The importance of data curation on QSAR Modeling: PHYSPROP open data as a case study

Kamel Mansouri,
Christopher Grulke,
Ann Richard,
Richard Judson
Antony Williams

NCCT, U.S. EPA



QSAR 2016

14 June 2016, Miami, FL

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 development at NCCT

- 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
  - RapidTox
  - Read-across
  - Predictive modeling
  - Non-targeted screening

Developing “NCCT 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

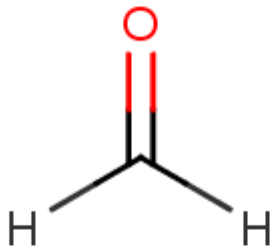
HENRYWIN Data files used in training & validation ... (includes Meylan and Howard (1991) Data document) - Download file is: HENRYWIN_Data_EPI.zip (531 K)

Click here to download HENRYWIN_Data_EPI.zip

- Water solubility
- Melting Point
- Boiling Point
- LogP (KOWWIN: Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Ready biodegradability
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOA (Octanol/Air Partition Coefficient)
- LogKOC (Soil Adsorption Coefficient)
- Vapor Pressure

Data Files

- 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

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

The Approach

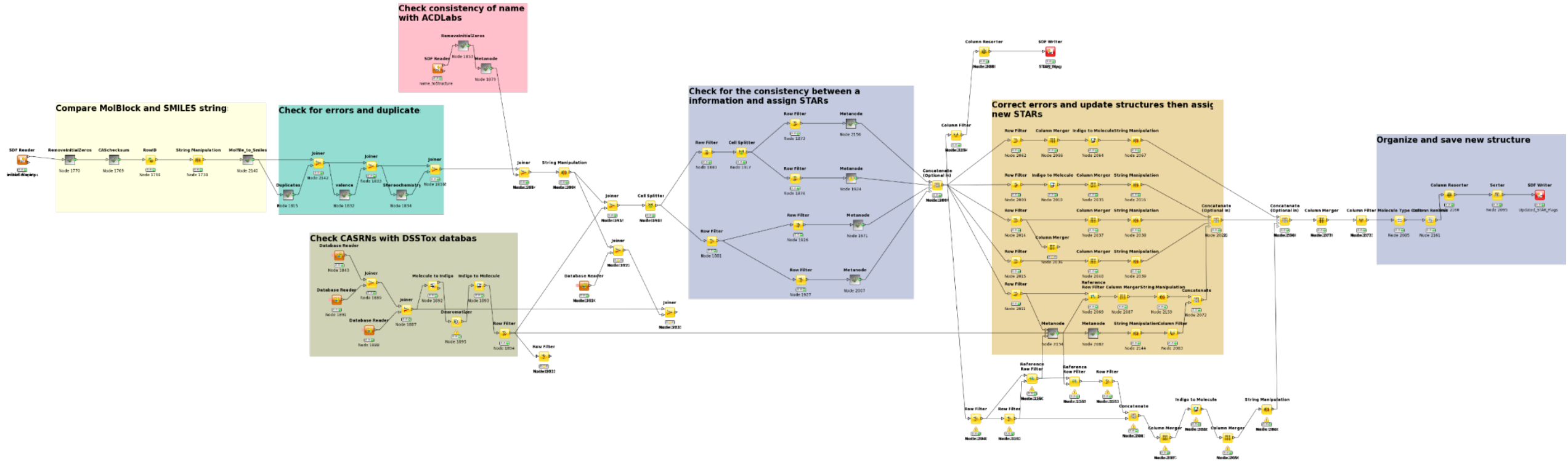
- To build models we need the set of chemicals and their property series

- 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

General Observations from LogP dataset

- CAS Numbers not matching structure
- Some SMILES won’t convert (non-standard SMILES)
- Valence and charge imbalance issues
- Stereochemistry poorly depicted if not totally absent
- Multiple duplicate pairs for a particular chemical compound
- Majority of duplicates from structure representations not matching the chemical.

KNIME workflow to evaluate the dataset



LogP dataset: 15,809 chemicals (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%)**

United States Environmental Protection Agency

Invalid CASRNs

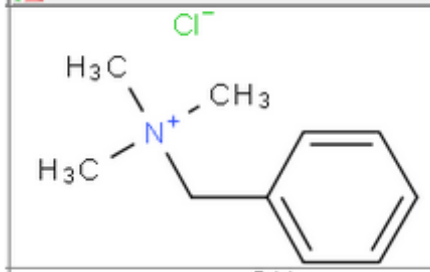
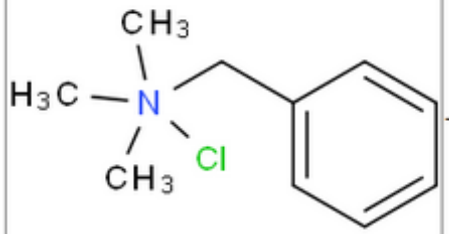
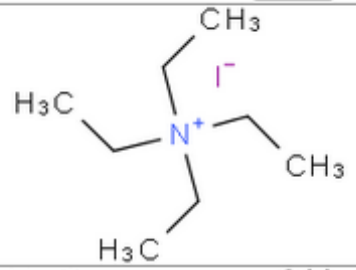
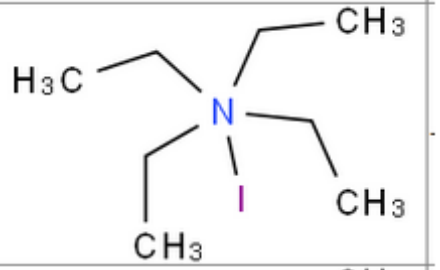
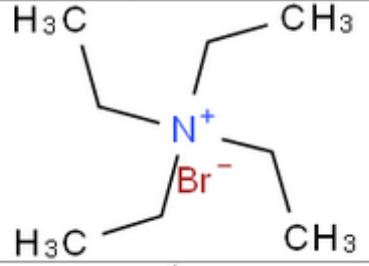
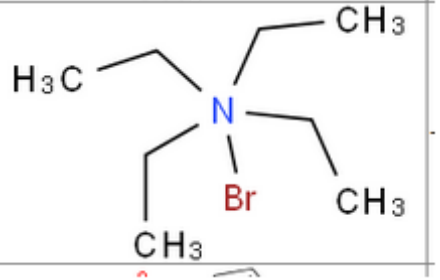
Truncated names

Missing SMILES

SRC000-01-7	Ethanaminium, 2-[(chloroacetyl)oxy]-N,N,N-trimet
SRC000-02-3	2-Trifluoroethane-picrate
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

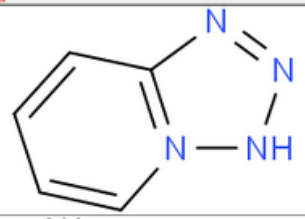
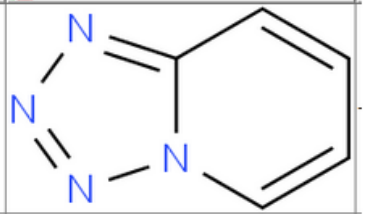
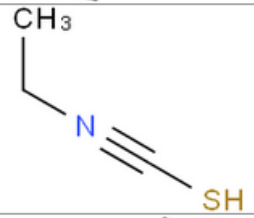
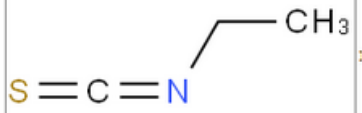
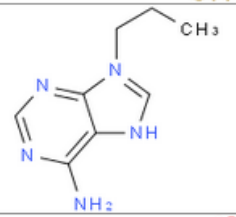
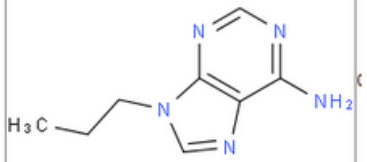
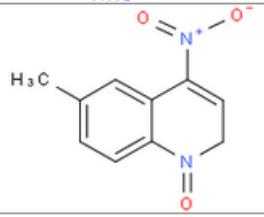
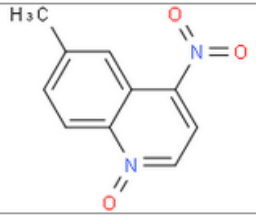
Examples of errors

- 362 Halogens bonded to nitrogen

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

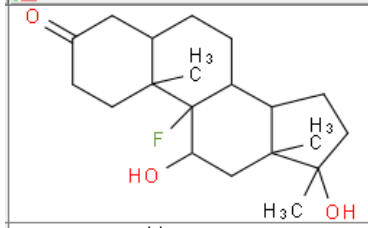
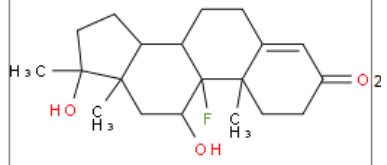
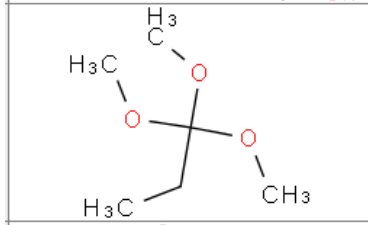
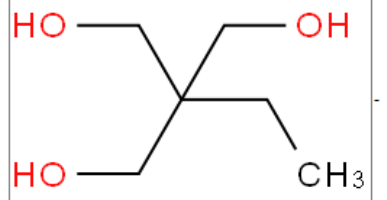
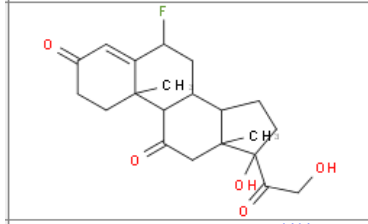
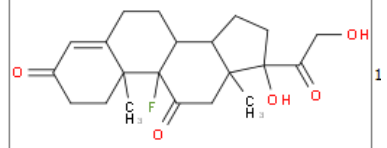
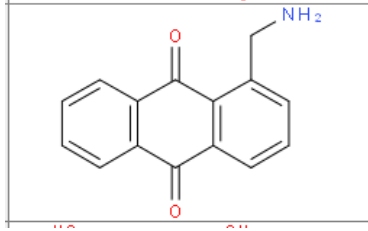
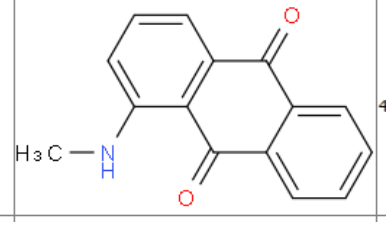
Examples of errors

- 191 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	

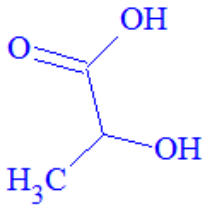
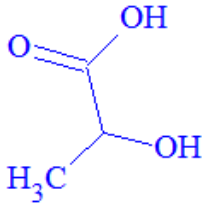
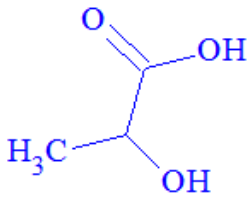
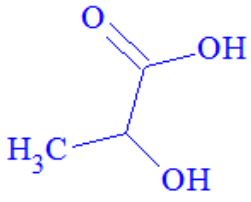
Examples of errors

- 463 completely 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	1.68000000000000e+001	2.26600000000000e+001	5.86000000000000e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.30000000000000e+001	2.26600000000000e+001	-3.03400000000000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.80000000000000e+001	2.26600000000000e+001	4.66000000000000e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.28000000000000e+001	2.26600000000000e+001	-3.01400000000000e+001

Quality flags: 1-4 STARs

4 levels of consistency exists between:

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

Quality FLAGS into LogP data

- 4 Stars ENHANCED: 4 levels of consistency with stereo information
- 4 Stars: 4 levels of consistency, stereo ignored.
- 3 Stars Plus: 3 out of 4 levels. The 4th is a tautomer.
- 3 Stars ENHANCED: 4 levels of consistency with stereo information
- 3 Stars: 3 levels of consistency, stereo ignored.
- 2 Stars PLUS : 2 out of 4 levels. The 3th is a tautomer.
- 1 Star - What's left.

Improved structures and updated flags

- 3 STAR and 2 STAR Plus are "upgraded" to a higher level of consistency

Done by correcting the mismatching field(s), or by generating a name or smiles string when missing or unreadable.

- 3 STAR to 4 Star:
- Available: Molblock, Name, CAS: Smiles generated from Molblock (DSSTOX)
- Available: Molblock, Smiles, CAS: Name retrieved from DSSTOX
- Available: Name, Smiles, CAS: Molblock retrieved from DSSTOX
- Available: Molblock, Smiles, Name: CAS retrieved when available in DSSTOX (no stereoisomers)

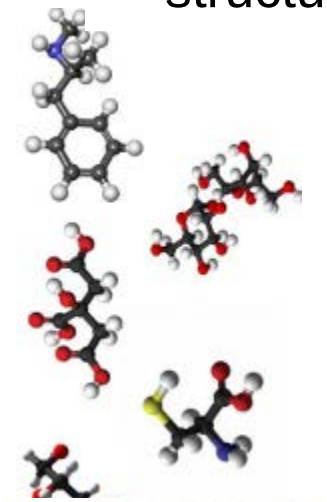
- 2 Star Plus with Unreadable Smiles, name or CAS

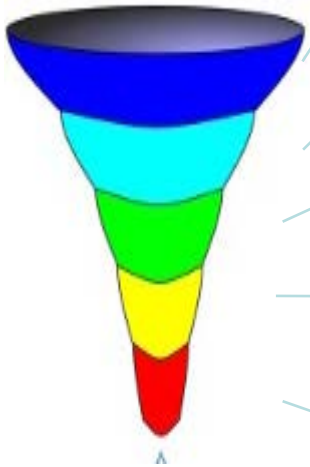
- Total upgraded chemicals for LogP data: **1740 chemicals**
- Total chemicals with 3 STAR levels of consistency for LogP data: **7910 chemicals**
- Total chemicals with 4 STAR levels of consistency for LogP data: **6525 chemicals**

Only part considered For QSAR

EPA United States Environmental Protection Agency

Initial structures





QSAR-ready structures

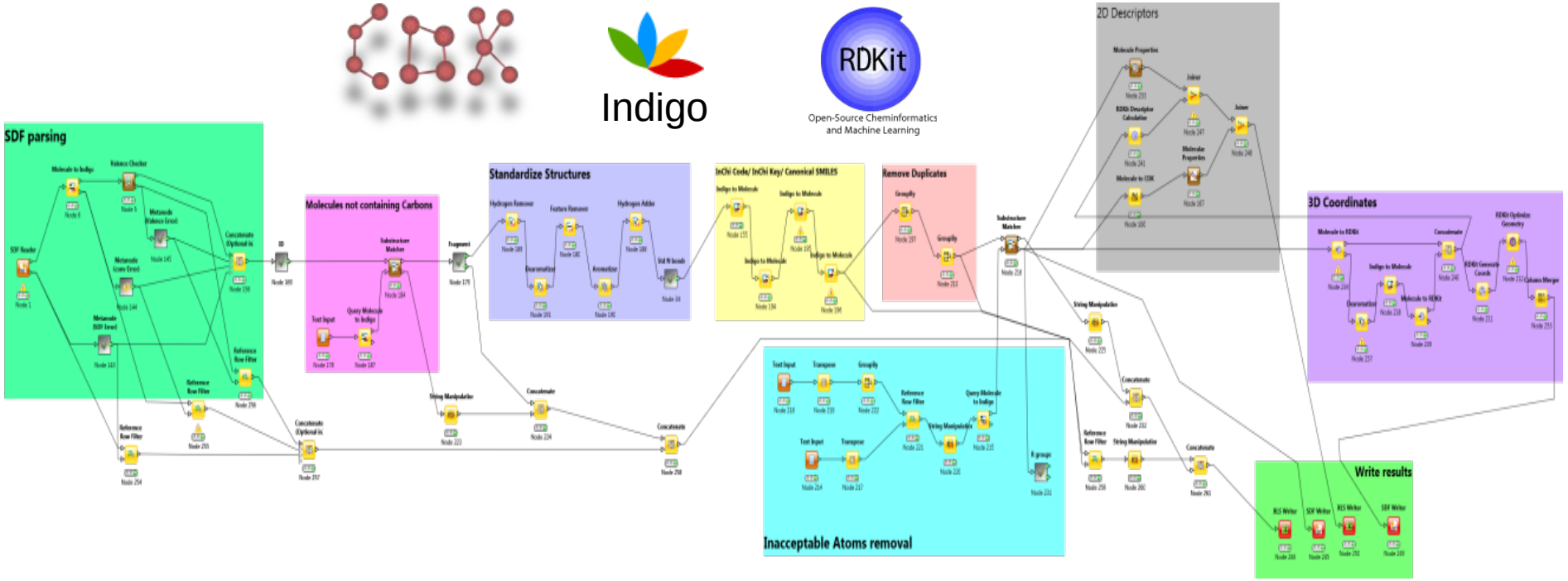
- Remove inorganics and mixtures
- Cleaning salts and counterions
- Normalization of tautomers
- Removal of duplicates
- Final inspection

Structure standardization

KNIME workflow
UNC, DTU, EPA Consensus

Aim of the workflow:

- Combine (not reproduce) different procedures and ideas
- Minimize the differences between the structures used for prediction by different groups
- Produce a flexible free and open source workflow to be shared



Mansouri et al. (http://ehp.niehs.nih.gov/15-10267/)

United States Environmental Protection Agency

Summary:

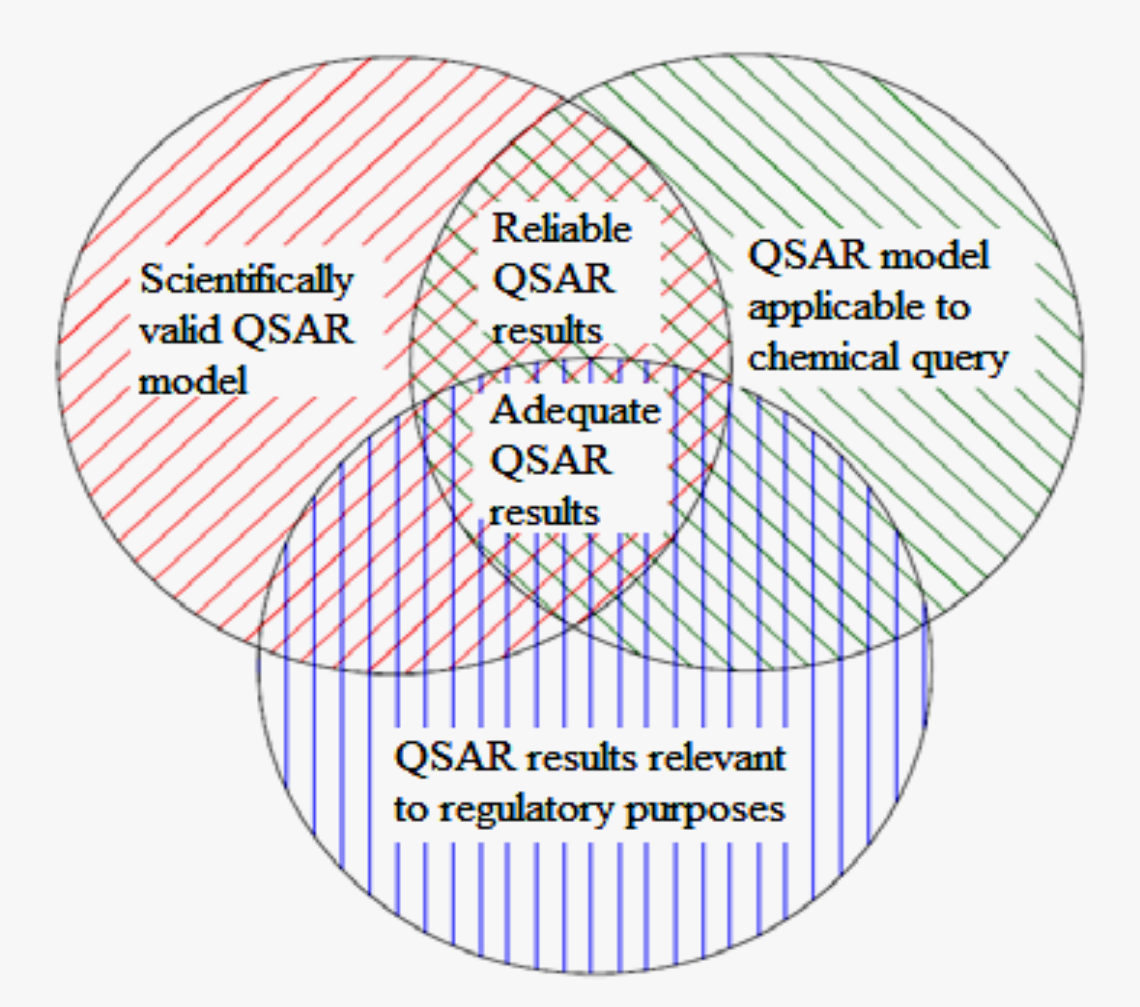
Property	Initial file flagged	Updated 3-4 STAR	Curated QSAR ready
AOP	818	818	745
BCF	685	618	608
BioHC	175	151	150
Biowin	1265	1196	1171
BP	5890	5591	5436
HL	1829	1758	1711
KM	631	548	541
KOA	308	277	270
LogP	15809	14544	14041
MP	10051	9120	8656
PC	788	750	735
VP	3037	2840	2716
WF	5764	5076	4836
WS	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.



The conditions for the validity of QSARs

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**.

EPA United States Environmental Protection Agency

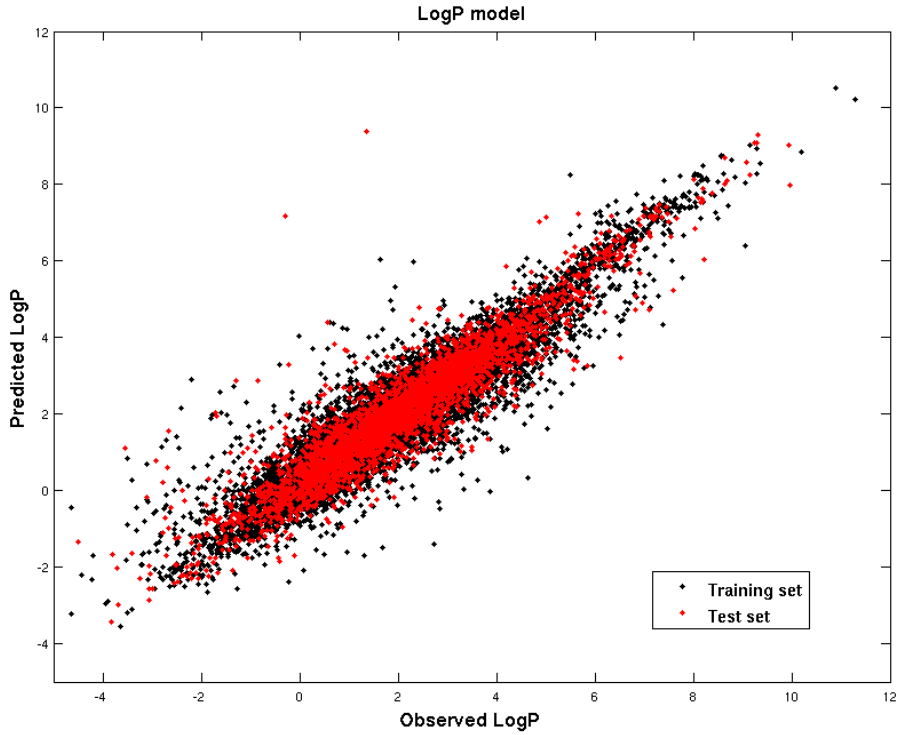
NCCT models

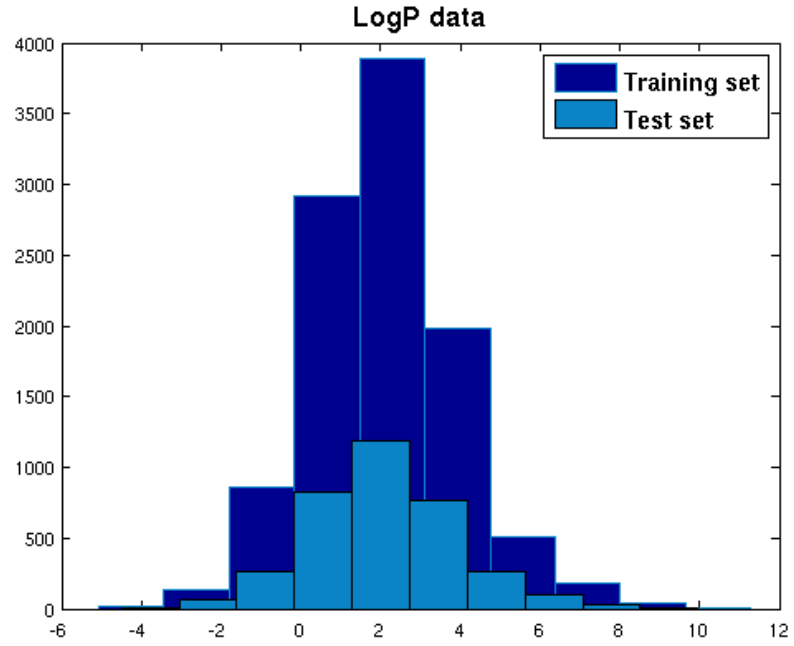
Prop	Vars	5-fold CV (75%)		Training (75%)			Test (25%)		
		Q2	RMSE	N	R2	RMSE	N	R2	RMSE
BCF	10	0.84	0.55	465	0.85	0.53	161	0.83	0.64
BP	13	0.93	22.46	4077	0.93	22.06	1358	0.93	22.08
LogP	9	0.85	0.69	10531	0.86	0.67	3510	0.86	0.78
MP	15	0.72	51.8	6486	0.74	50.27	2167	0.73	52.72
VP	12	0.91	1.08	2034	0.91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

NCCT 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 Cross-validation:
Q2=0.85 RMSE=0.69
Fitting:
R2=0.86 RMSE=0.67
Test:
R2=0.86 RMSE=0.78

Standalone application:

Input:

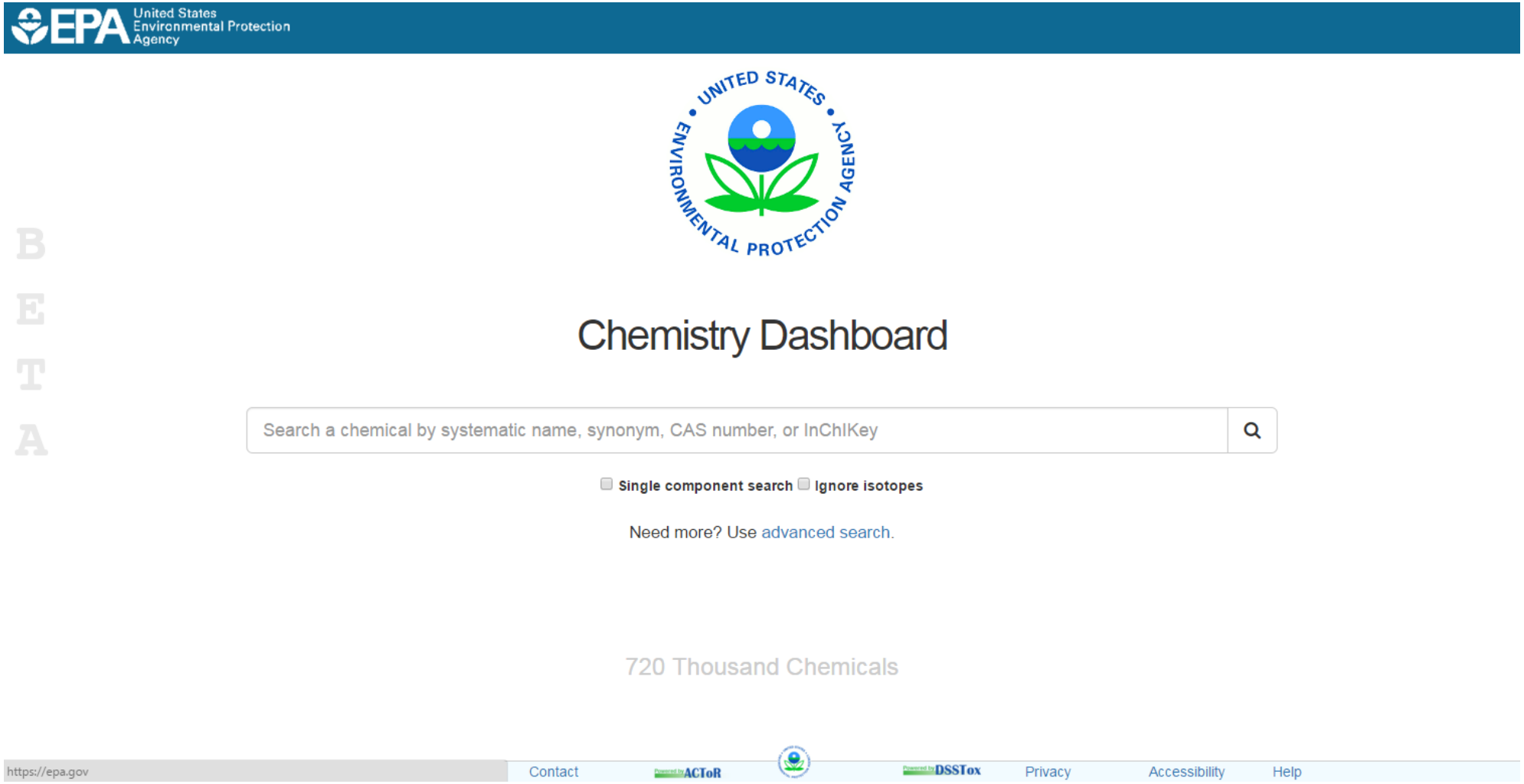
- MATLAB .mat file, an ASCII file with only a matrix of variables
- SDF file or SMILES strings of QSAR-ready structures. In this case the program will calculate PaDEL 2D descriptors and make the predictions.

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

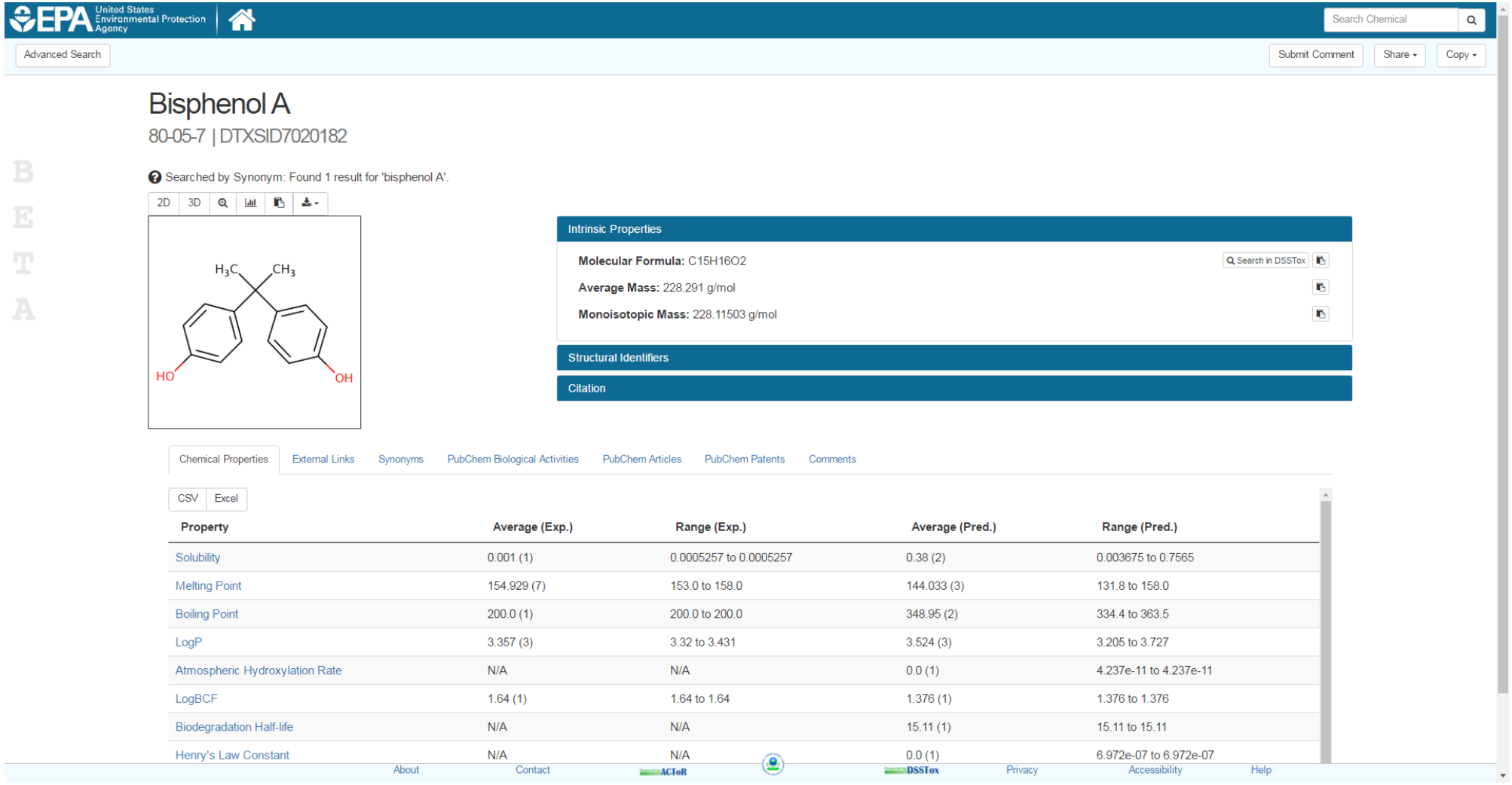
Output

- 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

The iCSS Chemistry Dashboard at https://comptox.epa.gov



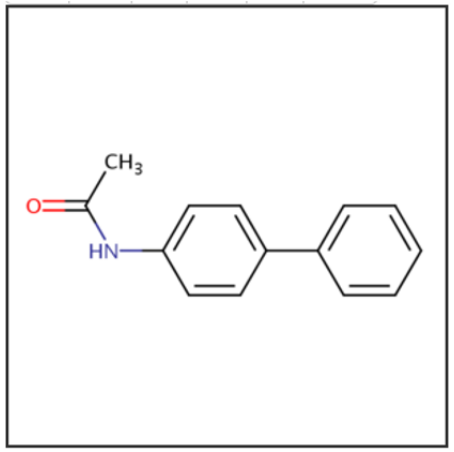
The iCSS Chemistry Dashboard



EPA United States Environmental Protection Agency

4-Acetylaminobiphenyl

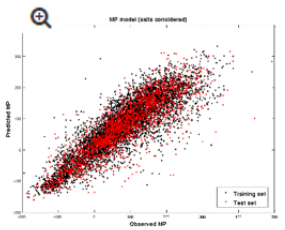
4075-79-0 | DTXSID8039243

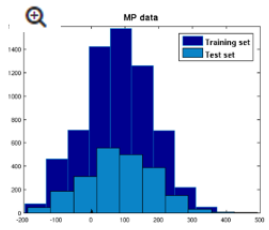


NCCT Models: Melting point (MP)

Save Report

Model Performance





QMRF

Weighted KNN model
15 molecular descriptors

5-fold CV (75%)		Training (75%)			Test (25%)		
Q2	RMSE	N	R2	RMSE	N	R2	RMSE
0.72	51.8	6486	0.74	50.27	2167	0.73	52.72

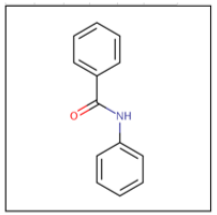
Model Results

Predicted value: 143 °C
Observed value in training set: Not available
Global applicability domain: Inside the AD
Local applicability domain index: 0.88
Confidence level: 0.70

5 nearest neighbors from the training set

Benzanilide

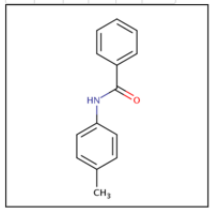
93-98-1 | DTXSID9059096



Observed: 163 C
Predicted: 141 C

4'-Methylbenzanilide

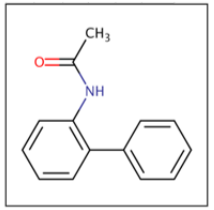
582-78-5 | DTXSID20206946



Observed: 158 C
Predicted: 141 C

2-Acetamidobiphenyl

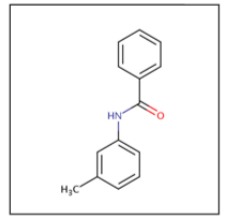
2113-47-5 | DTXSID6036837



Observed: 121 C
Predicted: 150 C

3'-Methylbenzanilide

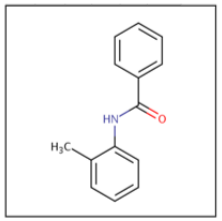
582-77-4 | DTXSID60206945



Observed: 125 C
Predicted: 150 C

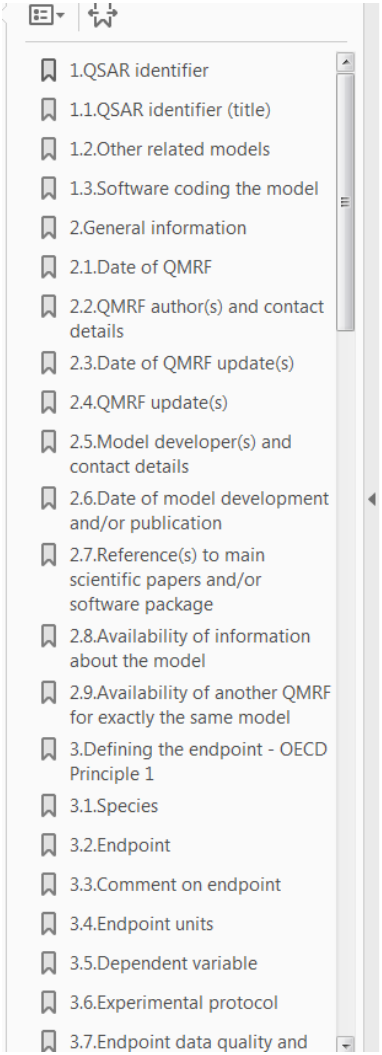
2'-Methylbenzanilide

584-70-3 | DTXSID00207113



Observed: 145 C
Predicted: 143 C

QMRF for LogP model



QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **LogP:** Octanol-water partition coefficient prediction from the NCCT_Models Suite.
Printing Date: **May 4, 2016**

1.QSAR identifier

1.1.QSAR identifier (title):

LogP: Octanol-water partition coefficient prediction from the NCCT_Models Suite.

1.2.Other related models:

No related models

1.3.Software coding the model:

NCCT_models V1.02

Suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals

Kamel Mansouri (mansouri.kamel@epa.gov; mansourikamel@gmail.com);

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

PaDEL descriptors V2.21

Open source software to calculate molecular descriptors and fingerprints.

Chun Wei Yap (phayapc@nus.edu.sg)

Conclusion

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

Acknowledgements

National Center for Computational Toxicology



Thank you for your attention

