Automated workflows for data curation and standardization of chemical structures for QSAR modeling

American Chemical Society meeting
New Orleans, LA

March 19, 2018

Kamel Mansouri: ScitoVation LLC
Andrew McEachran, NCCT/ US EPA (ORISE).
Christopher Grulke, Ann Richard, 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

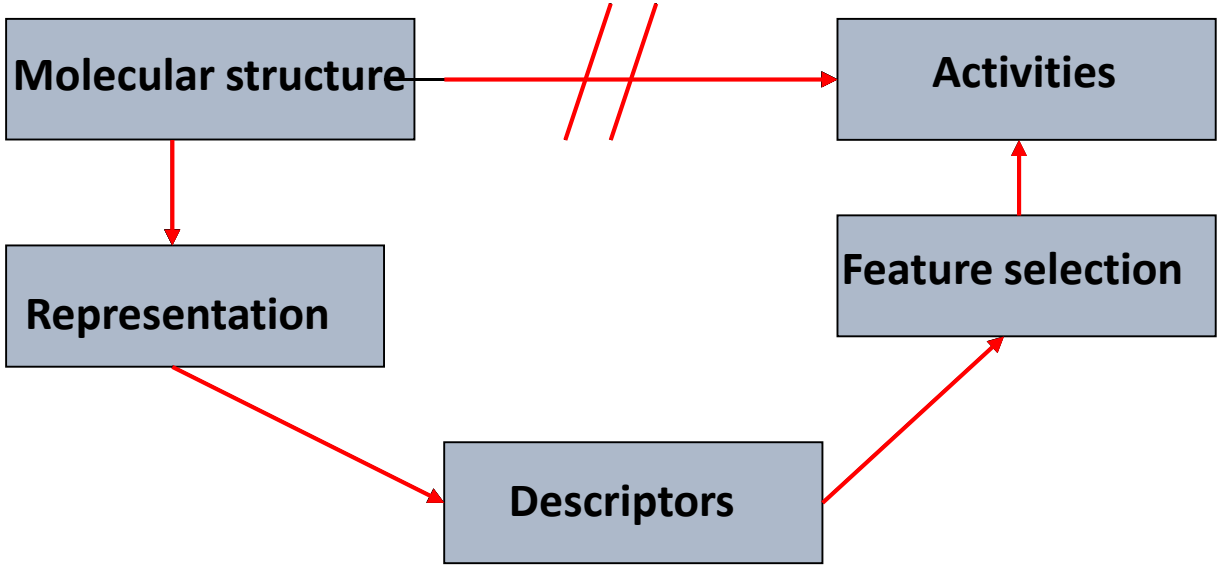
Kamel Mansouri, PhD
Research Investigator
Computational Chemistry
919-558-1356
kmansouri@scitovation.com

ScitoVation
INNOVATIVE CELL BASED SCIENCE
www.scitovation.com

Quantitative Structure Activity/Property Relationships (QSAR/QSPR)

QSARs correlate, within congeneric series of compounds, their chemical or biological activities, either with certain structural features or with atomic, group or molecular descriptors.

Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Chem. Soc. Rev. ***1995****, 279-287*



Development of a QSAR model

- Curation of experimental data
- **Generation of QSAR-ready structures**
- Preparation of training and test sets
- Calculation of an initial set of descriptors
- Selection of a mathematical method
- Variable selection technique
- Validation of the model's predictive ability
- Define the Applicability Domain

Chemical structures online

ChemSpider
The free chemical database
About | More Searches | Web APIs

chemexper.com
Enter a name, molecular formula, cas number, InChI, InChIKey or SMILES
Search

United States National Library of Medicine
ChemIDplus Advanced
News | SIS Home | Site | About Us

Revision 7464 by midnighter checked in on 2012-05-29 19:03:16. Built from
Online chemical database
with modeling environment
Home | Database | Models
Compounds properties browser
Search for numerical compounds properties linked to scientific articles

ECOTOX Database
Recent Additions | Contact Us | Search: All EPA
You are here: EPA Home » ECOTOX > Data Download >
Data Downloads
Download the Adobe® Reader®

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
Chemistry Dashboard

PubChem
BioAssay | Compound | Substance
GO | Advanced Search

PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed | pubmed id
RSS

OpenTox
Home | Applications | Downloads
About | Reading Room | FP7 | REACH

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

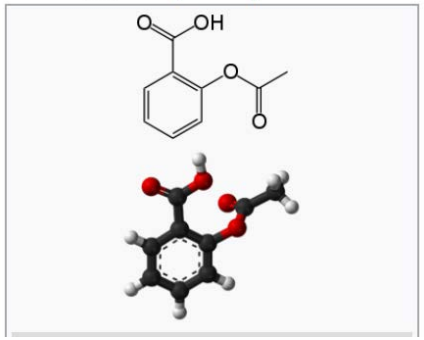
HENRYWIN Data files used in training & validation ... (includes Meylan and Howard (1991) Data document) - Download file is: HENRYWIN_Data_EPI.zip (531 K)

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

Chemical Identifiers

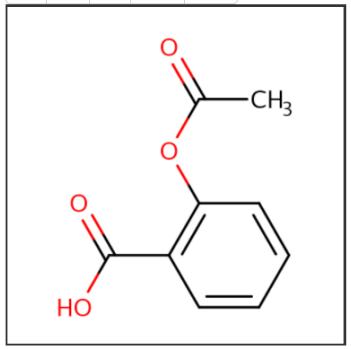
- CAS Registry Number
- Names:
  - IUPAC
  - Commercial
  - Synonyms…
- SMILES
- InChI
- InChI keys
- …



Identifiers		
IUPAC name		[show]
CAS Number	50-78-2 ✓	
PubChem CID	2244	
IUPHAR/BPS	4139	
DrugBank	DB00945 ✓	
ChemSpider	2157 ✓	
UNII	R16CO5Y76E	
KEGG	D00109 ✓	
ChEBI	CHEBI:15365 ✓	
ChEMBL	CHEMBL25 ✓	
PDB ligand	AIN (PDBe, RCSB PDB)	
ECHA InfoCard	100.000.059	

Aspirin
50-78-2 | DTXSID5020108

Searched by Approved Name: Found 1 result for 'Aspirin'.



Wikipedia

Intrinsic Properties

Structural Identifiers

IUPAC Name: 2-(Acetyloxy)benzoic acid

SMILES: CC(=O)OC1=C(C=CC=C1)C(O)=O

InChI String: InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

InChIKey: BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Search Google for: Structural Skeleton | Full Structure

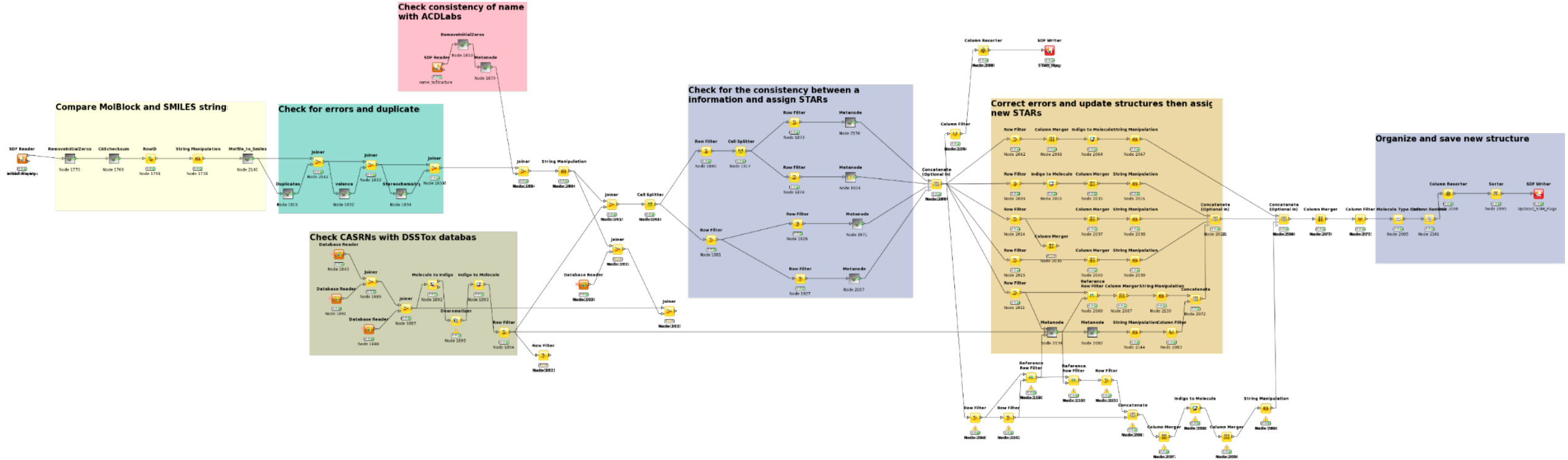
Copy All

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

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

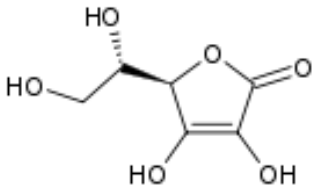
International Chemical Identifier

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

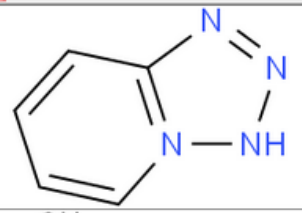
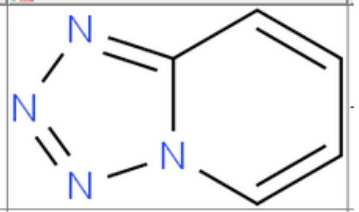
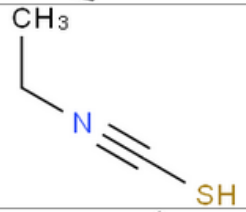
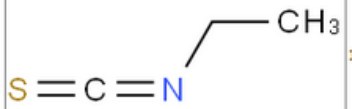
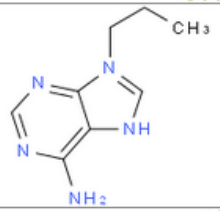
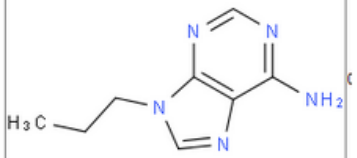
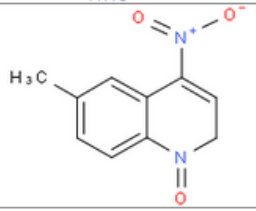
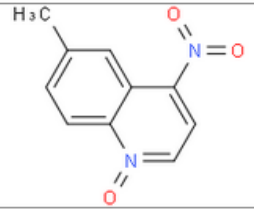
`InChI=1/C6H8O6/c7-1-2(8)5-3(9)4(10)6(11)12-5/h2,5,7-10H,1H2/t2-,5+/m0/s1`

LogP dataset: 15,809 chemicals (structures)

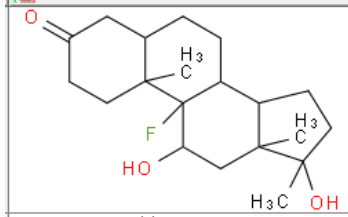
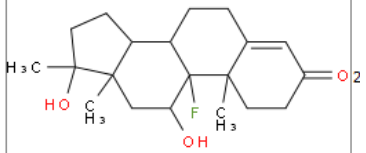
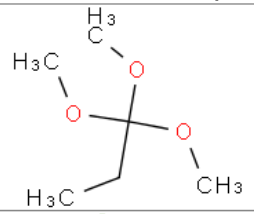
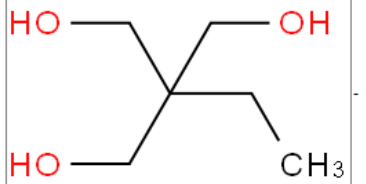
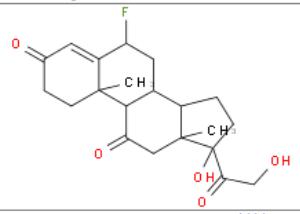
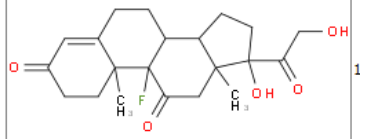
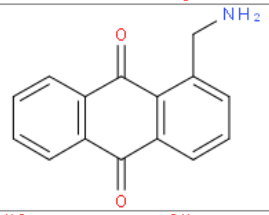
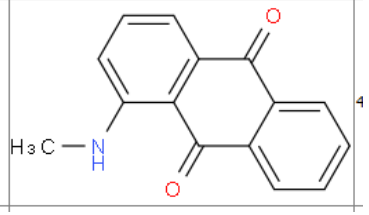
- CAS Checksum: 12163 valid, 3646 invalid **(>23%)**
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES **(>24%)**
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in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

Examples of Errors

Valence Errors

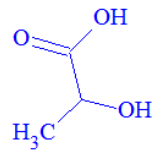
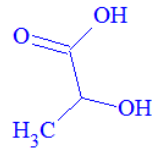
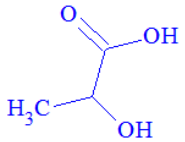
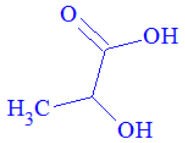
Mol Block	CAS	NAME	Smiles
	000274-87-3	TETRAZOLO[1,5-A]PYRIDINE	
	000542-85-8	ETHYL ISOTHIOCYANATE	
	000707-98-2	9-PROPYL ADENINE	
	000715-48-0	6-METHYL-4-NITROQUINOLINE-1-OXIDE	

Different Compounds

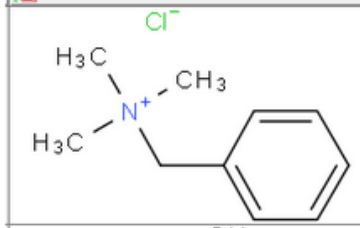
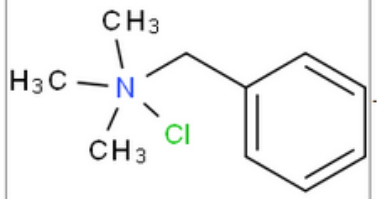
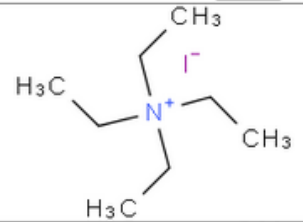
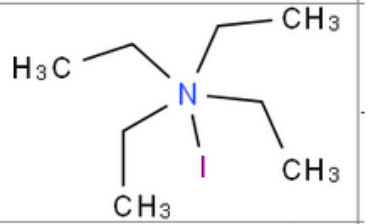
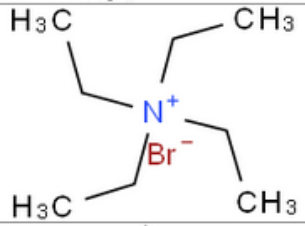
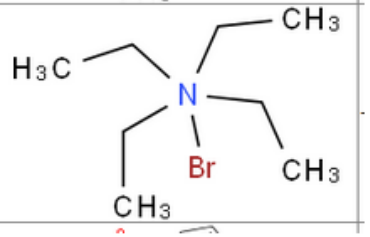
Mol Block	CAS	NAME	Smiles
	000076-43-7	FLUOXYMESTERONE	
	000077-99-6	1,1,1-TRIS(HYDROXYMETHYL)PROPANE	
	000079-60-7	CORTISONE-9A-FLUORO	
	000082-38-2	DISPERSE RED 9	

Examples of Errors

Duplicate Structures

Structure	Formula	FW	CAS	NAME	MP	EstMP	ErrorMP
	$C_3H_6O_3$	90.0779	000050-21-5	LACTIC ACID	1.6800000000000000e+001	2.2660000000000000e+001	5.8600000000000000e+000
	$C_3H_6O_3$	90.0779	000079-33-4	L-LACTIC ACID	5.3000000000000000e+001	2.2660000000000000e+001	-3.034000000000000e+001
	$C_3H_6O_3$	90.0779	000598-82-3	A-HYDROXYPROPIONIC ACID	1.8000000000000000e+001	2.2660000000000000e+001	4.6600000000000000e+000
	$C_3H_6O_3$	90.0779	010326-41-7	D-LACTIC ACID	5.2800000000000000e+001	2.2660000000000000e+001	-3.014000000000000e+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

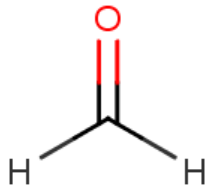
Truncated names

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)

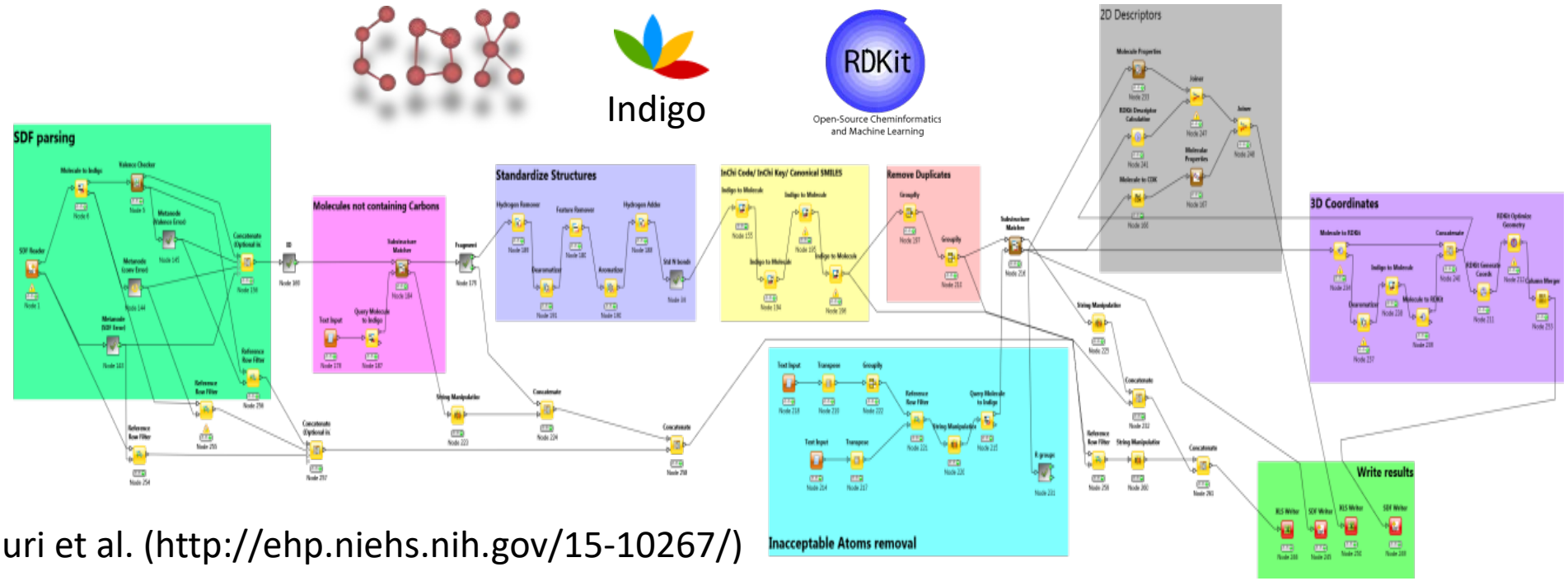
http://esc.syrres.com/interkow/EpiSuiteData.htm

Quality FLAGS and curated structures

QSAR-ready KNIME workflow

Aim of the workflow:

- Combine standardization procedures
- Minimize the differences between the structures used for prediction
- Produce a flexible free and open source workflow to be shared

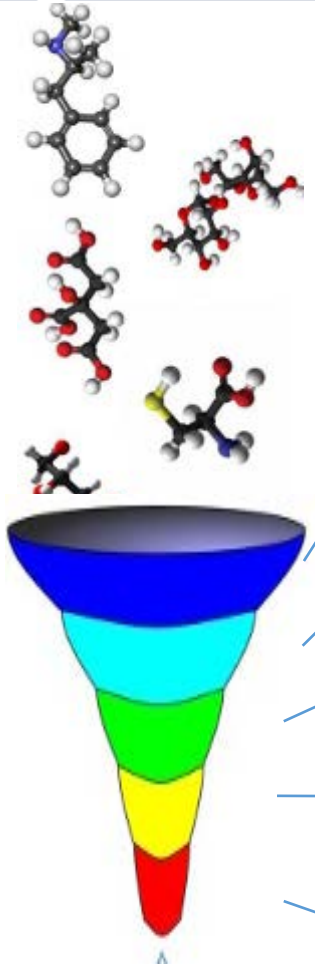


Mansouri et al. (http://ehp.niehs.nih.gov/15-10267/)

Standardization steps

Initial structures

Structure standardization



QSAR-ready structures

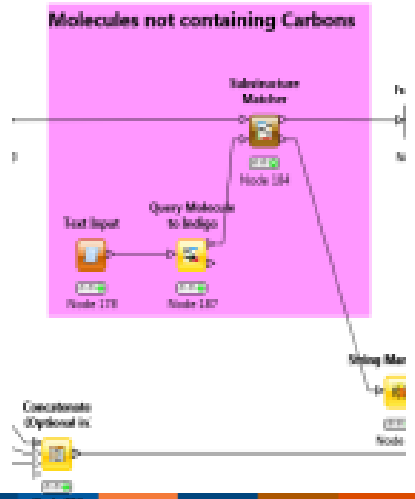
- Removal of inorganics and mixtures
- Cleaning salts and counter ions
- Normalization of tautomers
- Removal of duplicates
- Final inspection

Parsing and 1st filter

SDF Parser: original structures

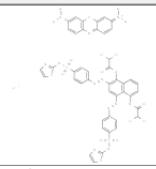
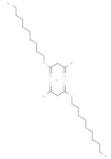
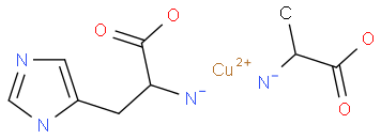
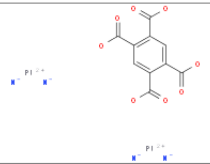
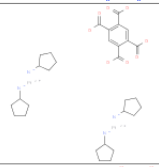
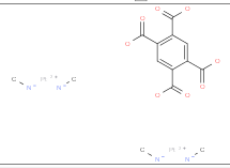
(Webservices: Pubchem, Chemspider)

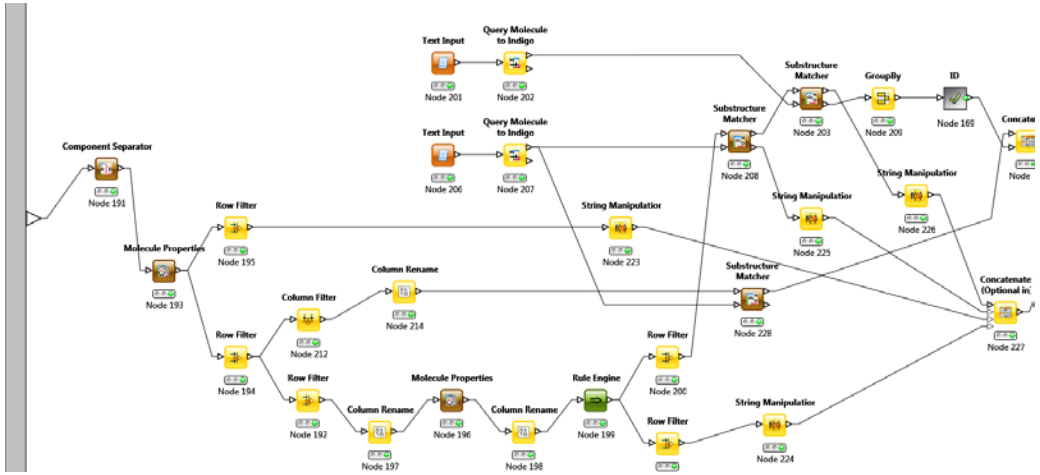
parsed compounds
Unique IDs



Errors reported

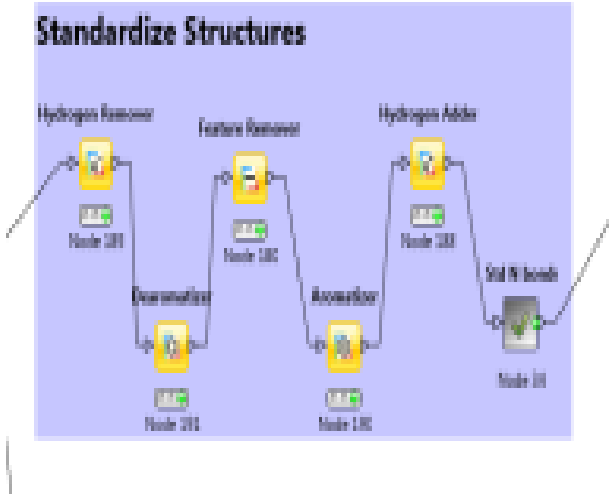
Unconnected structures (salts, solvents, mixtures)

Mol Block	gsid	dsstox_subs...	casrn	preferred_name
	639220	DTXSID70639220	74352-04-8	7-(Dimethylamino)...
	639221	DTXSID30639221	22757-23-9	PUBCHEM_24198890
	639250	DTXSID20639250	73655-89-7	Copper(2+) 1-carb...
	639266	DTXSID30639266	82422-15-9	Platinum(2+) azani...
	639267	DTXSID90639267	82422-14-8	Platinum(2+) cyclo...
	639268	DTXSID50639268	82422-13-7	Platinum(2+) meth...



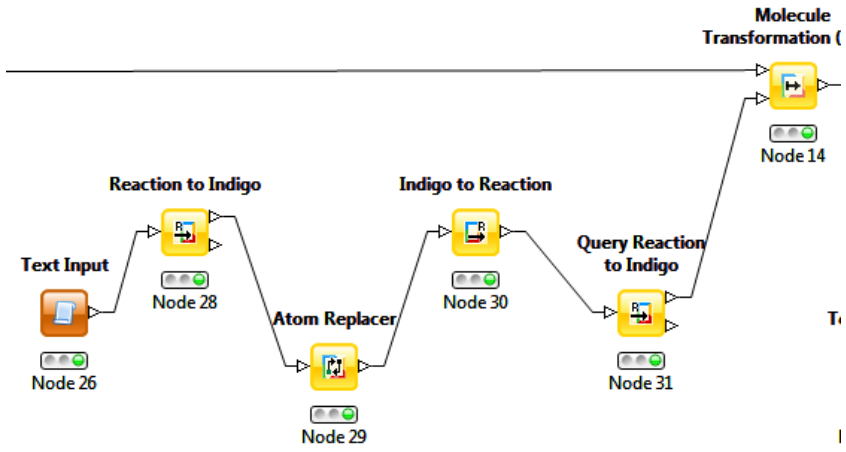
Standardization of structures

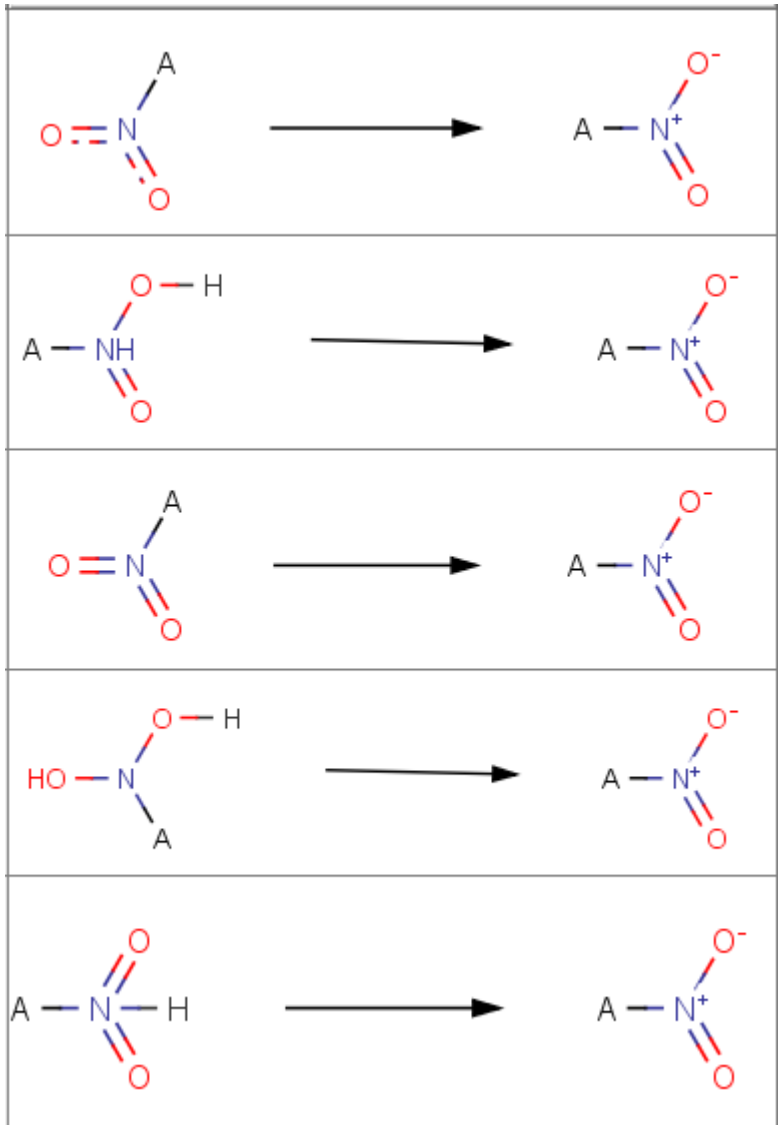
- Explicit hydrogen removed
- Dearomatization
- Removal of chirality/stereochemistry info, isotopes and pseudo-atoms
- Aromatization + add explicit hydrogen atoms
- Standardize Nitro groups
- Other: tautomerize/mesomerize
- Neutralize (when possible)



Standardize Nitro mesomers

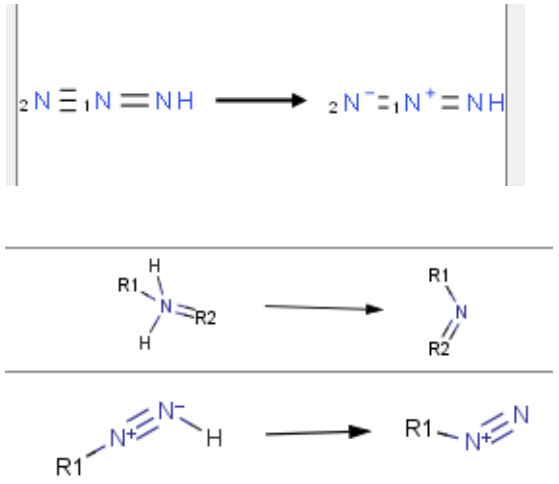
SMARTS query to reaction

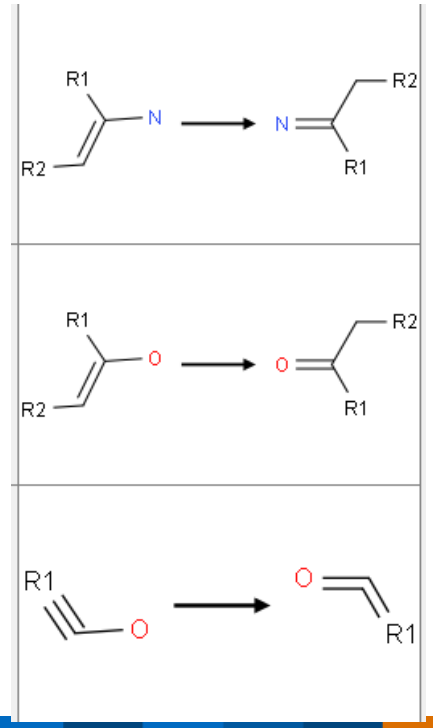


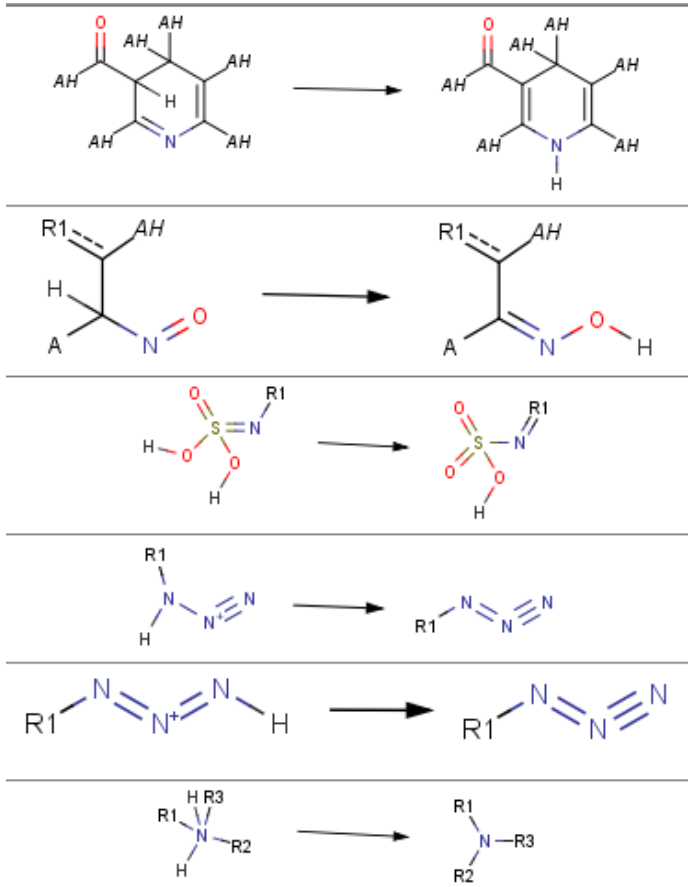


Mesomerization/tautomerization

- Azide mesomers
- Exo-enol tautomers
- Enamine-Imine tautomers
- Ynol-ketene tautomers
- ....



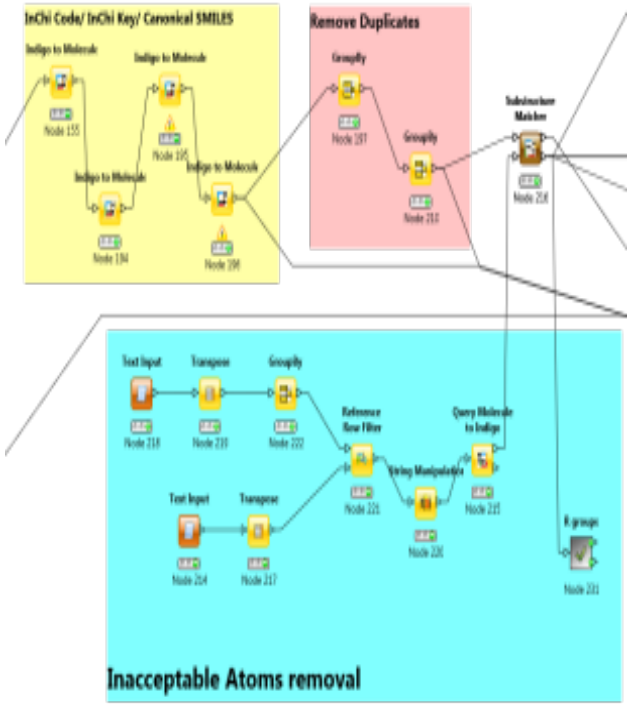




Neutralize Structures

Filter metalorganics

- Generate InChI, InChI Key and Canonical Smiles.
- Remove duplicates (InChIs & canonical SMI)
- Remove molecules with atoms. Other then:
  H, C, N, O, P, S, Se, F, Cl, Br, I, Li, Na, K, B, Si

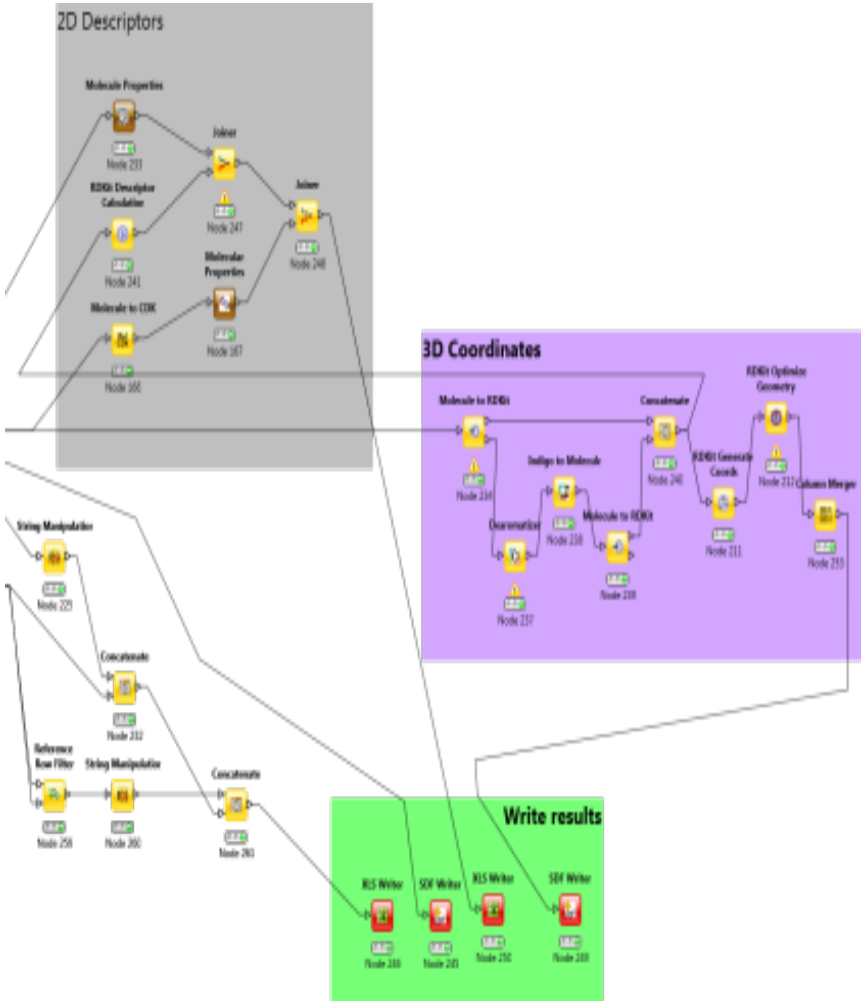


Write results

- Calculate 2D descriptors (Indigo, CDK, RDKit)
- Generate 3D conformers
- Optimize geometry (MMFF94S)

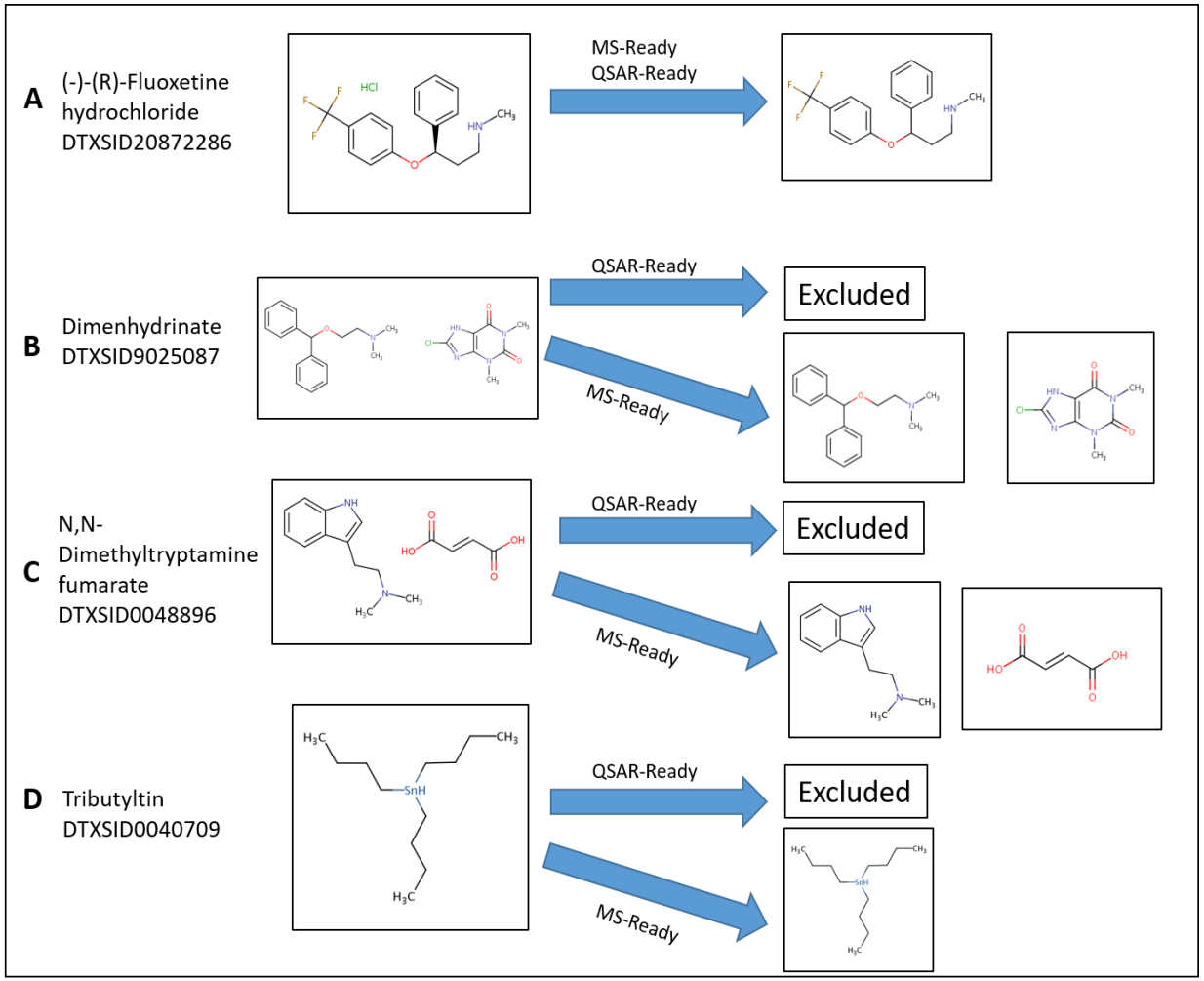
<u>Generated files:</u>

- Sdf file containing the 2D structures
- Excel file containing 2D descriptors
- Sdf file containing the 3D structures
- Excel file for error messages

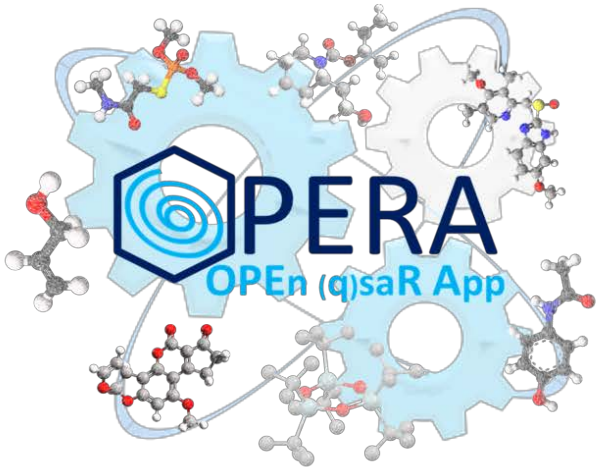


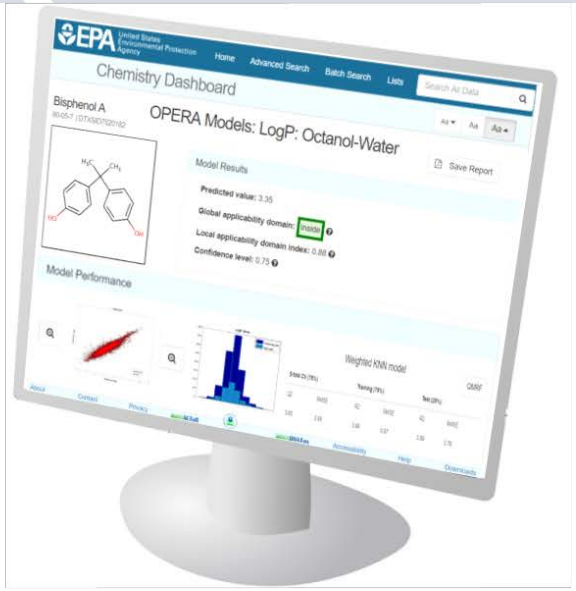
“MS-Ready” structures

- The related QSAR-Ready workflow was modified to produce an MS-ready workflow



Applications: OPERA models, Endocrine disruption screening…





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

CERAPP

Collaborative Estrogen Receptor Activity Prediction Project

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

CoMPARA

Collaborative **M**odeling **P**roject for **A**ndrogen **R**eceptor **A**ctivity

Mansouri et al. (https://www.researchgate.net/profile/Mansouri_Kamel)

Thank you for your attention

Question OR Comment