

OPERA: A free and open source QSAR tool for predicting physicochemical properties and environmental fate endpoints

American Chemical Society meeting
New Orleans, LA

March 20, 2018

Kamel Mansouri: ScitoVation LLC

Christopher Grulke, Richard Judson, Antony Williams: NCCT/ US EPA

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

Kamel Mansouri, Ph.D.

Investigator

919-558-1282

kmansouri@scitovation.com

www.scitovation.com



OPERA Models

- Our approach to modeling:
 - Obtain high quality training sets
 - Apply appropriate modeling approaches
 - Validate performance of models
 - Define the applicability domain and limitations of the models
 - Use models to predict properties across our full datasets
- Work has been initiated using available **physicochemical data then extend to additional endpoints**



PHYSPROP Data: Available from:

<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](#)

BCFBFAF 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)

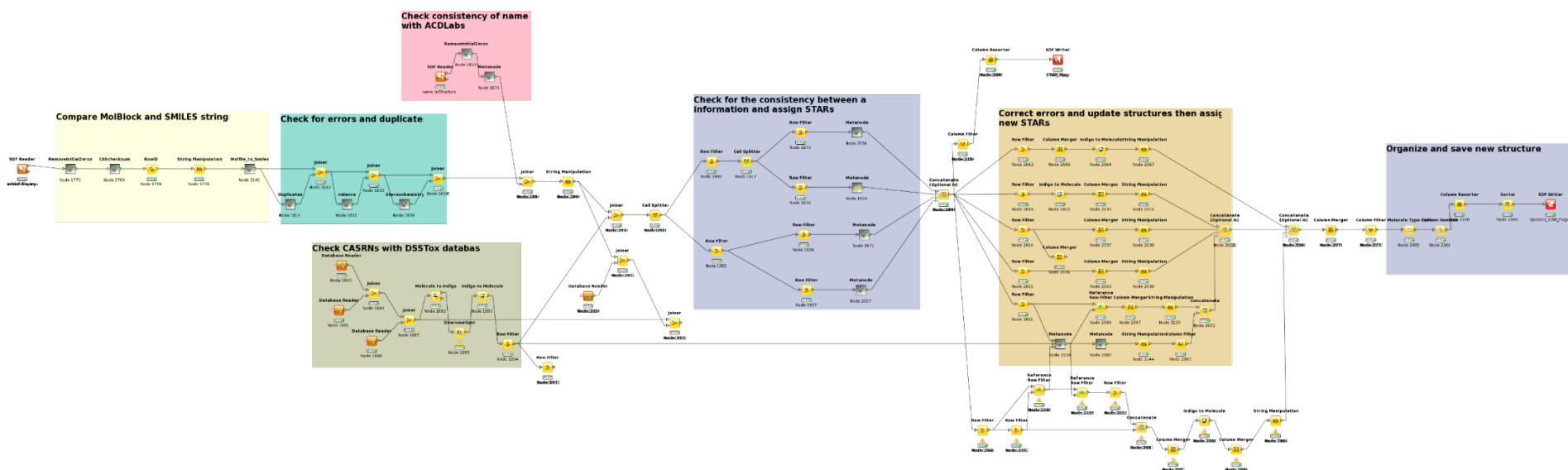
[Click here to download Data_for_BCFBAF.zip](#)

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](#)

Abbreviation	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

KNIME Workflow to Evaluate the Dataset



Mansouri et al. (<https://www.tandfonline.com/doi/abs/10.1080/1062936X.2016.1253611>)

The InChI Identifier

- Unique code managed by IUPAC: No variability as with SMILES
- InChI Strings can be reversed to structures: same as with SMILES
- Adopted by the community (databases, blogs, Wikipedia): good for searching the internet

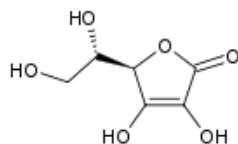
International Chemical Identifier

From Wikipedia, the free encyclopedia
(Redirected from [InChI](#))

The **IUPAC International Chemical Identifier (InChI)**, pronounced "INchee") is a textual [identifier](#) for [chemical substances](#), designed to provide a standard and human-readable way to encode molecular information and to facilitate the search for such information in databases and on the web. Developed by [IUPAC](#) and [NIST](#) during 2000-2005, the format and algorithms are non-proprietary and the software is freely available under the [open source LGPL](#) license (though the term "InChI" is a [trademark](#) of IUPAC).^[1]

CH₃CH₂OH
[ethanol](#)

InChI=1/C2H6O/c1-2-3/h3H,2H2,1H3

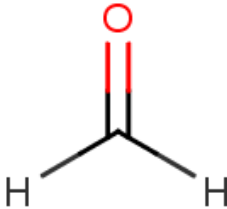


[L-ascorbic acid](#)

InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1

Check and Curate Public Data

- Public data should always be checked and curated prior to modeling. This dataset was no different.
- The data files have **FOUR** representations of a chemical, plus the property value.

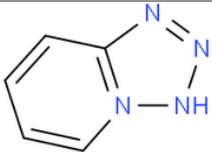
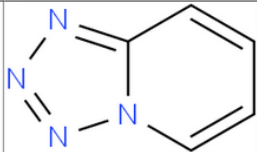
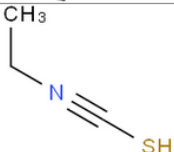
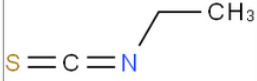
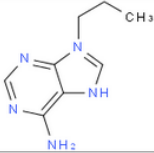
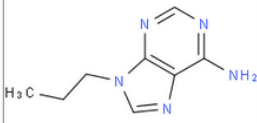
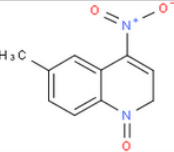
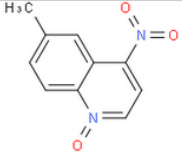
sdf Molecule	Mol Mol Block	S Smiles	S CAS	S NAME	D Kow
<pre>-ISIS- 09141018452D 4 3 0 0 0 0 0 0 0 0 0999 V2000 2.4667 -0.0833 0.0000 O 0 0 ... 2.4667 -0.9125 0.0000 C 0 0 ... 1.7500 -1.3292 0.0000 H 0 0 ... 3.1833 -1.3292 0.0000 H 0 0 ... 2 1 2 0 0 0 0 3 2 1 0 0 0 0 4 2 1 0 0 0 0 M END > <CAS> (000050-00-0)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

LogP dataset: 15,809 structures

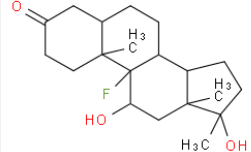
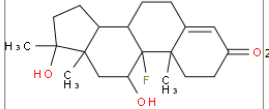
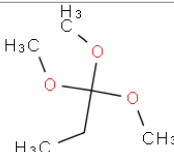
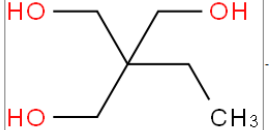
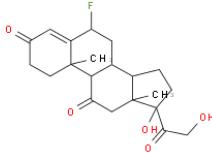
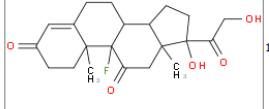
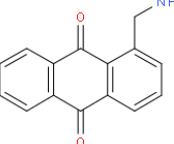
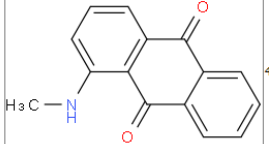
- CAS Checksum: 12163 valid, 3646 invalid (**>23%**)
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES (**>24%**)
- Duplicates check:
 - 31 DUPLICATE MOLFILES
 - 626 DUPLICATE SMILES
 - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
 - 1279 differ in stereochemistry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

Examples of Errors

Valence Errors

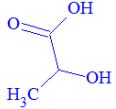
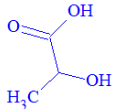
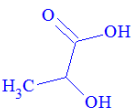
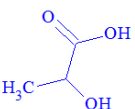
Mol Block	S CAS	S NAME	Smiles
	000274-87-3	TETRAZOLO[1,5-A]PYRIDINE	
	000542-85-8	ETHYL ISOTHIOCYANATE	
	000707-98-2	9-PROPYL ADENINE	
	000715-48-0	6-METHYL-4-NITROQUINOLINE-1-OXIDE	

Different Compounds

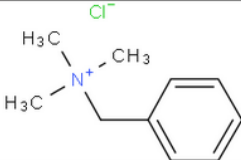
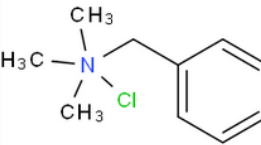
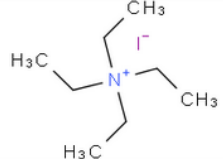
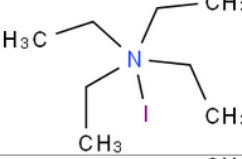
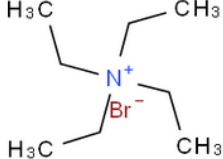
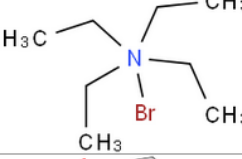
Mol Block	S CAS	S NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Examples of Errors

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	2.2660000000000000e+001	5.8600000000000000e+000
	C ₃ H ₆ O ₃	90.0779	000079-33-4	L-LACTIC ACID	5.3000000000000000e+001	2.2660000000000000e+001	-3.0340000000000000e+001
	C ₃ H ₆ O ₃	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.8000000000000000e+001	2.2660000000000000e+001	4.6600000000000000e+000
	C ₃ H ₆ O ₃	90.0779	010326-41-7	D-LACTIC ACID	5.2800000000000000e+001	2.2660000000000000e+001	-3.0140000000000000e+001

Covalent Halogens

Mol Block	S CAS	S NAME	Smiles
	000056-93-9	BENZYL TRIMETHYL AMMONIUM CHLORIDE	
	000068-05-3	TETRAETHYL AMMONIUM IODIDE	
	000071-91-0	TETRAETHYL AMMONIUM BROMIDE	

Other issues

Invalid CASRNs

Truncated names

Missing SMILES

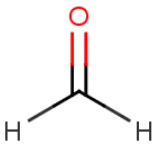
SRC000-02-7	Ethanaminium, N,N,N-trimethyl-2-[(1-oxo-2-propen
SRC000-04-3	Guanidine, N-hydroxy-N'-[4-(methylthio)benzeneme
SRC000-04-4	Hydrazinecarboximidamide, N'-[4-(methylthio)benz
SRC000-04-5	NNN5-TeMe-N-(3FuranMe),ammon Br
SRC000-04-6	Benzenamine, 4-bromo-N,N-bis(2,2,2-trifluoroethy
SRC000-04-7	2-Propenoic acid, 3-(2-chlorophenoxy)-, methyl e
SRC000-05-1	9H-Purine-9-acetaldehyde, a-(1-formyl-2-hydroxye
SRC000-05-2	N1-Pr-N2-CN-N3-Me guanidine
SRC000-05-3	1-(2-OHEt)-2-Me imidazoline HCL
SRC000-06-3	Propanoic acid, 3-[[[(4-cyanophenyl)methyl]seleno

Data Files & Quality flags

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

4 levels of consistency exists among:

- The Molblock
- The SMILES string
- The chemical name (based on ACD/Labs dictionary)
- The CAS Number (based on a DSSTox lookup)

sdm Molecule	Mol Mol Block	S Smiles	S CAS	S NAME	D Kow
<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>		O=C	000050-00-0	FORMALDEHYDE	0.35

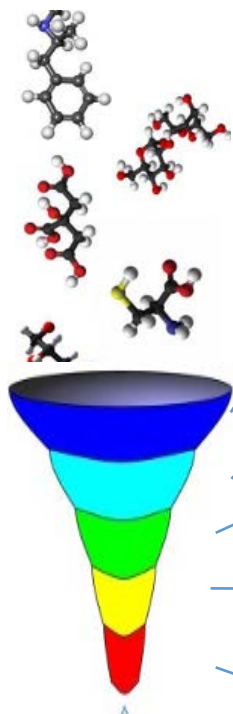
<http://esc.syrres.com/interkow/EpiSuiteData.htm>



Quality FLAGS and curated structures

QSAR-ready standardization procedure

Initial
structures



QSAR-ready
structures

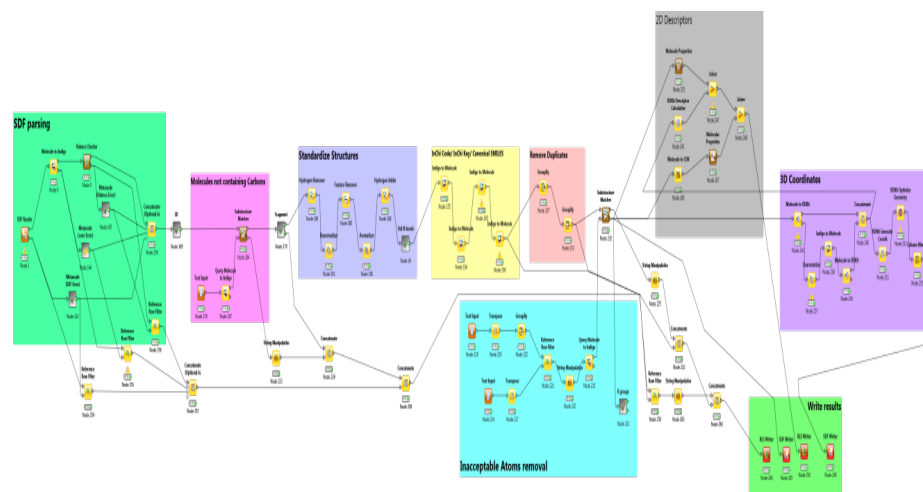
Remove inorganics
and mixtures

Clean salts and
counterions

Normalize of
tautomers

Remove of
duplicates

Final inspection



KNIME workflow
UNC, DTU, EPA Consensus

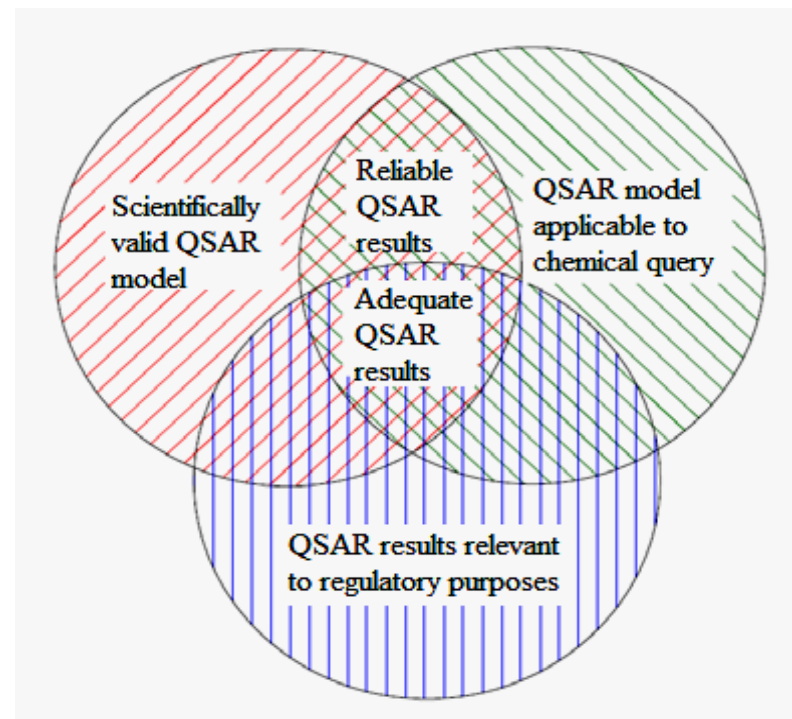
Curation to QSAR Ready Files

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>)

Development of a QSAR model

- 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 2D descriptors and fingerprints generated and shared*
- 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*



Following the 5 OECD Principles*

Principle	Description
1) A defined endpoint	Any physicochemical, biological or environmental effect that can be measured and therefore modelled.
2) An unambiguous algorithm	Ensure transparency in the description of the model algorithm.
3) A defined domain of applicability	Define limitations in terms of the types of chemical structures , physicochemical properties and mechanisms of action for which the models can generate reliable predictions .
4) Appropriate measures of goodness-of-fit, robustness and predictivity	a) The internal fitting performance of a model b) the predictivity of a model, determined by using an appropriate external test set .
5) Mechanistic interpretation, if possible	Mechanistic associations between the descriptors used in a model and the endpoint being predicted .

OPERA Models

Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
BCF	10	0.84	0.55	465	0.85	0.53	161	0.83	0.64
BP	13	0.93	22.46	4077	0.93	22.06	1358	0.93	22.08
LogP	9	0.85	0.69	10531	0.86	0.67	3510	0.86	0.78
MP	15	0.72	51.8	6486	0.74	50.27	2167	0.73	52.72
VP	12	0.91	1.08	2034	0.91	1.08	679	0.92	1
WS	11	0.87	0.81	3158	0.87	0.82	1066	0.86	0.86
HL	9	0.84	1.96	441	0.84	1.91	150	0.85	1.82

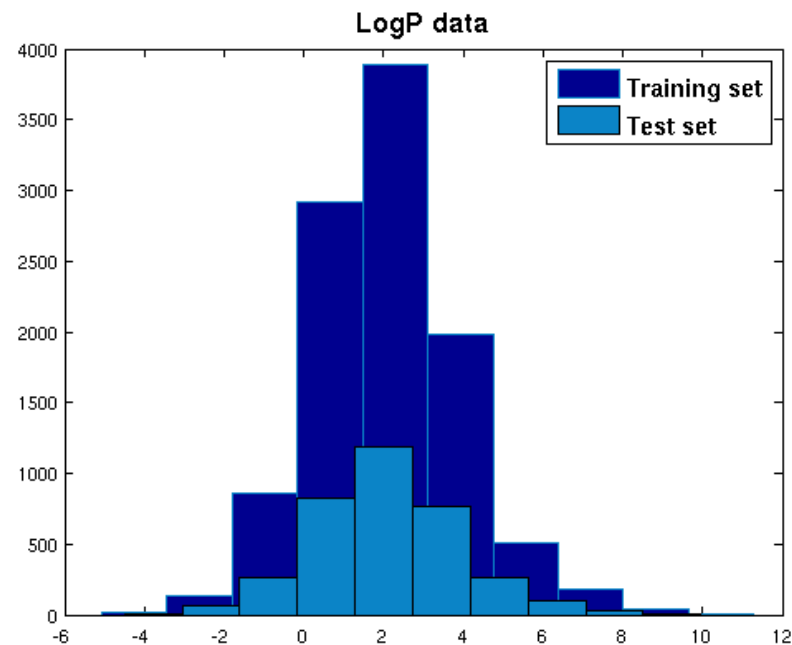
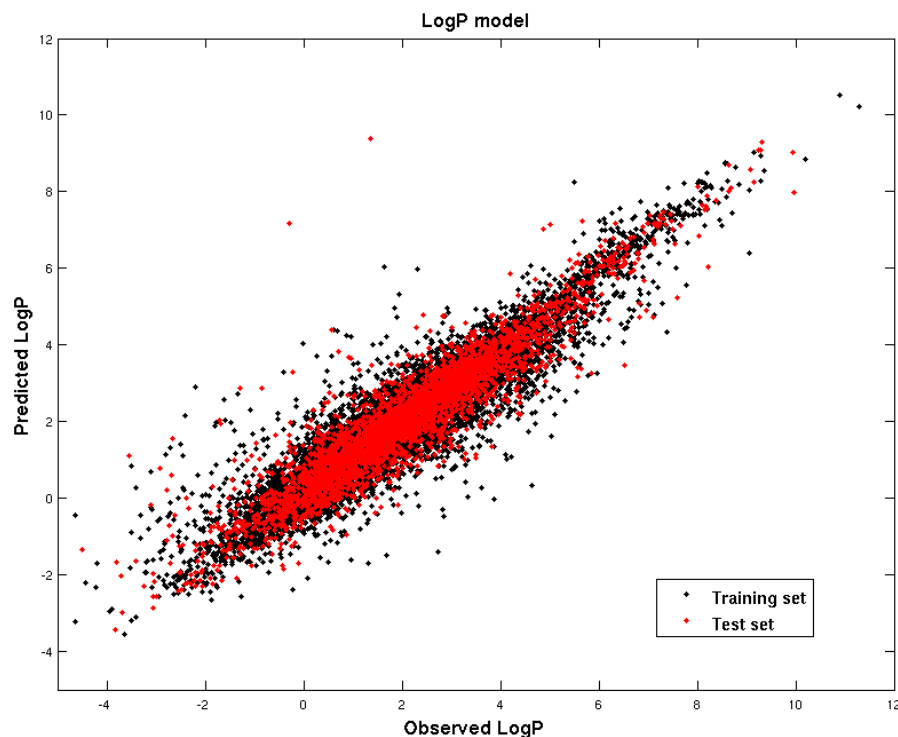
Mansouri et al. OPERA models. (<https://link.springer.com/article/10.1186/s13321-018-0263-1>)

OPERA Models statistics

Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
AOH	13	0.85	1.14	516	0.85	1.12	176	0.83	1.23
BioHL	6	0.89	0.25	112	0.88	0.26	38	0.75	0.38
KM	12	0.83	0.49	405	0.82	0.5	136	0.73	0.62
KOC	12	0.81	0.55	545	0.81	0.54	184	0.71	0.61
KOA	2	0.95	0.69	202	0.95	0.65	68	0.96	0.68
		BA	Sn-Sp		BA	Sn-Sp		BA	Sn-Sp
R-Bio	10	0.8	0.82-0.78	1198	0.8	0.82-0.79	411	0.79	0.81-0.77

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

LogP Model: weighted kNN Model, 9 descriptors



Weighted 5-nearest neighbors

9 Descriptors

Training set: 10531 chemicals

Test set: 3510 chemicals

5 fold CV: $Q^2=0.85$, $RMSE=0.69$

Fitting: $R^2=0.86$, $RMSE=0.67$

Test: $R^2=0.86$, $RMSE=0.78$

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

OPERA Standalone application:



Input:

- SDF file or SMILES strings of QSAR-ready structures. In this case the program will calculate PaDEL 2D descriptors and make the predictions.
- Calculated PaDEL descriptors

Output

- A list of molecules IDs and predictions
- Applicability domain
- Accuracy of the prediction
- Similarity index to the 5 nearest neighbors
- The 5 nearest neighbors from the training set: Exp. value, Prediction, InChi key

OPERA on Github

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

 This repository Search Pull requests Issues Marketplace Explore

 kmansouri / OPERA Watch 1 Star 2 Fork 0

[Code](#) [Issues 1](#) [Pull requests 0](#) [Projects 0](#) [Wiki](#) [Insights](#) [Settings](#)

Command line application providing QSAR models predictions as well as applicability domain and accuracy assessment for physicochemical properties and environmental fate endpoints. [Edit](#)

[Add topics](#)

49 commits1 branch0 releases1 contributorMIT

Branch: master New pull requestCreate new fileUpload filesFind fileClone or download

 kmansouri OPERA 1.5 Linux Latest commit 4492331 on Dec 18, 2017

 Icon.png	OPERA 1.2 icon	10 months ago
 LICENSE	Initial commit	a year ago
 Logo.png	Added logo and icon	a year ago
 Matlab_Source_code.zip	OPERA 1.5 MATLAB source code	3 months ago
 OPERA_CLi_Linux.tar.gz	OPERA 1.5 Linux	3 months ago
 OPERA_C_library.tar.gz	OPERA 1.5 C Library	3 months ago
 OPERA_Data_SDF.zip	OPERA 1.5 Datasets	5 months ago
 OPERA_py.zip	OPERA 1.5 Python Library	3 months ago
 OPERA_win.zip	OPERA 1.5 Windows	3 months ago

OPERA on EPA's Comptox Chemistry Dashboard



<https://comptox.epa.gov>

 United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads

20182

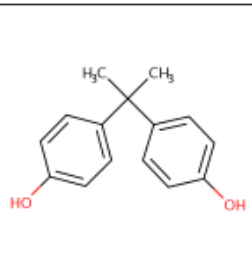
Chemistry Dashboard

OPERA Models: LogP: Octanol-Water

Save PDF

Bisphenol A

80-05-7 | DTXSID7020182



Model Results

Predicted value: 3.35

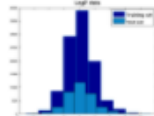
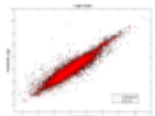
Global applicability domain: Inside

Local applicability domain index: 0.88

Confidence level: 0.75

Calculation Result for a chemical

Model Performance



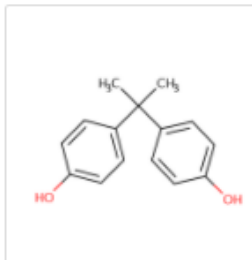
Model Performance with full QMRF

Weighted KNN model

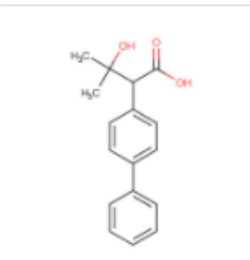
QMRF

5-fold CV (76%)		Training (76%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.86	0.67	0.86	0.78

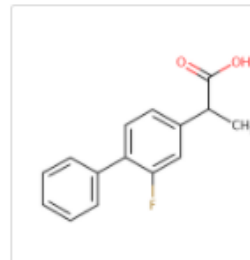
Nearest Neighbors from the Training Set



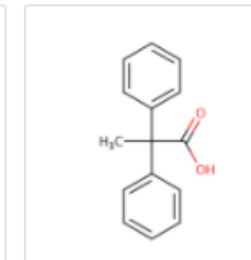
Bisphenol A
Measured: 3.32
Predicted: 3.35



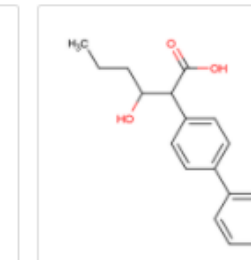
BUTANOIC ACID 2-(4-BIPHENYL)-3-HYDROXY-2-METHYL-
Measured: 3.25
Predicted: 3.45



Flurbiprofen
Measured: 4.16
Predicted: 3.83



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



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

Nearest Neighbors from Training Set

 UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

Discover.

About/Disclaimer
Accessibility
Privacy

Connect.

ACToR
DSSTox
OpenTox

Ask.

Contact
Help

OPERA on EPA's Comptox Chemistry Dashboard

<https://comptox.epa.gov>

The screenshot displays the EPA Comptox Chemistry Dashboard interface. At the top, the EPA logo and navigation tabs (Home, Advanced Search, Batch Search, Lists, Predictions, Downloads) are visible. A search bar labeled "Search All Data" is on the right. Below the navigation bar, the "Chemistry Dashboard" title is shown. A progress bar at the top indicates the current step: "Step Five: Choose Data Fields to Download".

Below the progress bar, the "Select Output Format" dropdown is set to "Excel". Under "Customize Results", there are checkboxes for "Select All" and "Select All In Lists".

The main section is titled "Intrinsic And Predicted Properties" and contains a list of checkboxes with information icons:

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

Below this is the "Metadata" section with checkboxes:

- ☐ OPERA Model Predictions **i**
- ☐ TEST Model Predictions **i**

On the right side, there are checkboxes for "ITN ANTIBIOTIC L" and "KEML List of Subst".

A tooltip box on the right contains the following text:

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).

At the bottom right, there are checkboxes for "Massbank.EU Collection: Special Cases", "National Environmental Methods Index", and "National-Scale Air Toxics Assessment".

QMRF Reports



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



European
Commission

JOINT RESEARCH CENTRE

The European Commission's science and knowledge service

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			
	Q17-23a-0014	OPERA-model for Readily 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 rate 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>

Real Time Predictions in Development

 United States Environmental Protection Agency

Home Advanced Search Batch Search Lists Predictions Downloads

Search All Data

Chemistry Dashboard

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

130%

Select properties to predict

T.E.S.T. 18 OPERA **In Development**

Toxicological properties + -

- ☒ 96 hour fathead minnow LC50
- ☒ 48 hour D. magna LC50
- ☒ 48 hour T. pyriformis IGC50
- ☒ Developmental toxicity
- ☒ Ames mutagenicity
- ☒ Estrogen Receptor RBA
- ☒ Estrogen Receptor Binding

Physical properties

- ☒ Normal boiling point
- ☒ Melting point
- ☒ Flash point
- ☒ Surface tension
- ☒ Thermal conductivity
- ☒ Viscosity
- ☒ Water solubility

Chemical Structure

Cc1ccc(O)cc1C(C)c2ccc(O)cc2

Chemical structure of 2,2-bis(4-hydroxyphenyl)propane (BPA) is shown. The structure consists of two phenol rings connected by a central carbon atom bonded to two methyl groups.

Calculate

Future Work

- Structural properties:
Hybridization Ratio, nHBAcc, nHBDON, LipinskiFailures, Topo
PSA, Molar refractivity, Polarizability, electronegativity
- HPLC retention time
- pKa
- Log D
- Bioaccumulation factor
- Estrogen and Androgen Receptor activity
- Acute toxicity

Thank you for your attention

26



SAR and QSAR in Environmental Research



ISSN: 1062-936X (Print) 1029-046X (Online) Journal homepage: <http://www.tandfonline.com/loi/gsar20>

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



[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

23

Shares

525

Downloads