EPA United States Environmental Protection Agency

OPERA: A QSAR tool for physicochemical properties and environmental fate predictions

Kamel Mansouri
Chris Grulke
Richard Judson
Antony Williams

NCCT, U.S. EPA

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
COMPUTATIONAL TOXICOLOGY

American Chemical Society Meeting, Spring 2017

2-6 April 2017, San Francisco, CA

Office of Research and Development

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

Recent Cheminformatics Developments

- We are building a new cheminformatics architecture
- PUBLIC dashboard gives access to “curated chemistry”
- Focus on integrating EPA ***and*** external resources
- Aggregating and curating data, visualization elements and “services” to underpin other efforts
  - QSAR
  - Read-across
  - Non-targeted screening

OPERA Models

- Interest in physicochemical properties to include in exposure modeling, augmented with ToxCast HTS *in vitro* data etc.

- 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**

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

The Approach

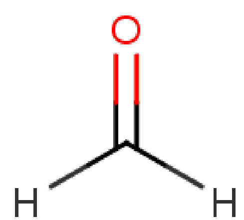
- To build models we need the set of chemicals and their property series

- Our curation process
  - Decide on the “chemical” by checking levels of consistency
  - We did NOT validate each measured property value
  - Perform initial analysis manually to understand how to clean the data (chemical structure and ID)
  - Automate the process (and test iteratively)
  - Process all datasets using final method

EPA United States Environmental Protection Agency

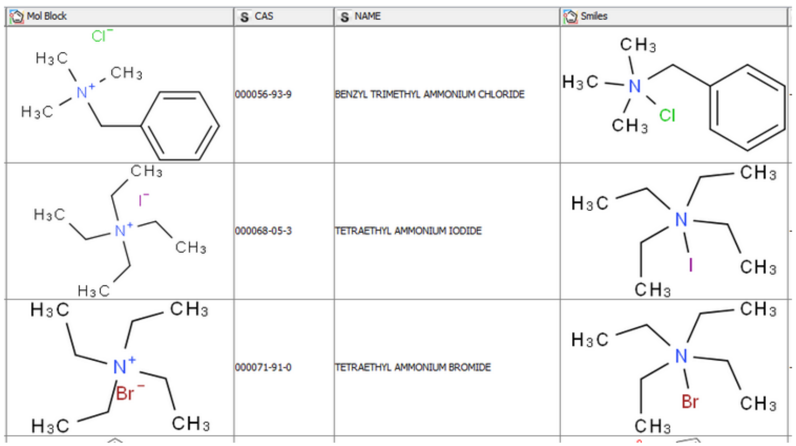
Check and Curate Public Data

- 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

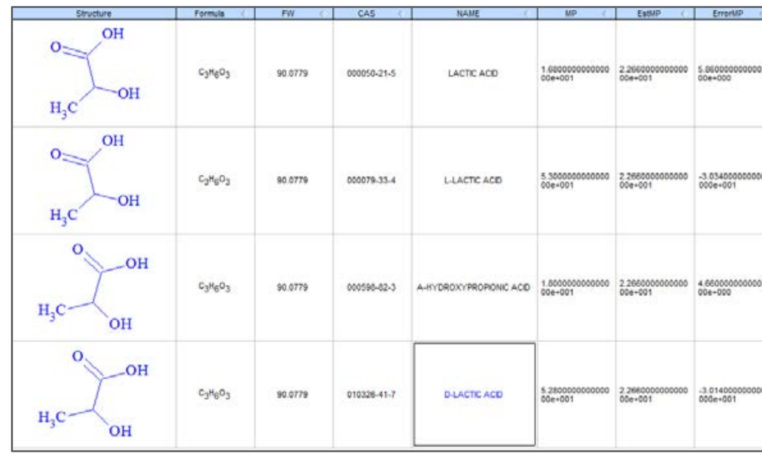
Check and Curate Public Data

- Public data should always be checked and curated prior to modeling. This dataset was no different.



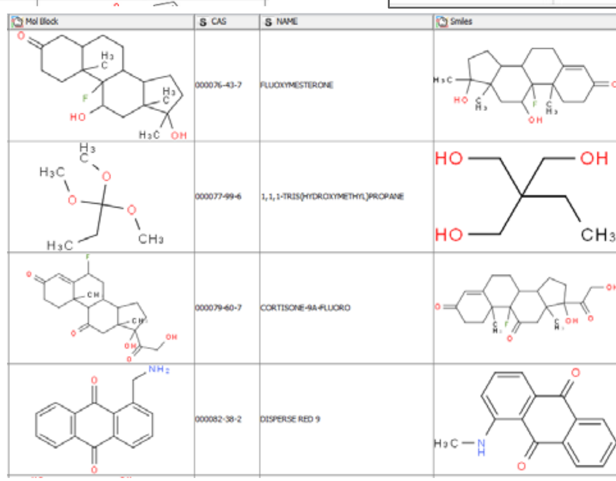
Mol Block	CAS	NAME	Smiles
	000056-93-9	BENZYL TRIMETHYL AMMONIUM CHLORIDE	
	000068-05-3	TETRAETHYL AMMONIUM IODIDE	
	000071-91-0	TETRAETHYL AMMONIUM BROMIDE	

Covalent Halogens



Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_3H_6O_3$	90.0779	000050-21-5	LACTIC ACID	1.680000000000000e+001	2.266000000000000e+001	5.860000000000000e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.300000000000000e+001	2.266000000000000e+001	-3.034000000000000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.800000000000000e+001	2.266000000000000e+001	4.660000000000000e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.280000000000000e+001	2.266000000000000e+001	-3.014000000000000e+001

Identical Chemicals

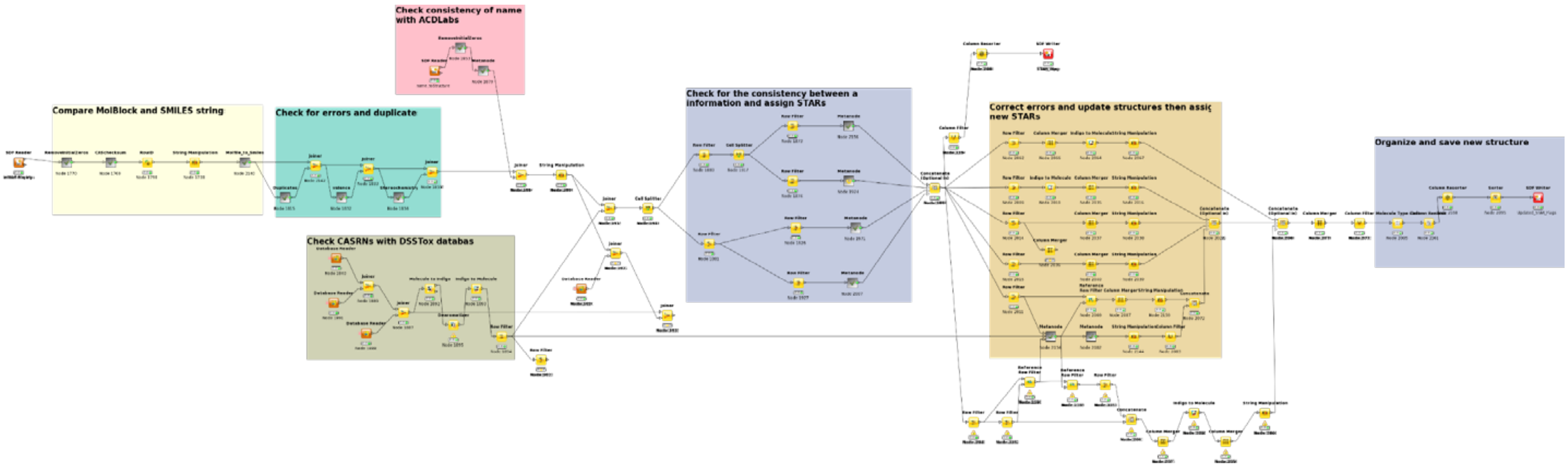


Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Mismatches

EPA
United States
Environmental Protection
Agency

KNIME Workflow to Evaluate the Dataset

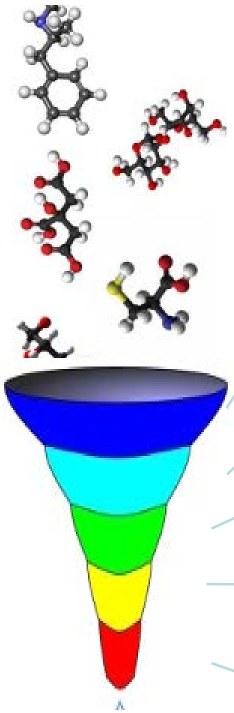


LogP dataset: 15,809 structures

- 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%)**

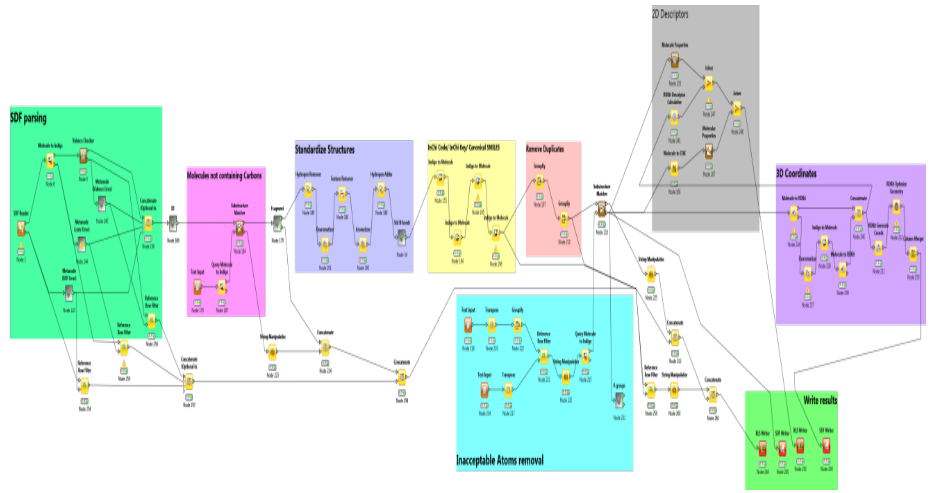
QSAR-ready standardization procedure

Initial structures



- Remove inorganics and mixtures
- Clean salts and counterions
- Normalize of tautomers
- Remove of duplicates
- Final inspection

QSAR-ready structures



KNIME workflow
UNC, DTU, EPA Consensus

Curation to QSAR Ready Files

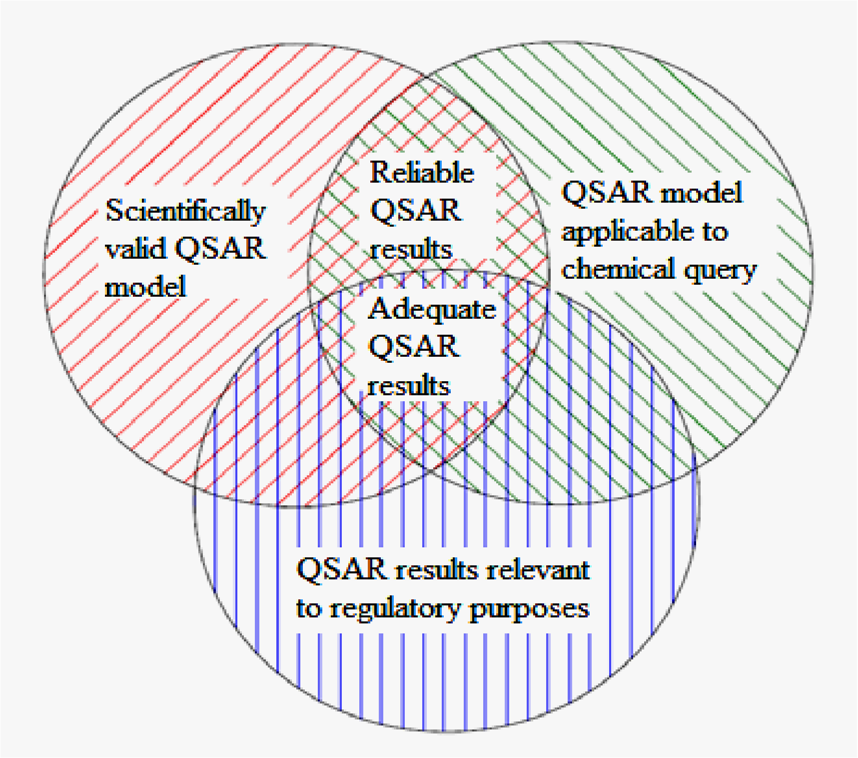
Property	Initial file	Curated Data	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	14544	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	2348	2046	2010

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*

QSARs validity, reliability, applicability and adequacy for regulatory purposes

ORCHESTRA. Theory, guidance and application on QSAR and REACH; 2012. http://home.deib.polimi.it/gini/papers/orchestra.pdf.



Following the 5 OECD Principles*

United States Environmental Protection Agency

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

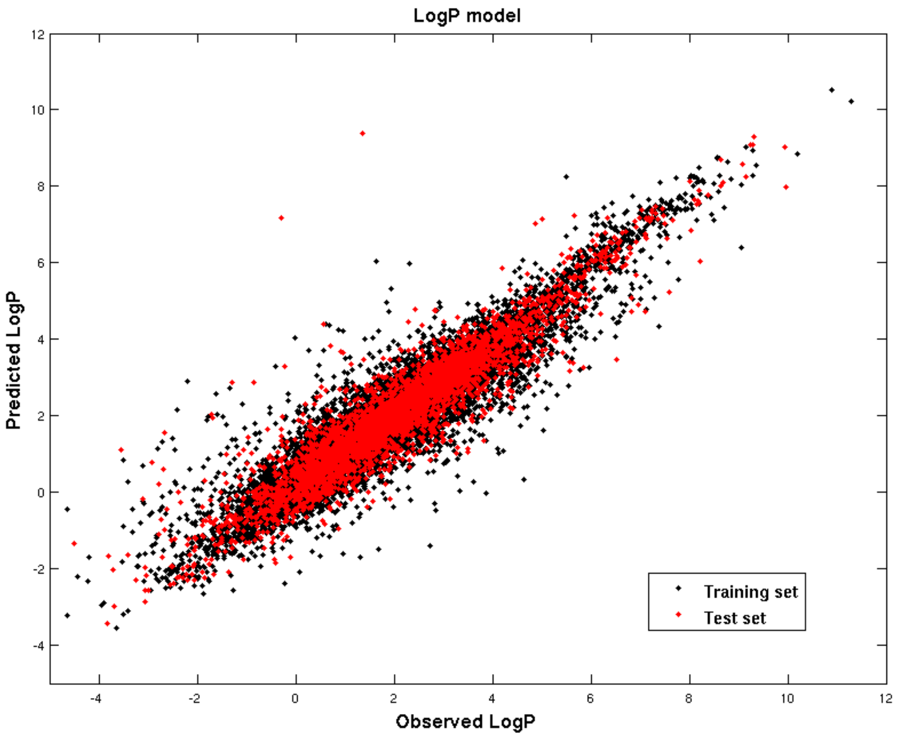
OPERA Models
What you would report in a paper

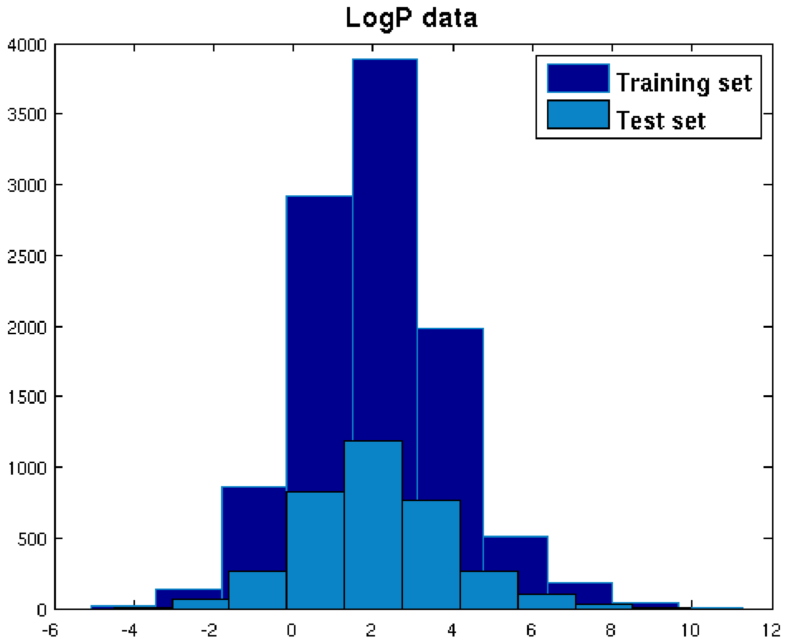
Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
BCF	10	0.84	0.55	465	0.85	0.53	161	0.83	0.64
BP	13	0.93	22.46	4077	0.93	22.06	1358	0.93	22.08
LogP	9	0.85	0.69	10531	0.86	0.67	3510	0.86	0.78
MP	15	0.72	51.8	6486	0.74	50.27	2167	0.73	52.72
VP	12	0.91	1.08	2034	0.91	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

OPERA Models

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

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:

- MATLAB .mat file, an ASCII file with only a matrix of variables
- SDF file or SMILES strings of QSAR-ready structures. In this case the program will calculate PaDEL 2D descriptors and make the predictions.

- The program will extract the molecules names from the input csv or SDF (or assign arbitrary names if not) As IDs for the predictions.

Output

- Depending on the extension, the can be text file or csv with
  - 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

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

Predictions for >720,000 Chemicals

- OPERA predictions were built on curated training sets
- All chemicals in DSSTox, accessed via the CompTox Dashboard, were pushed through all predictive models
- Predicted data made available, with detailed **MODEL REPORTS**

CompTox Chemistry Dashboard

https://comptox.epa.gov

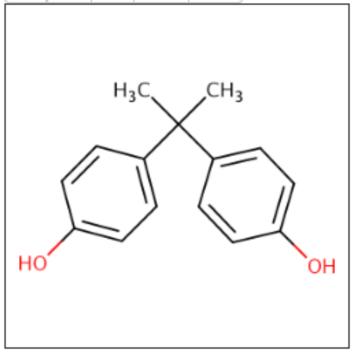
States
mental Protection | Home | Advanced Search | Lists

y Dashboard

Bisphenol A

80-05-7 | DTXSID7020182

Searched by CAS-RN: Found 1 result for '80-05-7'.



Wikipedia

Bisphenol A (BPA) is an organic synthetic compound with the chemical formula (CH3)2C(C6H4OH)2 belonging to the group of diphenylmethane derivatives and bisphenols, wi colorless solid that is soluble in organic solvents, but poorly soluble in water. It has been in commercial use since 1957. BPA is employed to make certain plastics and epoxy re tough... Read more

Intrinsic Properties

Structural Identifiers

Related Compounds (Beta)

Presence in Lists

Record Information

Chemical Properties | Synonyms | External Links | Env. Fate/Transport | Toxicity Values (Beta) | Bioassays | Exposure | Literature | Similar Molecules (Beta) | Comments

- Summary
- **LogP: Octanol-Water**
- Water Solubility
- Density
- Melting Point
- Boiling Point
- Surface Tension
- Vapor Pressure
- LogKoa: Octanol-Air
- Henry's Law
- Index of Refraction

LogP: Octanol-Water

	Average	Median	Range
Experimental	3.32 (1)	3.32	3.32
Predicted	3.24 (4)	3.24	2.40 to 3.73

Download as: TSV | Excel | SDF

Experimental

Source	Result
PhysPropNCCT	3.32

Predicted

Source	Result	Calculation Details	QMRF
EPISUITE	3.64	Not Available	Not Available
OPERA	3.21	OPERA Model Report	Available

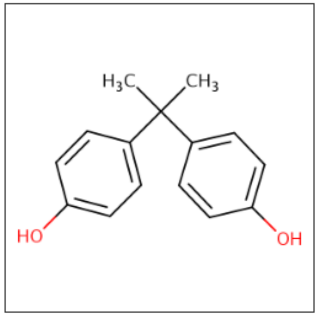
Predicted Data

https://comptox.epa.gov

OPERA Models: LogP: Octanol-Water

Bisphenol A

80-05-7 | DTXSID7020182



Model Results

Predicted value: 3.21

Global applicability domain: Inside

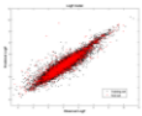
Local applicability domain index: 0.54

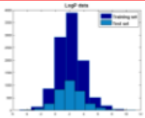
Confidence level: 0.79

Calculation Result for a chemical

Model Performance with full QMRF

Model Performance





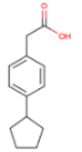
Weighted KNN model

QMRF

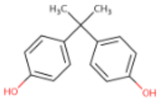
5-fold CV (75%)		Training (75%)		Test (25%)	
Q2	RMSE	R2	RMSE	R2	RMSE
0.85	0.69	0.86	0.67	0.86	0.78

Nearest Neighbors from the Training Set

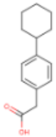
Nearest Neighbors from Training Set



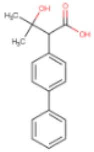
BENZENEACETIC ACID, 4-CYCLOPENTYL-
Measured: 3.44
Predicted: N/A



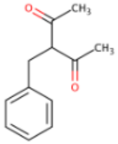
Bisphenol A
Measured: 3.32
Predicted: 3.21



4-Cyclohexylphenylacetic acid
Measured: 3.91
Predicted: 3.90

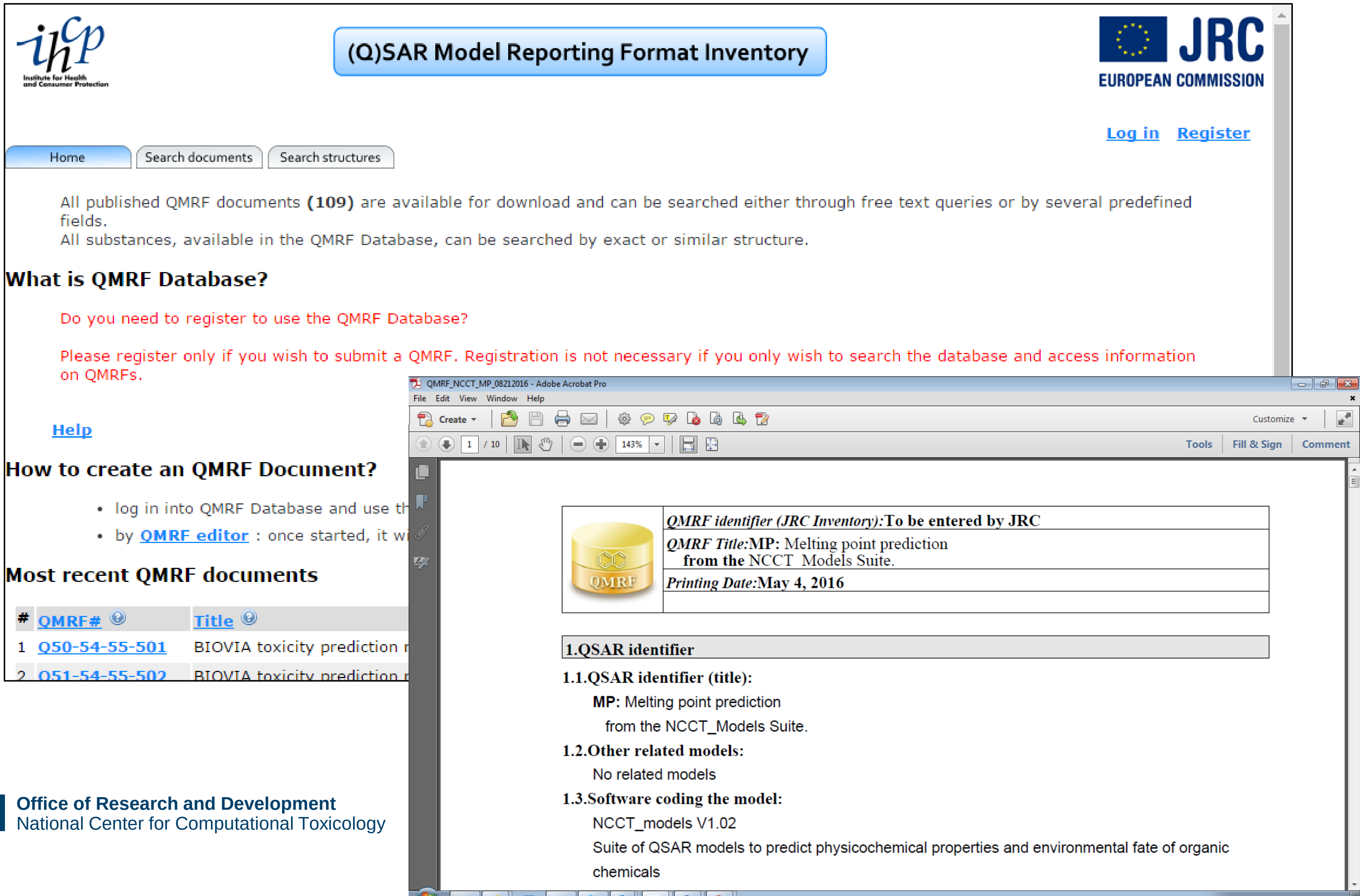


BUTANOIC ACID,2-(4-BIPHENYLYL)-3-HYDROXY-3-MET...
Measured: 3.25
Predicted: N/A



2,4-Pentanedione, 3-benzyl- (8CI)
Measured: 1.89
Predicted: 1.89

QMRF Reports



Conclusion

- QSAR prediction models (kNN) produced for all properties
- 700k chemical structures pushed through OPERA
- Supplementary data will include appropriate files with flags – full dataset plus QSAR ready form
- Full performance statistics available for all models
- OPERA Models will be deployed as prediction engines in the future – one chemical at a time and on the fly batch processing

Acknowledgements

EPA NCCT
Antony Williams
Imran Shah
Chris Grulke
Jeff Edwards
Ann Richard
Jordan Foster
Jennifer Smith
Richard Judson
Grace Patlewicz
John Wambaugh
Michelle Krzyzanowski

Thank you for your attention