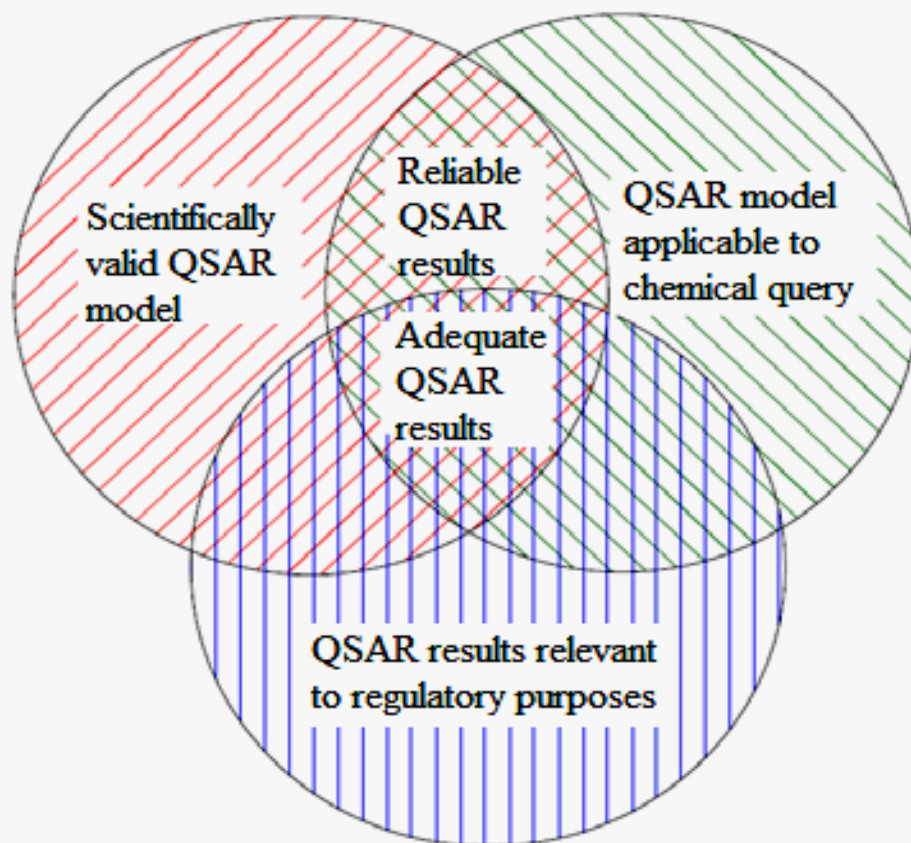
1) A defined endpoint

2) An unambiguous algorithm

3) A defined domain of applicability

4) Appropriate measures of goodness-of-fit, robustness and predictivity

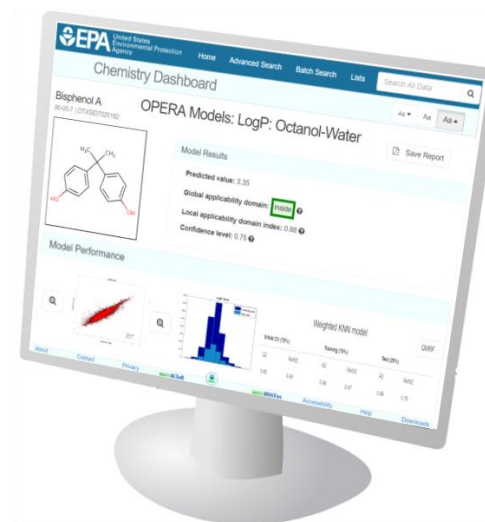
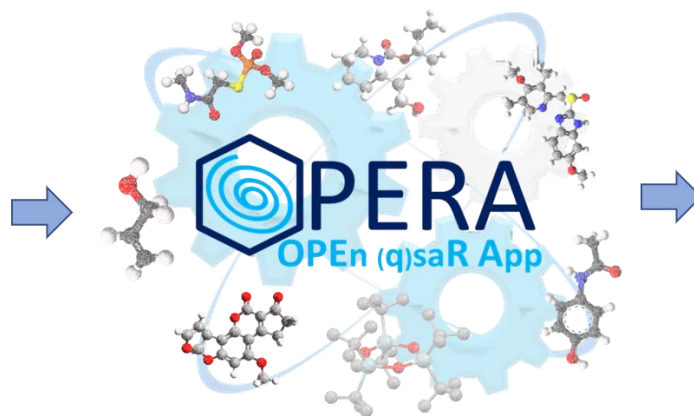
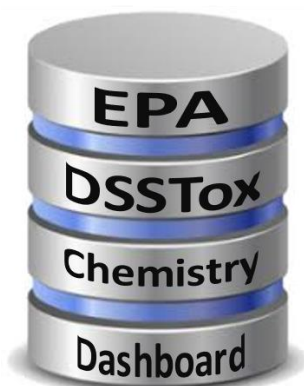
5) Mechanistic interpretation, if possible





OPERA approach

- Curated **open** access datasets (<https://doi.org/10.1186/s13321-018-0263-1>)
- **Open-source** code (github.com/NIEHS/OPERA)
- **Transparent** unambiguous algorithms (<https://qsardb.jrc.ec.europa.eu/qmrf/>)
- **Transparent** validated performances (<https://doi.org/10.1080/1062936X.2016.1253611>)
- **Defined** applicability domain and limitations of the models
- Predictions **available** through:
 - The EPA's CompTox Dashboard (<https://comptox.epa.gov/dashboard>)
 - Free and open-source standalone application (github.com/NIEHS/OPERA)





OPERA modeling steps and considerations

Step	Description
Curation of the data	Flagged and curated files available for sharing
Preparation of training and test sets	Inserted as a field in SDF files and csv data files
Calculation of an initial set of descriptors	PaDEL & CDK 2D descriptors and fingerprints
Selection of a mathematical method	Several approaches tested: KNN, PLS, SVM...
Variable selection technique	Genetic algorithm
Validation of the model's predictive ability	5-fold cross validation and external test set
Define the Applicability Domain	Local (nearest neighbors) and global (leverage) approaches



[Journal of Cheminformatics](#)
December 2018, 10:10 | [Cite as](#)

OPERA models for predicting physicochemical properties and environmental fate endpoints





OPERA models

Example of public data

- **PHYSPROP** <http://esc.syrres.com/interkow/EpiSuiteData.htm>

EPI Suite Data

The downloaded files are provided in "zip" format ... the downloaded file must be "un-zipped" with common utility programs such as [WinZip](#).

Basic Instructions:

- (1) Download the zip file
- (2) Un-Zip the file

WSKOWWIN Program Methodology & Validation Documents (includes Training & Validation datasets) - Download file is: WSKOWWIN_Datasets.zip (180 KB)

[Click here to download WSKOWWIN_Datasets.zip](#)

WATERNT (Water Solubility Fragment) Program Methodology & Validation Documents (includes Training & Validation datasets) - Download file is: WaterFragmentDataFiles.zip (511 KB)

[Click here to download WaterFragmentDataFiles.zip](#)

MPBPWIN (Melting Pt, Boiling Pt, Vapor Pressure) Program Test Sets - Download file is: MP-BP-VP-TestSets.zip (1983 KB)

[Click here to download MP-BP-VP-TestSets.zip](#)

BCFBAF Excel spreadsheets of BCF and kM data used in training & validation ... (includes the Jon Arnot Source BCF DB with multiple BCF values) - Download file is: Data_for_BCFBAF.zip (1.4 MB)

[Click here to download Data_for_BCFBAF.zip](#)

HENRYWIN Data files used in training & validation ... (includes Meylan and Howard (1991) Data document) - Download file is: HENRYWIN_Data_EPI.zip (531 K)

[Click here to download HENRYWIN_Data_EPI.zip](#)

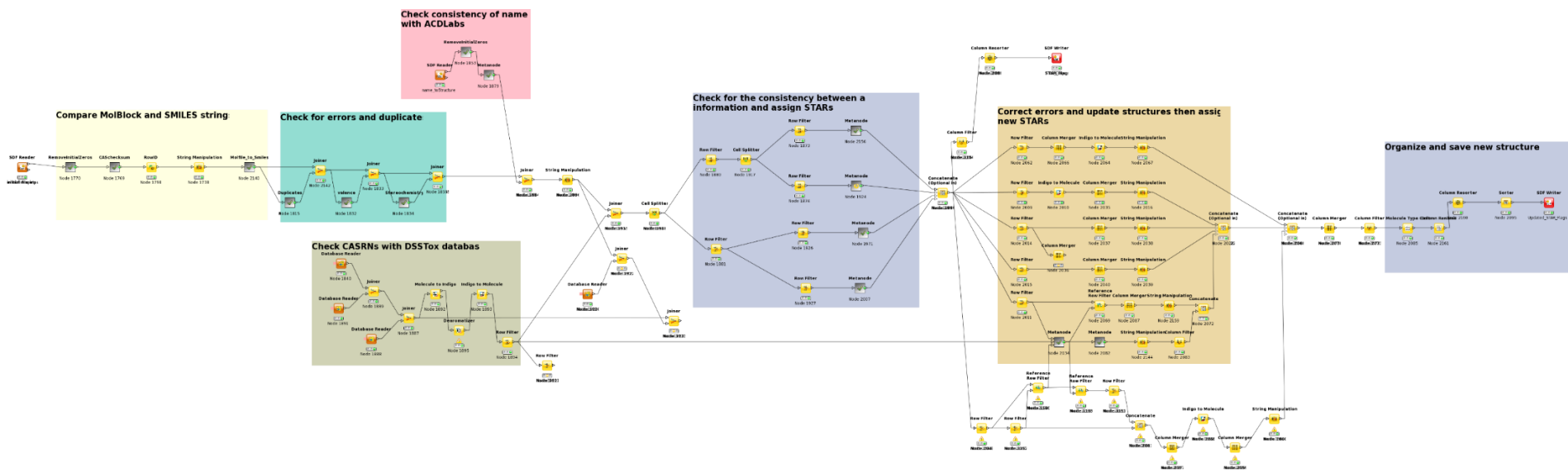
SDP Molecule	Mol Mol Block		
<pre>-ISIS- 09141018452D 4 3 0 0 0 0 0 0 0 0999 V2000 2.4667 -0.0833 0.0000 O 0 0 ... 2.4667 -0.9125 0.0000 C 0 0 ... 1.7500 -1.3292 0.0000 H 0 0 ... 3.1833 -1.3292 0.0000 H 0 0 ... 2 1 2 0 0 0 0 3 2 1 0 0 0 0 4 2 1 0 0 0 0 M END > <CAS> (000050-00-0) 000050-00-0 > <NAME> (000050-00-0) FORMALDEHYDE > <Kow> (000050-00-0) 3.5000000000000000e-001</pre>			
S Smiles	S CAS	S NAME	D Kow
O=C	000050-00-0	FORMALDEHYDE	0.35

- Molblock
- SMILES
- Name
- CASRN

The data files have **FOUR** representations of a chemical, plus the property value.



KNIME Workflow to Evaluate the Data



Quality FLAGS and curated structures



SAR and QSAR in Environmental Research

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling

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



Examples of Errors

Valence Errors

Mol Block	CAS	NAME	Smiles
	000274-87-3	TETRAZOLO[1,5-A]PYRIDINE	<chem>C1=CN2C=NC=N2N1</chem>
	000542-85-8	ETHYL ISOTHIOCYANATE	<chem>CCN=C=S</chem>
	000707-98-2	9-PROPYL ADENINE	<chem>CCCN1=NC=NC2=C1N=CN2</chem>
	000715-48-0	6-METHYL-4-NITROQUINOLINE-1-OXIDE	<chem>Cc1ccc2c(c1)[n+]([O-])c3ccc(O2)=O</chem>

Mismatching structures

Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	<chem>CC12CCC3C(C1)C(=O)CC4C(C3)C(O)C(C4)C</chem>
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	<chem>CC(CO)C(O)CO</chem>
	000079-60-7	CORTISONE-9A-FLUORO	<chem>CC12CCC3C(C1)C(=O)CC4C(C3)C(O)C(C4)C(F)C</chem>
	000082-38-2	DISPERSE RED 9	<chem>Cc1ccc2c(c1)c3ccccc3C(=O)c2=O</chem>

Duplicate Structures

Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	C ₃ H ₆ O ₃	90.0779	000050-21-5	LACTIC ACID	1.6800000000000000e+001	2.2860000000000000e+001	5.9600000000000000e+000
	C ₃ H ₆ O ₃	90.0779	000079-33-4	L-LACTIC ACID	5.3000000000000000e+001	2.2860000000000000e+001	-3.0340000000000000e+001
	C ₃ H ₆ O ₃	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.8000000000000000e+001	2.2860000000000000e+001	4.9600000000000000e+000
	C ₃ H ₆ O ₃	90.0779	010326-41-7	D-LACTIC ACID	5.2000000000000000e+001	2.2860000000000000e+001	-3.0140000000000000e+001

Covalent Halogens

Mol Block	CAS	NAME	Smiles
	000056-93-9	BENZYL TRIMETHYL AMMONIUM CHLORIDE	<chem>C[N+](C)(C)Cc1ccccc1.[Cl-]</chem>
	000068-05-3	TETRAETHYL AMMONIUM IODIDE	<chem>CC[N+](C)(C)CC.[I-]</chem>
	000071-91-0	TETRAETHYL AMMONIUM BROMIDE	<chem>CC[N+](C)(C)CC.[Br-]</chem>



LogP dataset: 15,809 structures

- CAS Checksum: 12163 valid, 3646 invalid (>23%)
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES (>24%)
- Duplicates check:
 - 31 DUPLICATE MOLFILES
 - 626 DUPLICATE SMILES
 - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
 - 1279 differ in stereochemistry (~8%)
 - 362 “Covalent Halogens”
 - 191 differ as tautomers
 - 436 are different compounds (~3%)



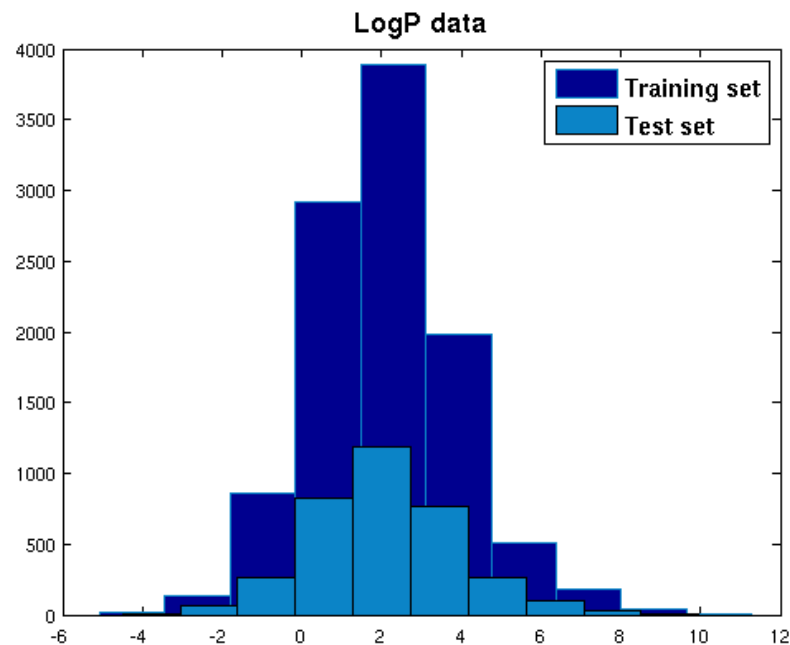
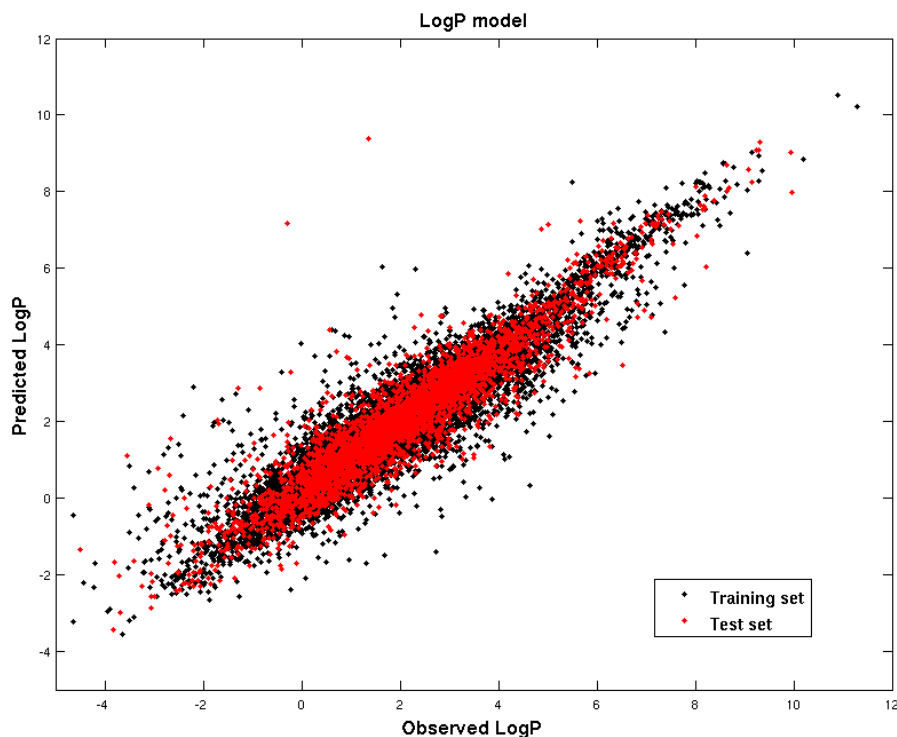
Curation to QSAR Ready Files

Property	Initial file	Curated Data	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	14544	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	2348	2046	2010

Mansouri et al. OPERA models. (<https://link.springer.com/article/10.1186/s13321-018-0263-1>)



LogP Model: weighted kNN



Weighted 5-nearest neighbors

9 Descriptors

Training set: 10531 chemicals

Test set: 3510 chemicals

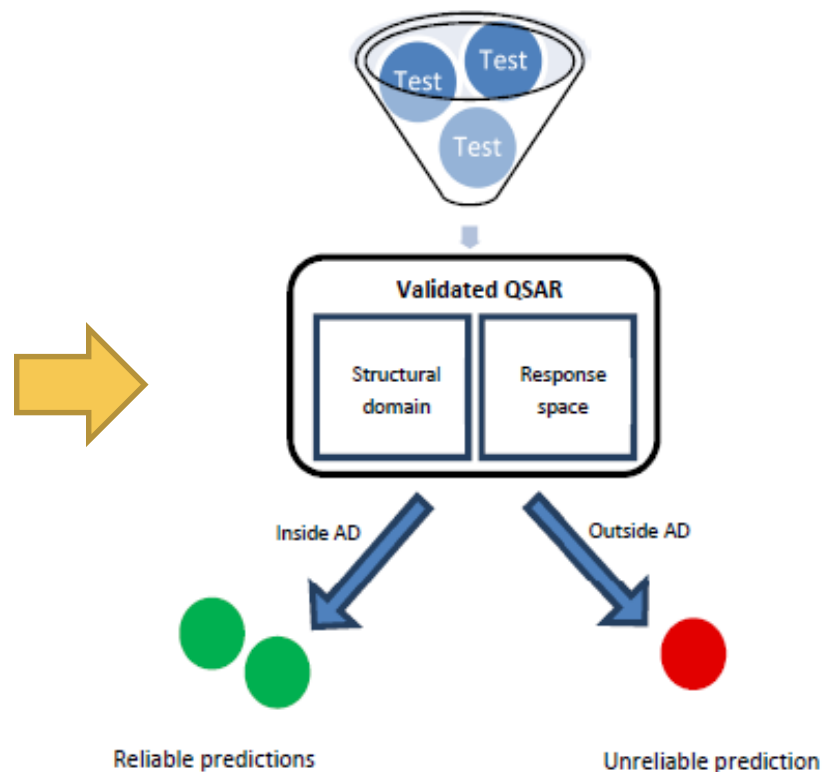
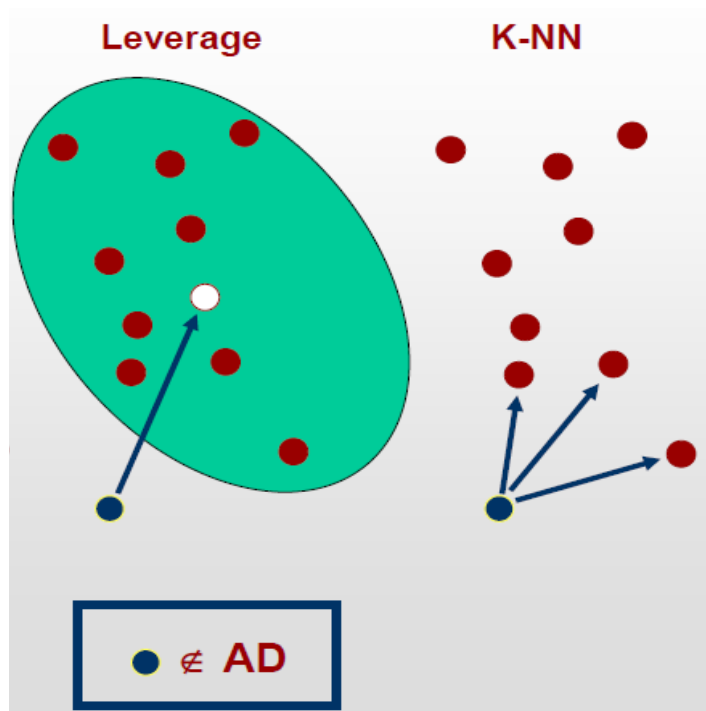
5 fold CV: Q2=0.85, RMSE=0.69

Fitting: R2=0.86, RMSE=0.67

Test: R2=0.86, RMSE=0.78



Chemical space and AD definition



Descriptor space based the response domain:

- Global applicability domain (leverage)
- Local applicability domain (kNN)
- Accuracy estimate based on the 5NN

Reliable predictions
for structurally similar
chemicals.



OPERA Standalone application:

Command line

```
OPERA
-----
OPERA models for physchem and environmental fate properties.
Version 1.5 (September 2017)

OPERA is a command line application developed in Matlab providing QSAR
models predictions as well as applicability domain and accuracy assessment.

Developed by:
Kamel Mansouri
mansourikamel@gmail.com

Developed at:
National Center of Computational Toxicology
United States Environmental Protection Agency

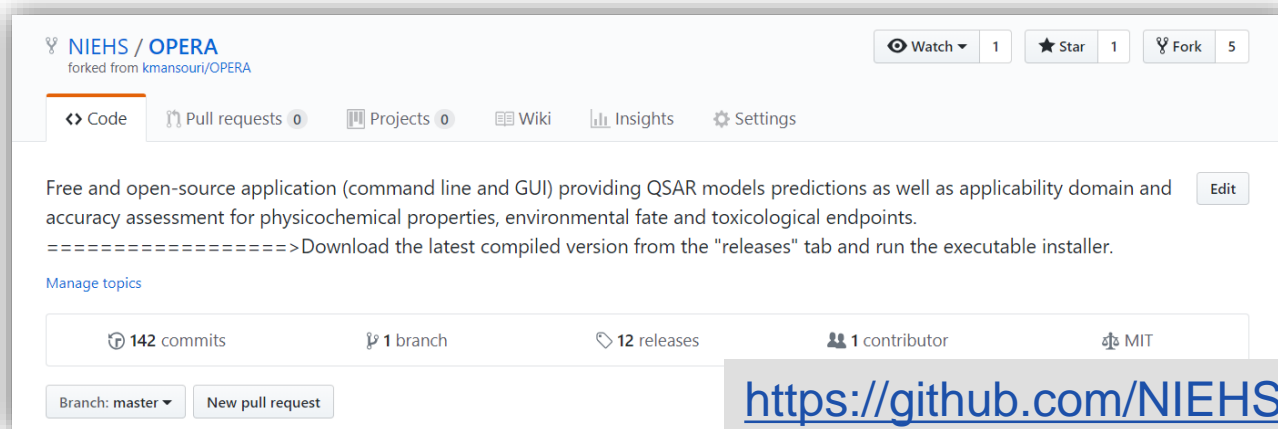
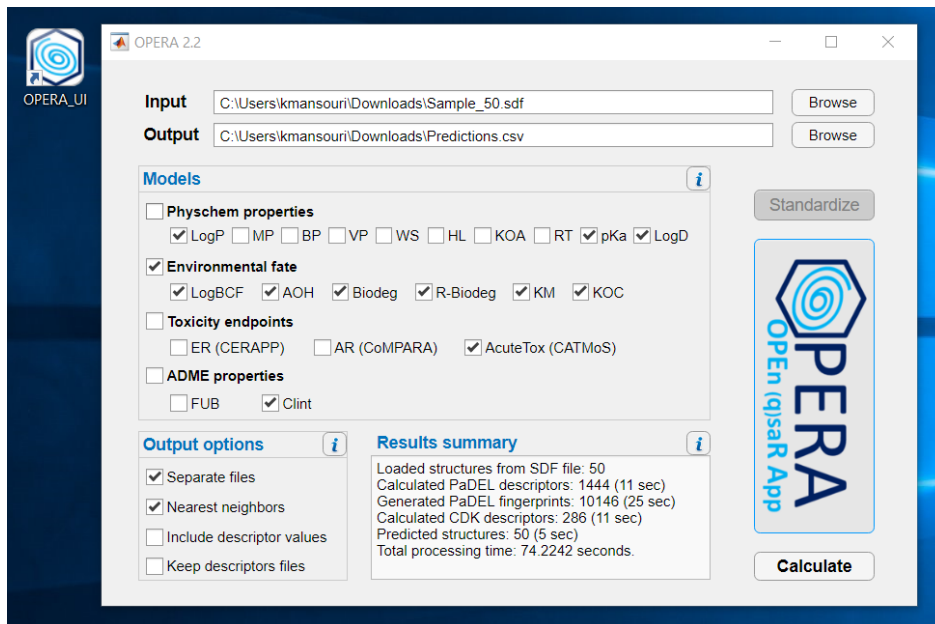
Usage: OPERA <argument_list>

Examples:
OPERA -s Sample_50.sdf -o predictions.csv -a -x -v 2
opera -d Sample_50.csv -o predictions.txt -e logP BCF -n -v

Type OPERA -h or OPERA --help for more info.
```



Graphical User Interface



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



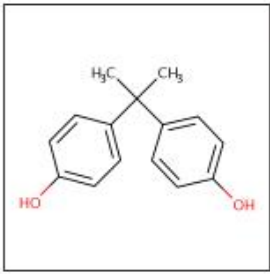
OPERA on the EPA Dashboard

Prediction report

United States Environmental Protection Agency | Home | Advanced Search | Batch Search | Lists | Predictions | Downloads | 20182

Chemistry Dashboard | Bisphenol A | OPERA Models: LogP: Octanol-Water | Save PDF

80-05-7 | DTXSID7020182



Model Results

Predicted value: 3.35


Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.75

Calculation Result for a chemical

Model Performance



Weighted KNN model

6-fold CV (75%) | Training (75%) | Test (25%)

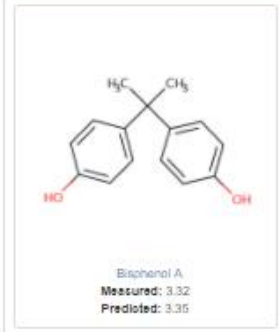
6-fold CV (75%)		Training (75%)		Test (25%)	
O2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.85	0.67	0.85	0.78

QMRP

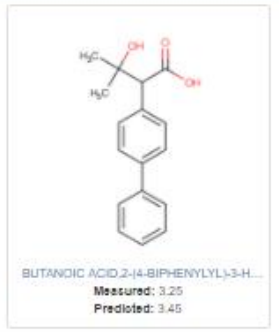
Prediction, AD and accuracy estimates

Model Performance with full QMRF

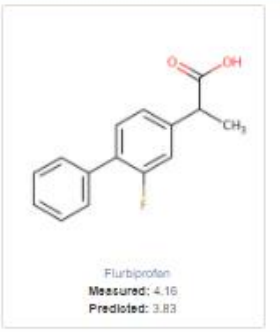
Nearest Neighbors from the Training Set



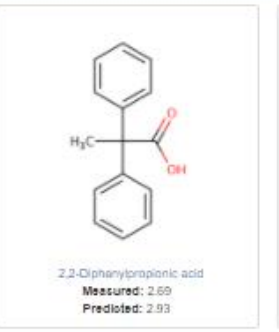
Bisphenol A
Measured: 3.32
Predicted: 3.35



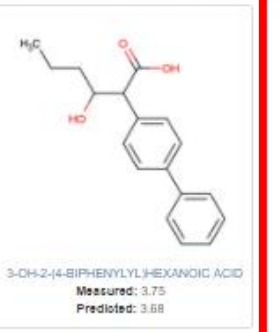
BUTANOIC ACID, 2-(4-BIPHENYLYL)-3-HYDROXY-2-METHYL-
Measured: 3.25
Predicted: 3.45



Flurbiprofen
Measured: 4.16
Predicted: 3.83



2,2-Diphenylpropionic acid
Measured: 2.69
Predicted: 2.93



3-OH-2-[4-BIPHENYLYL]-HEXANOIC ACID
Measured: 3.75
Predicted: 3.68

Nearest Neighbors from Training Set

Discover: About/Disclaimer | Accessibility | Privacy

Connect: ACToR | DSSTox | Downloads

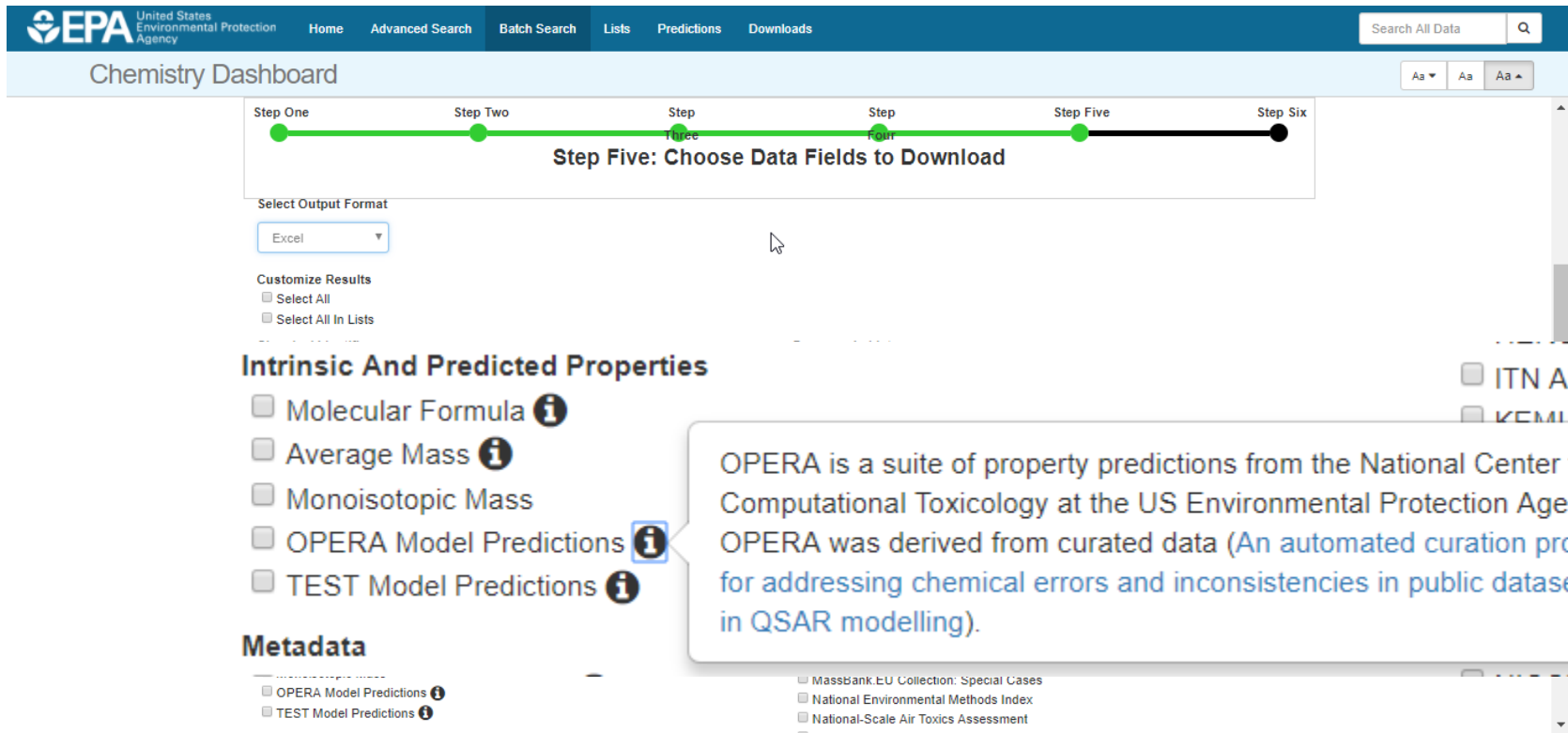
Ask: Contact | Help





OPERA on the EPA Dashboard

Batch download of predictions



United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads

Search All Data

Chemistry Dashboard

Step One Step Two Step Three Step Four Step Five Step Six

Step Five: Choose Data Fields to Download

Select Output Format

Excel

Customize Results

- Select All
- Select All In Lists

Intrinsic And Predicted Properties

- Molecular Formula **i**
- Average Mass **i**
- Monoisotopic Mass
- OPERA Model Predictions **i**
- TEST Model Predictions **i**

Metadata

- OPERA Model Predictions **i**
- TEST Model Predictions **i**
- Massbank.EU Collection: Special Cases
- National Environmental Methods Index
- National-Scale Air Toxics Assessment

ITN ANTIBIOTIC L

KEMList of Subst

er

ill

ne

r

le

OPERA is a suite of property predictions from the National Center for Computational Toxicology at the US Environmental Protection Agency. OPERA was derived from curated data (An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling).



OPERA QMRF Reports



JOINT RESEARCH CENTRE

The European Commission's science and knowledge service

Legal Notice | Cookies | Contact | Search | English (en) ▾



European Commission > EU Science Hub > EURL ECVAM > QSARDB > QMRF documents



Welcome

QMRF documents

Structures

Endpoints

Get QMRF Editor

User support

QMRF document search

- Title ?
- Free text ?
- Free text (boolean) ?
- Endpoint ?
- Author ?
- QMRF number ?

Max results

Display QMRF documents.

Search:

	QMRF Number	Title	Endpoint	Last updated	Download
	Q17-13-0012	OPERA-model for Water solubility	1.3.Water solubility	Sep 21 2017	
	Q17-14-0013	OPERA-model for Vapor pressure	1.4 Vapour pressure	Sep 21 2017	
	Q17-23a-0014	OPERA-model for Readil biodegradability			
	Q17-11-0015	OPERA-model for Melting point			
	Q17-16-0016	OPERA-model for Octanol-water partition coefficient			
	Q17-26-0017	OPERA-model for organic carbon sorption coefficient			
	Q17-18-0018	OPERA-model for octanol/air partition coefficient			
	Q17-66-0019	OPERA-model for biotransformation rat constant			
	Q17-19-0020	OPERA-model for Henry's Law constant			
	Q17-12-0021	OPERA-model for Boiling point			



QMRF identifier (JRC Inventory):Q17-18-0018

QMRF Title:OPERA-model for octanol/air partition coefficient

Printing Date:Oct 17, 2017

1.QSAR identifier

1.1.QSAR identifier (title):

OPERA-model for octanol/air 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.

Kamel Mansouri (mansourikamel@gmail.com)

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

<https://qsardb.jrc.ec.europa.eu/qmrf>



OPERA Standalone application:

OPERA v1.5: Physchem & Env. fate



Model	Property
AOH	Atmospheric Hydroxylation Rate
BCF	Bioconcentration Factor
BioHL	Biodegradation Half-life
RB	Ready Biodegradability
BP	Boiling Point
HL	Henry's Law Constant
KM	Fish Biotransformation Half-life
KOA	Octanol/Air Partition Coefficient
LogP	Octanol-water Partition Coefficient
MP	Melting Point
KOC	Soil Adsorption Coefficient
VP	Vapor Pressure
WS	Water solubility
RT	HPLC retention time

New in OPERA v2.2:

- Structural properties:
Hybridization Ratio, nHBAcc, nHBDon, LipinskiRule, Topo PSA, Molar refractivity, Polarizability, electronegativity...
- pKa
- Log D
- ER activity (CERAPP)
 - Agonist
 - Antagonist
 - Binding

(<https://ehp.niehs.nih.gov/15-10267/>)
- AR activity (CoMPARA)
 - Agonist
 - Antagonist
 - Binding

(<https://doi.org/10.13140/RG.2.2.19612.80009>,
<https://doi.org/10.13140/RG.2.2.21850.03520>)
- Acute toxicity (CATMoS)
 - NT
 - VT
 - EPA categories
 - GHS categories
 - LD50

(<https://doi.org/10.1016/j.comtox.2018.08.002>)
- ADME
 - FUB
 - Clint

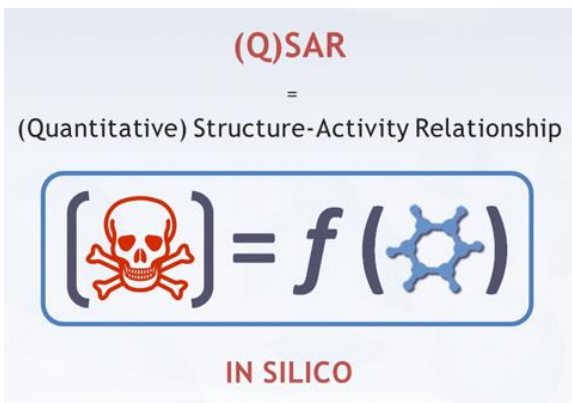
Models versioned separately from the tool



Toxicity prediction

Too many chemicals to test with standard animal-based methods
– Cost, time, animal welfare

Alternative



- Organic **pollutants** with exposure potential **accumulate** in body tissues
 - Cause **toxic effects** to wild life and humans
- Existence of **gaps in the experimental data** for environmental endpoints
 - Need to fill the data gaps and bridge the **lack of knowledge**
- **Regulatory** requirements:
 - Reduce **animal** testing, **time** and **costs**
 - **Methodology**: use of **QSAR/QSPR** to **predict** the **endpoints** of interest.



International collaborative projects

CERAPP

Collaborative Estrogen Receptor
Activity Prediction Project (2015/16)

CoMPARA

Collaborative Modeling Project for Androgen
Receptor Activity (2017/18)

CATMoS

Collaborative Acute Toxicity Modeling Suite
(2017/18)



Endocrine Disruptor Screening Program (EDSP)

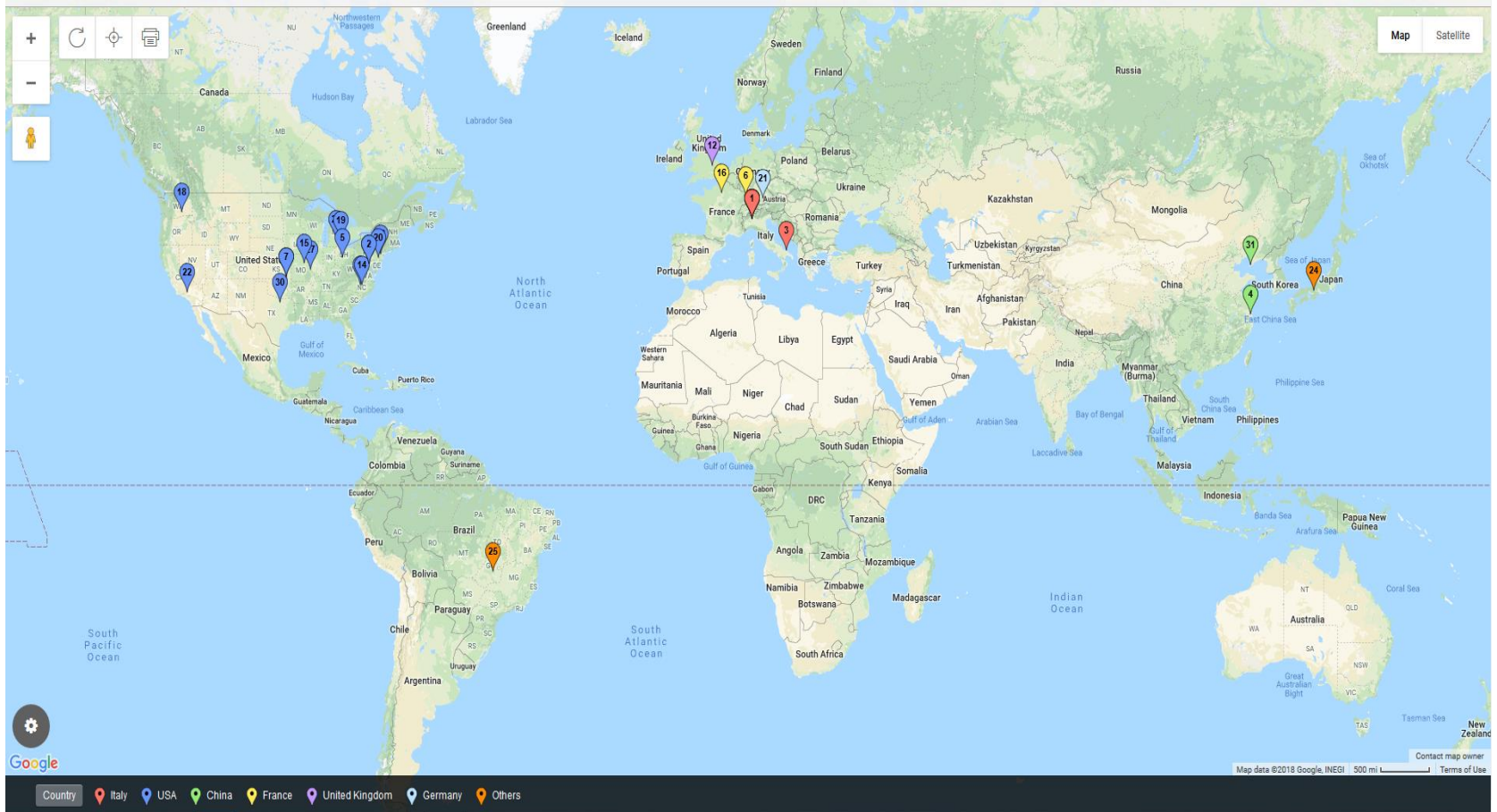


Acute Toxicity Workgroup: alternative methods



International consortium

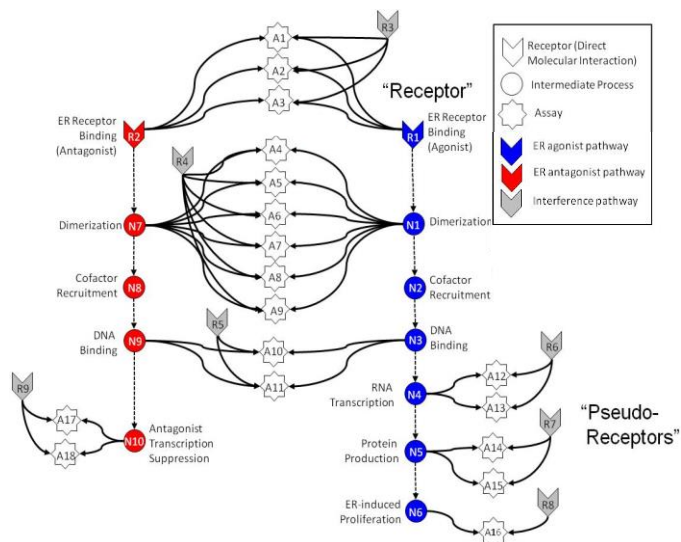
Over 100 collaborators from around the globe representing academia, industry, and government contributed.





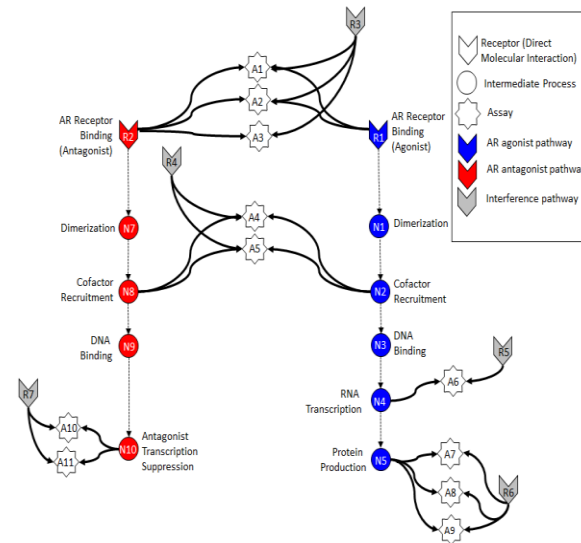
CERAPP (ER) & CoMPARA (AR)

Tox21/ToxCast ER Pathway Model



Judson et al Toxicol. Sci. (2015) 148: 137-154

Tox21/ToxCast AR Pathway Model



Kleinstreuer N. C. et al. 2017 30 (4), 946-964.

CERAPP consensus

	Binding		Agonist		Antagonist	
	Train	Test	Train	Test	Train	Test
Sn	0.93	0.58	0.85	0.94	0.67	0.18
Sp	0.97	0.92	0.98	0.94	0.94	0.90
BA	0.95	0.75	0.92	0.94	0.80	0.54

CoMPARA consensus

	Binding		Agonist		Antagonist	
	Train	Test	Train	Test	Train	Test
Sn	0.99	0.69	0.95	0.74	1.00	0.61
Sp	0.91	0.87	0.98	0.97	0.95	0.87
BA	0.95	0.78	0.97	0.86	0.97	0.74



Acute Oral Toxicity: CATMoS

Endpoints predicted:

Binary models



VT
T
NT

EPA Categories



I
II
III
IV

GHS Categories



OSHA®

I
II
III
IV
NC

LD50 point estimates (mg/kg)

	Very Toxic (32 models)		Non-Toxic (33 models)		EPA (26 models)		GHS (23 models)	
	Train	Eval	Train	Eval	Train	Eval	Train	Eval
Sn	0.87	0.67	0.93	0.70	0.73	0.50	0.63	0.45
Sp	0.94	0.96	0.96	0.88	0.96	0.91	0.91	0.92
BA	0.93	0.81	0.94	0.79	0.83	0.71	0.77	0.68
<i>In vivo</i>	0.81		0.89		0.82		0.79	

	LD50 (25 models)		LD50 values
	Train	Eval	<i>In Vivo</i>
R2	0.84	0.64	0.80
RMSE	0.32	0.51	0.42



Inform Regulatory Decisions



Environ Health Perspect, DOI:10.1289/ehp.1510267

CERAPP: Collaborative Estrogen Receptor Activity Prediction Project



Your Voice in Federal Decision-Making

FIFRA SAP Meeting on Integrated Endocrine Activity and Exposure-based Prioritization and Screening

Docket Folder Summary [View all documents and comments in this Docket](#)

Docket ID: EPA-HQ-OPP-2014-0614 Agency: Environmental Protection Agency (EPA)

Summary: Announcing nomination to consider for Appointment to the FIFRA SAP and requesting comment on individuals available and interested

[View More Docket Details](#)

Primary Documents

Meetings: Federal Insectic

Notice Posted: 11/05/2014

Meetings: Federal Insectic

Notice Posted: 09/16/2014

Supporting Documents

EPA US Environmental Protection Agency

Learn the Issues Science & Technology Laws & Regulations About EPA

Search EPA.gov

Related Topics: Safer Chemicals Research

Safer Chemicals Research Update June 2016

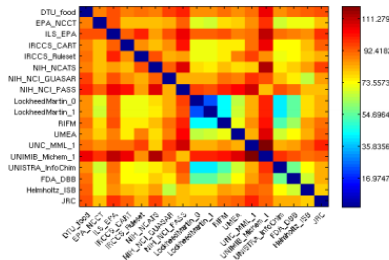
US EPA's Office of Research and Development provides quarterly updates, highlights, events and news about its chemical research. This is the June 2016 edition.

You will need Adobe Reader to view some of the files on this page. See EPA's [About PDF](#) page to learn more.

- [June 2016 CSS Pathways News Anticipating Impacts of Chemicals \(PDF\)](#) (13 pp, 1 MB)

Consensus Modeling: Powering Prediction Through Collaboration

Predictive computational models can efficiently help us prioritize thousands of chemicals for additional testing and evaluation. CSS scientists Kamel Mansouri and Richard Judson, from the U.S. EPA's National Center for Computational Toxicology (NCCT), led a large-scale modeling project called the [Collaborative Estrogen Receptor Activity Prediction Project \(CERAPP\)](#). CERAPP demonstrated the efficacy of using computational models with high-throughput screening (HTS) data to predict potential estrogen receptor (ER) activity of over 32,000 chemicals. This international collaborative effort (17 research groups from the United States and Europe) used both quantitative structure-activity relationship models and docking approaches to evaluate binding, agonist and antagonist activity of chemicals. A total of 48 models were developed. Each model was evaluated and weighed for its predictive accuracy using ToxCast and Tox21 ER HTS results along with data collected from



Correlation matrix of the CERAPP continuous ER models predictions

US Government Information

One stop source for US Government Information

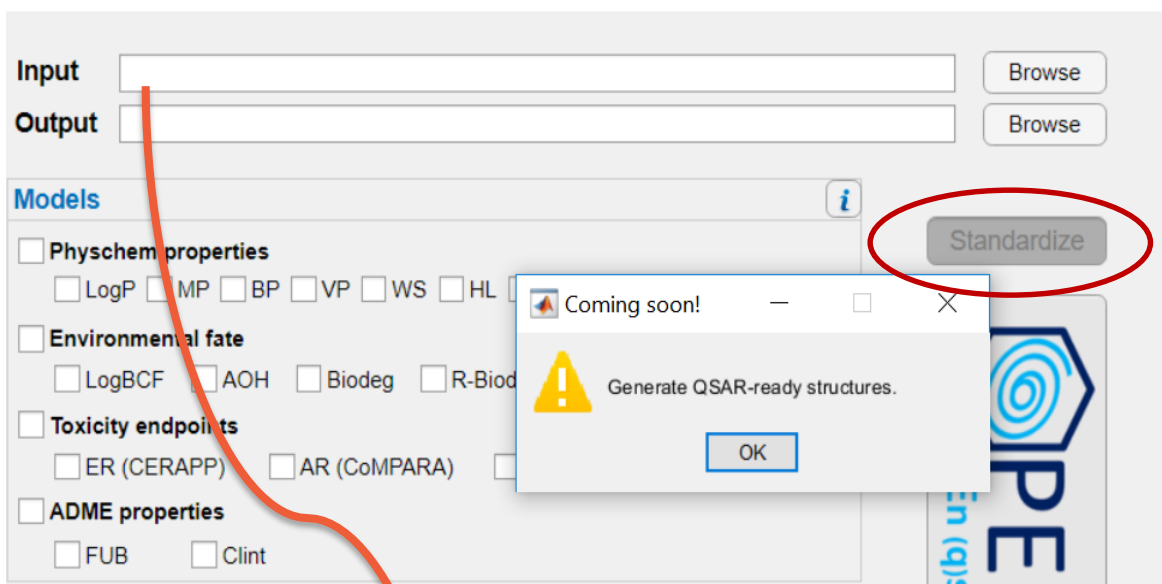
- HOME
- CONSUMER
- DEFENSE & INTERNATIONAL RELATIONS
- EDUCATION & EMPLOYMENT
- FAMILY, HOME, & COMMUNITY
- HEALTH
- MONEY
- PUBLIC SAFETY & LAW
- REFERENCE &
- SCIENCE & TECHNOLOGY
- ABOUT



EDSP Prioritization: Collaborative Estrogen Receptor Activity Prediction Project (CERAPP) (SOT)



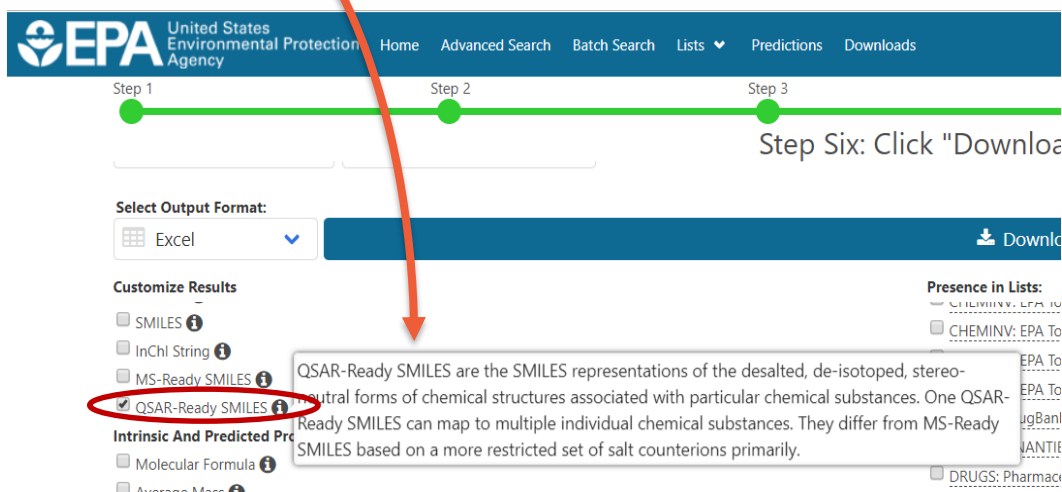
New features to be implemented



1. Integrate the QSAR-ready workflow to process any chemical structure

2. Calculate predictions using ONLY a chemical ID:

- CASRN,
- DTXSID,
- InChiKey



QSAR-ready SMILES from the EPA CompTox Dashboard:

https://comptox.epa.gov/dashboard/dsstoxdb/batch_search



Thank you for your attention!

Acknowledgements

- EPA/NCCT
- NTP/NICEATM
- ILS
- ScitoVation
- Sciome
- ICCVAM Acute Toxicity Workgroup
- All international collaborators

Funding

- EPA/ORD, Oak Ridge Institutes through U.S. DoE & EPA.
- The Lush Prize 2017, young researchers, supporting animal free testing.
- ILS/NICEATM under NIEHS contract HHSN273201500010C