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

ScitoVation
INNOVATIVE CELL BASED SCIENCE

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

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)

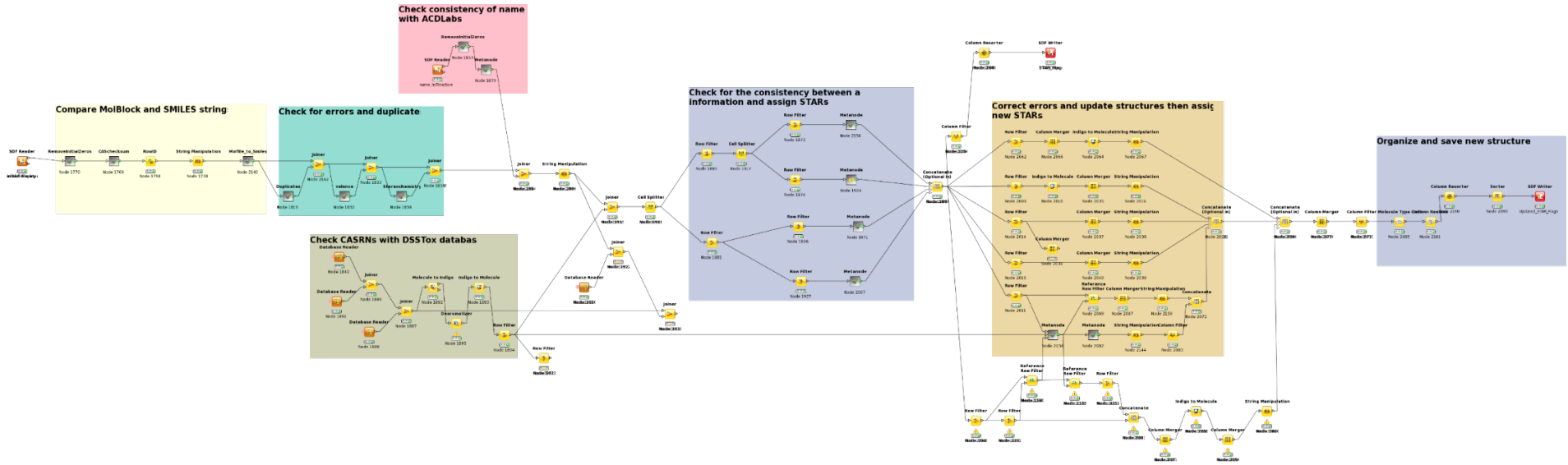
Click here to download Data_for_BCFBAF.zip

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

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

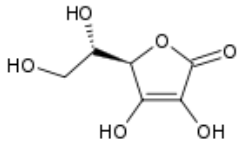
International Chemical Identifier

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]

CH_3CH_2OH
ethanol

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

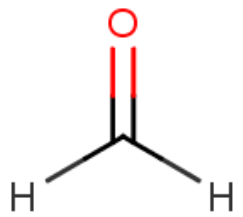


L-ascorbic acid

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.

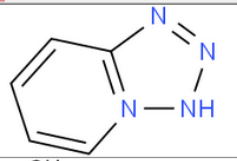
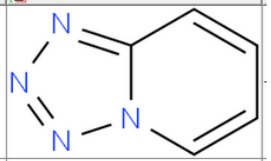
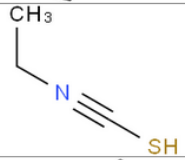
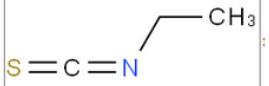
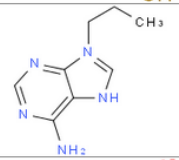
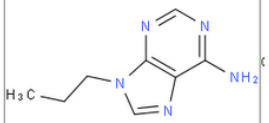
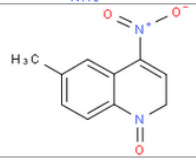
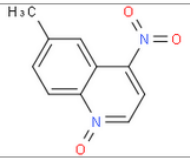
SDF Molecule	Mol Block	S Smiles	S CAS	S NAME	D Kow
-ISIS- 09141018452D 4 3 0 0 0 0 0 0 0 0999 V2000 2.4667 -0.0833 0.0000 O 0 0 ... 2.4667 -0.9125 0.0000 C 0 0 ... 1.7500 -1.3292 0.0000 H 0 0 ... 3.1833 -1.3292 0.0000 H 0 0 ... 2 1 2 0 0 0 0 3 2 1 0 0 0 0 4 2 1 0 0 0 0 M END > <CAS> (000050-00-0) 000050-00-0 > <NAME> (000050-00-0) FORMALDEHYDE > <Kow> (000050-00-0) 3.500000000000000e-001		O=C	000050-00-0	FORMALDEHYDE	0.35

LogP dataset: 15,809 structures

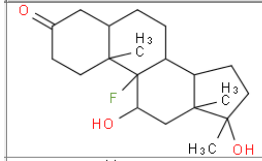
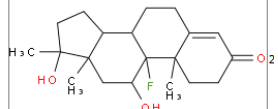
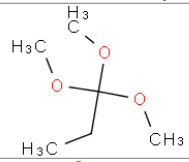
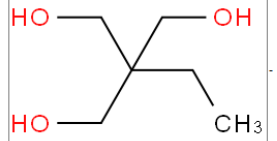
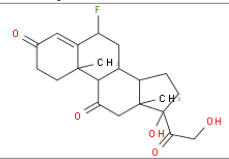
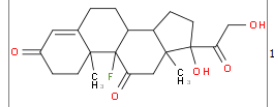
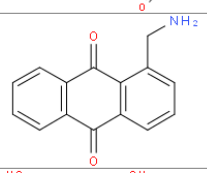
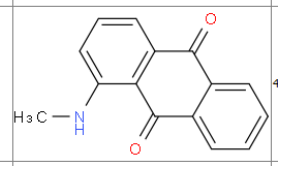
- CAS Checksum: 12163 valid, 3646 invalid **(>23%)**
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES **(>24%)**
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in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

Examples of Errors

Valence Errors

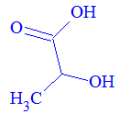
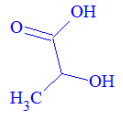
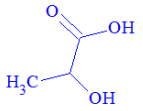
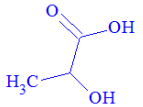
Mol Block	CAS	NAME	Smiles
	000274-87-3	TETRAZOLO[1,5-A]PYRIDINE	
	000542-85-8	ETHYL ISOTHIOCYANATE	
	000707-98-2	9-PROPYL ADENINE	
	000715-48-0	6-METHYL-4-NITROQUINOLINE-1-OXIDE	

Different Compounds

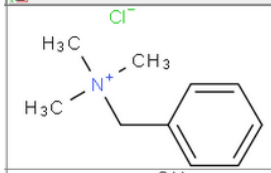
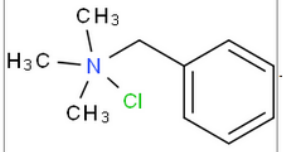
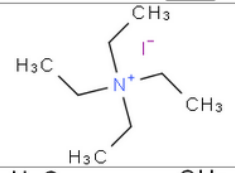
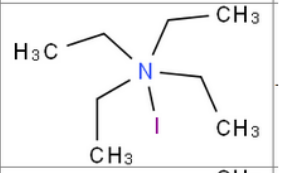
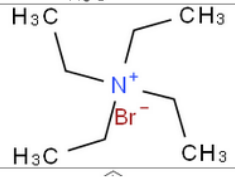
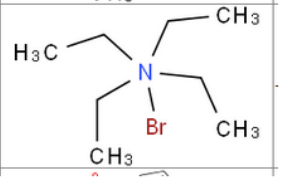
Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Examples of Errors

Duplicate Structures

Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_3H_6O_3$	90.0779	000050-21-5	LACTIC ACID	1.6800000000000 00e+001	2.2660000000000 00e+001	5.8600000000000 00e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.3000000000000 00e+001	2.2660000000000 00e+001	-3.034000000000 000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.8000000000000 00e+001	2.2660000000000 00e+001	4.6600000000000 00e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.2800000000000 00e+001	2.2660000000000 00e+001	-3.014000000000 000e+001

Covalent Halogens

Mol Block	CAS	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

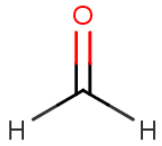
Truncated names

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.

SDF Molecule	Mol Mol Block	S Smiles	S CAS	S NAME	D Kow
-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) 3.500000000000000e-001		O=C	000050-00-0	FORMALDEHYDE	0.35

4 levels of consistency exists among:

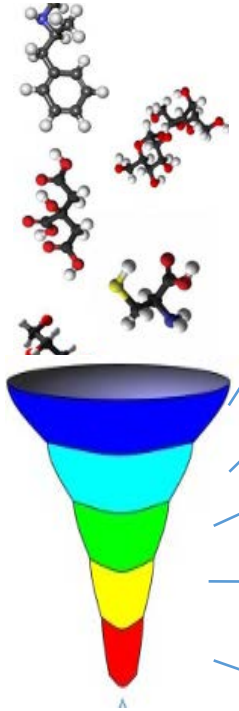
- The Molblock
- The SMILES string
- The chemical name (based on ACD/Labs dictionary)
- The CAS Number (based on a DSSTox lookup)

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

Quality FLAGS and curated structures

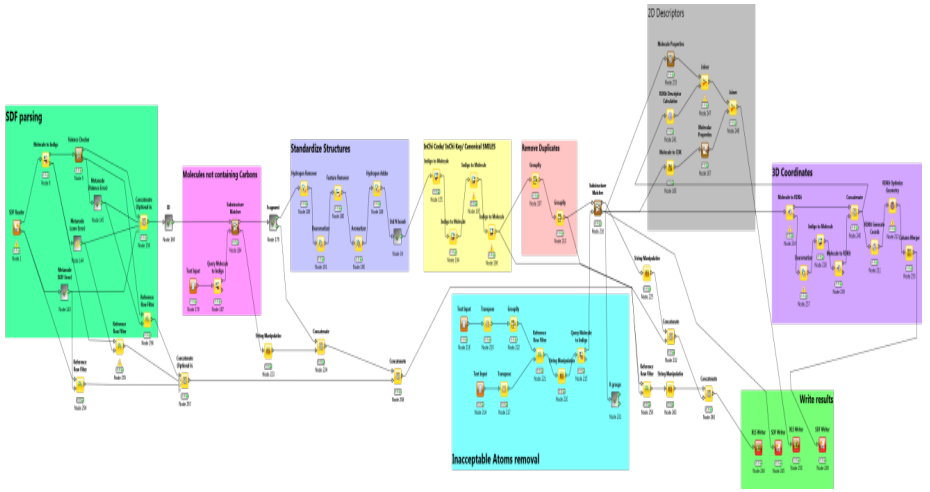
QSAR-ready standardization procedure

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

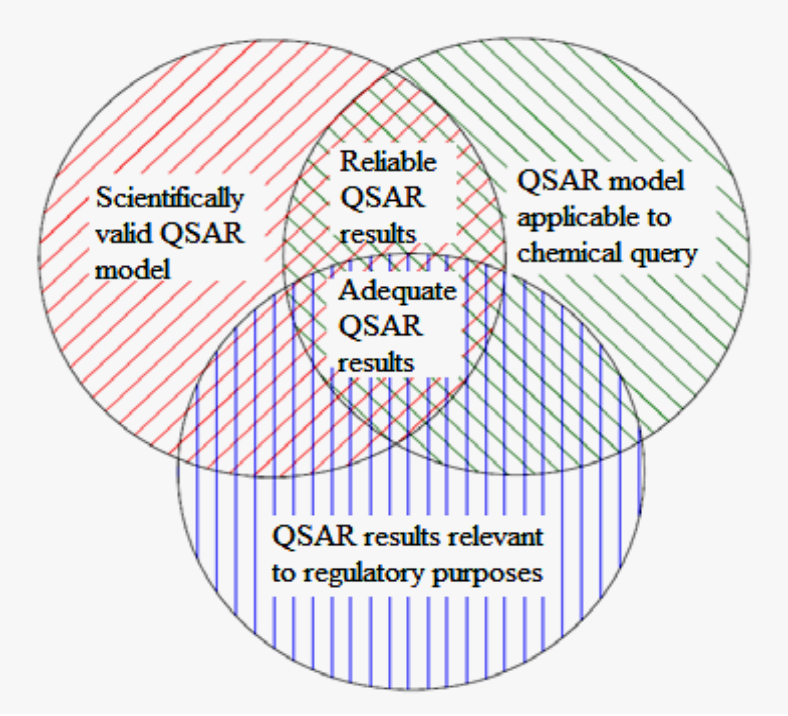
Curation to QSAR Ready Files

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 SDFiles 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**.

http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf

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

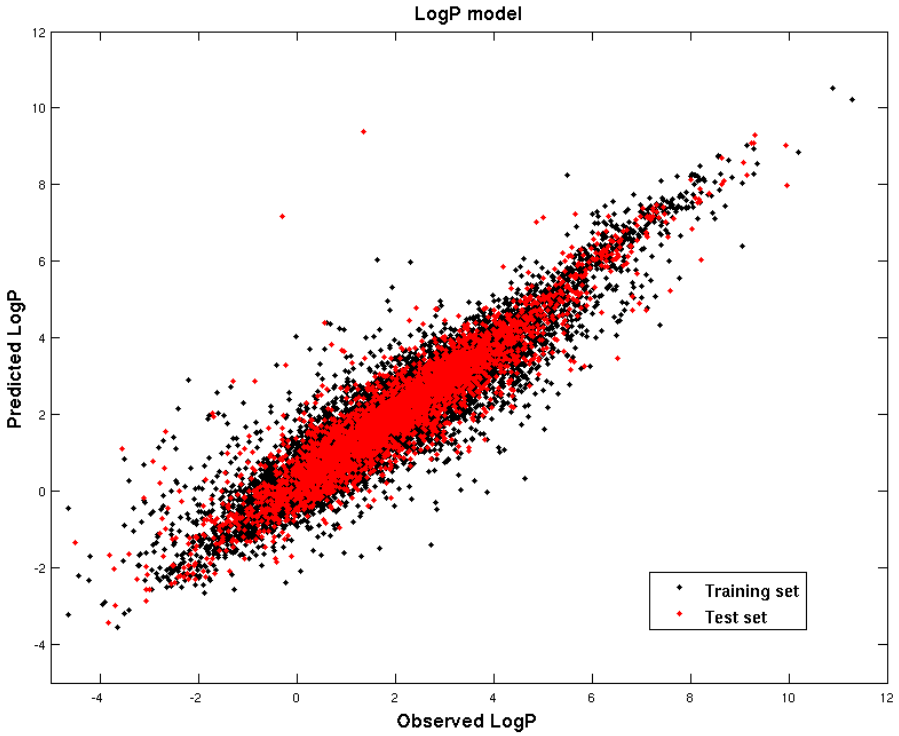
Mansouri et al. OPERA models. (https://link.springer.com/article/10.1186/s13321-018-0263-1)

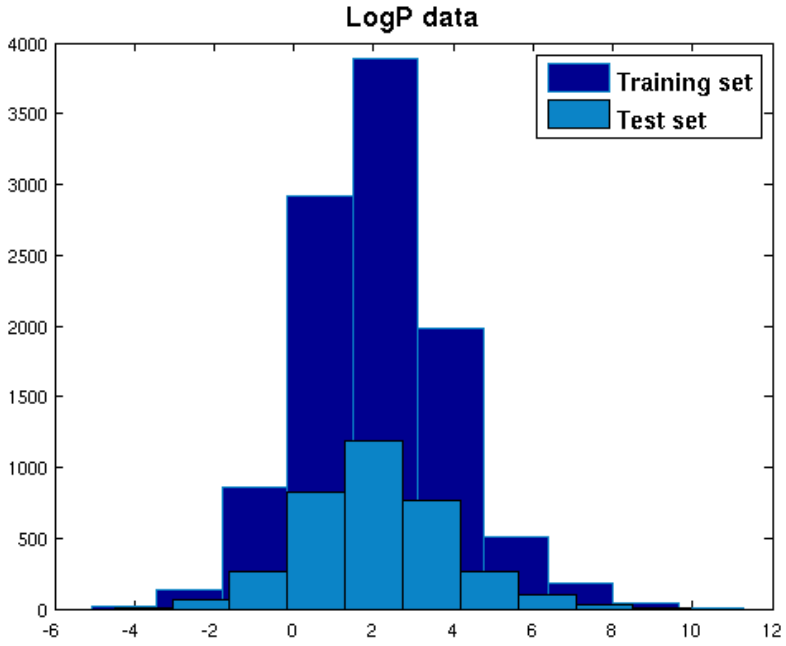
OPERA Models statistics

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: Q2=0.85,RMSE=0.69
Fitting: R2=0.86,RMSE=0.67
Test: R2=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

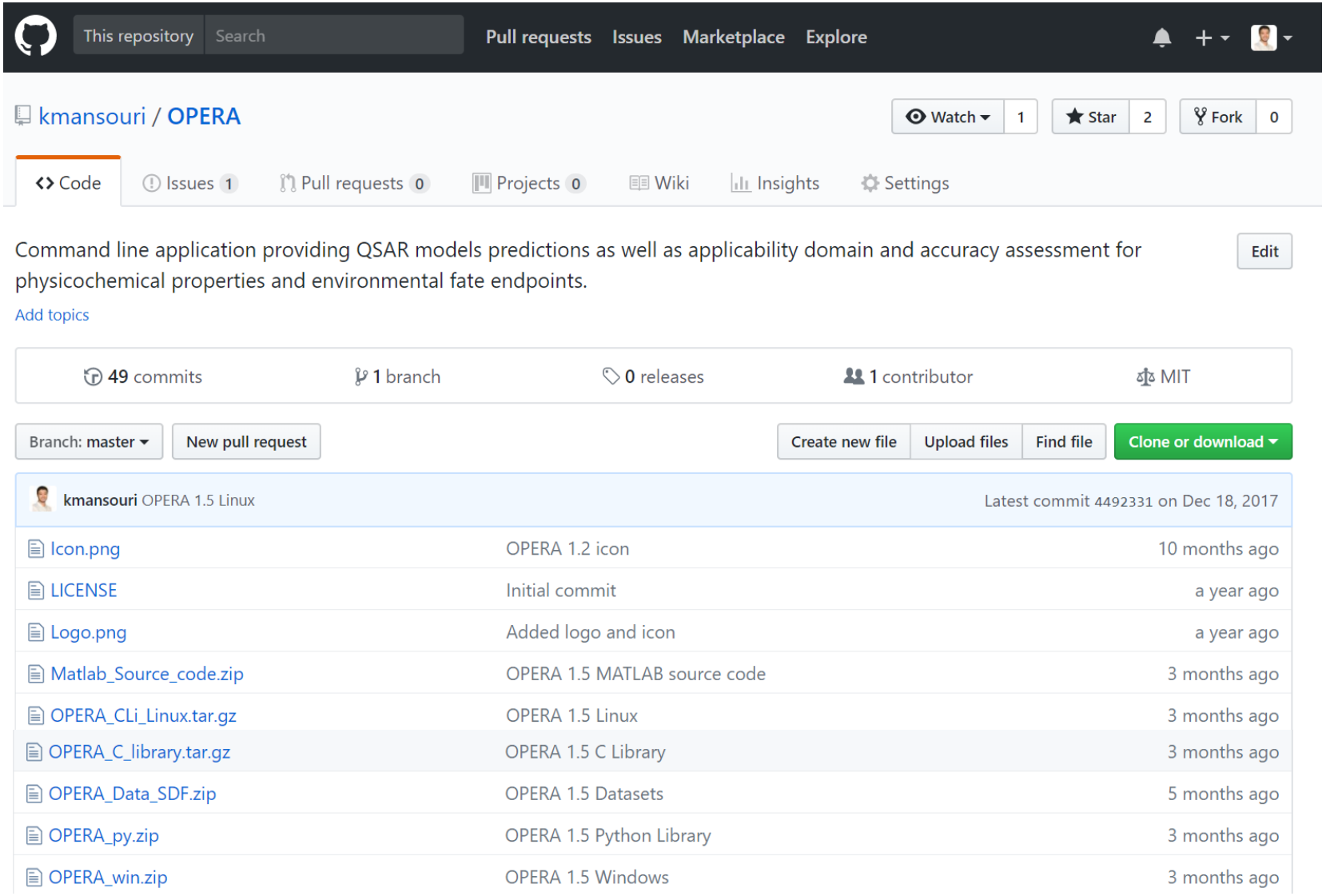
Output

- 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

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

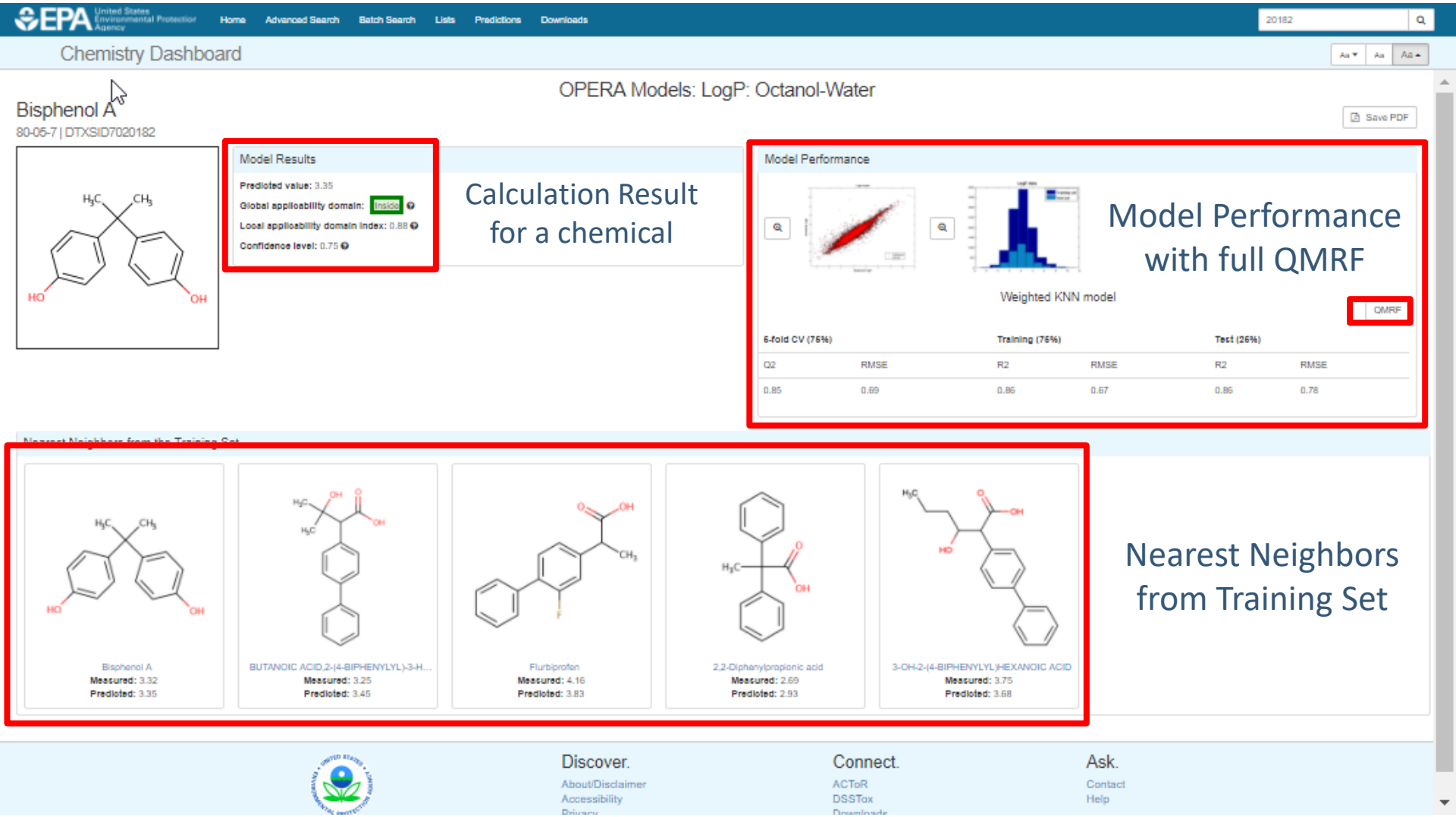
OPERA on Github

https://github.com/kmansouri/OPERA



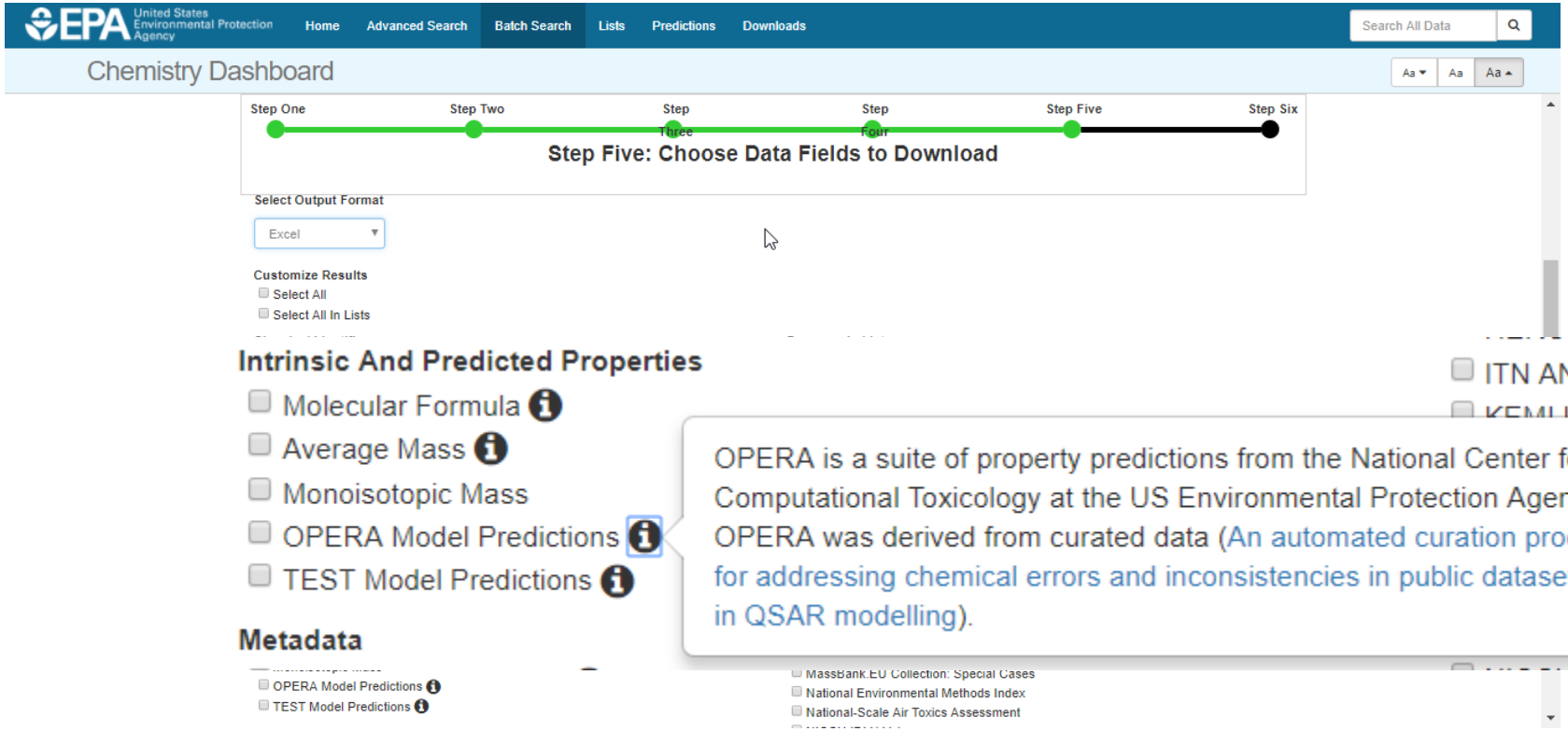
OPERA on EPA's Comptox Chemistry Dashboard

https://comptox.epa.gov

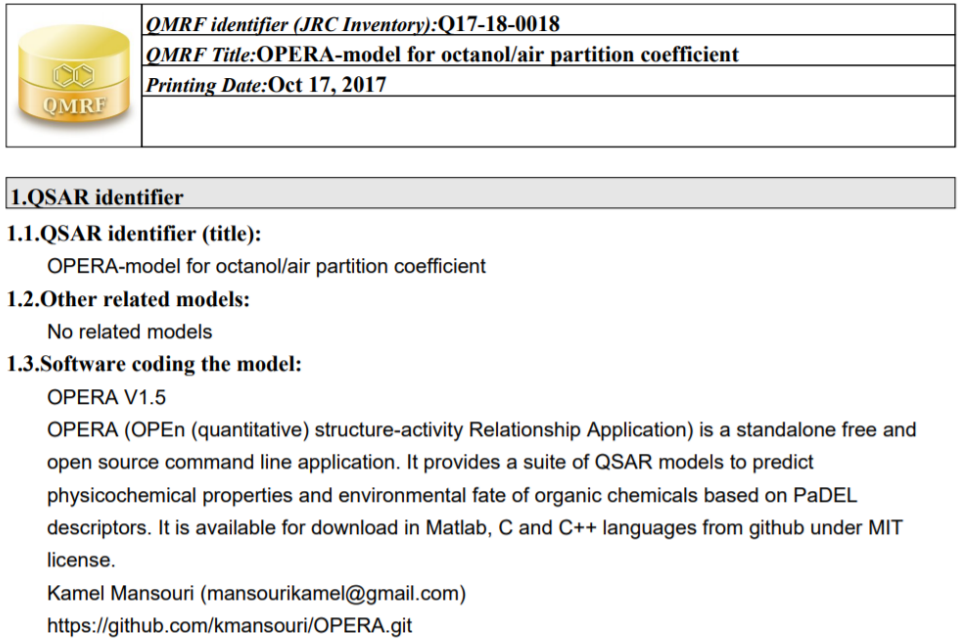


OPERA on EPA's Comptox Chemistry Dashboard

https://comptox.epa.gov

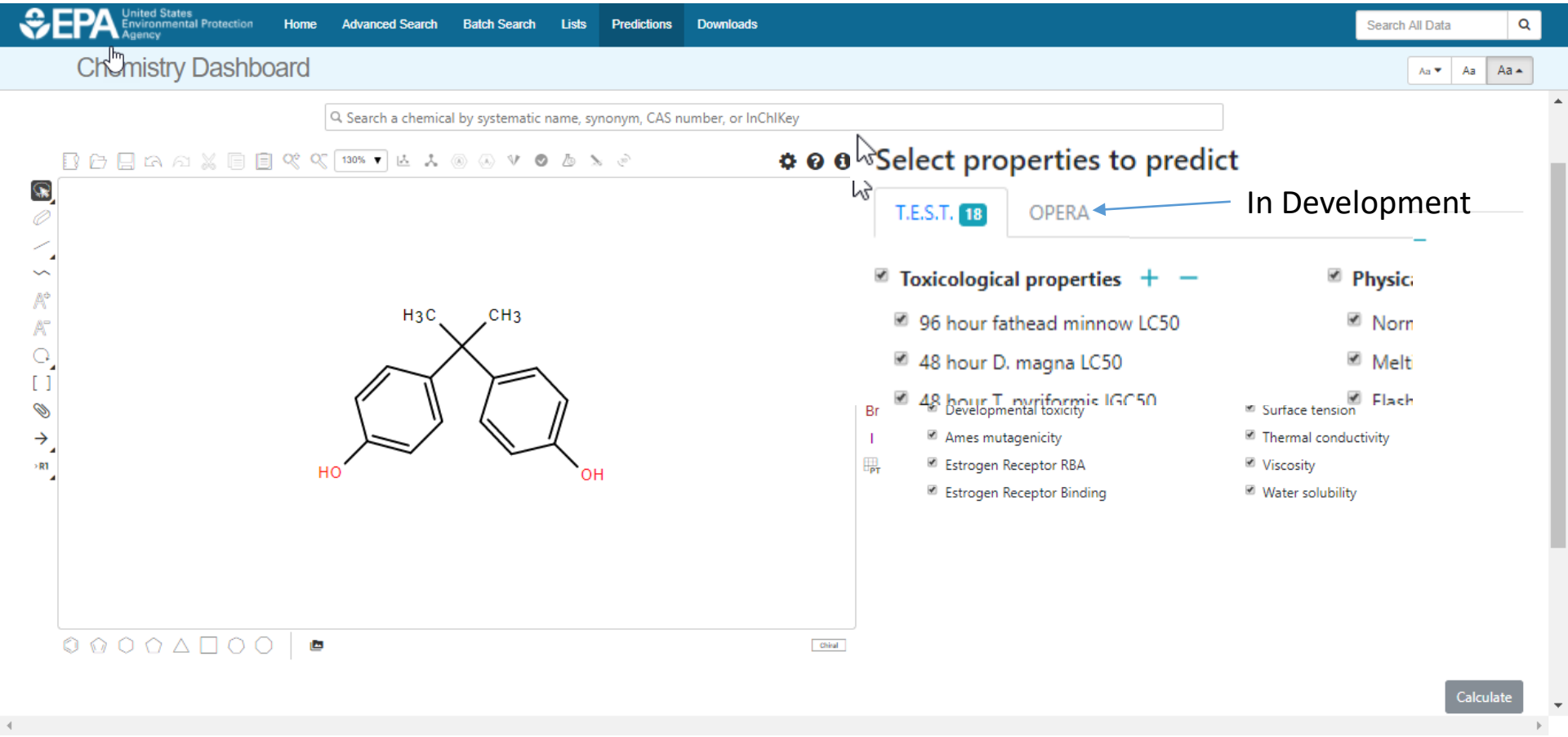


QMRF Reports



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

Real Time Predictions in Development



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

Taylor & Francis
Taylor & Francis Group

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