Free online access to experimental and predicted chemical properties through the EPA's CompTox Chemistry Dashboard

Kamel Mansouri[1,2], Christopher Grulke[2], Richard Judson[2] and Antony J. Williams[2]
1. Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC
2. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC

PID: 2655845
American Chemical Society Fall Meeting
San Francisco, CA.
April 2-6, 2017

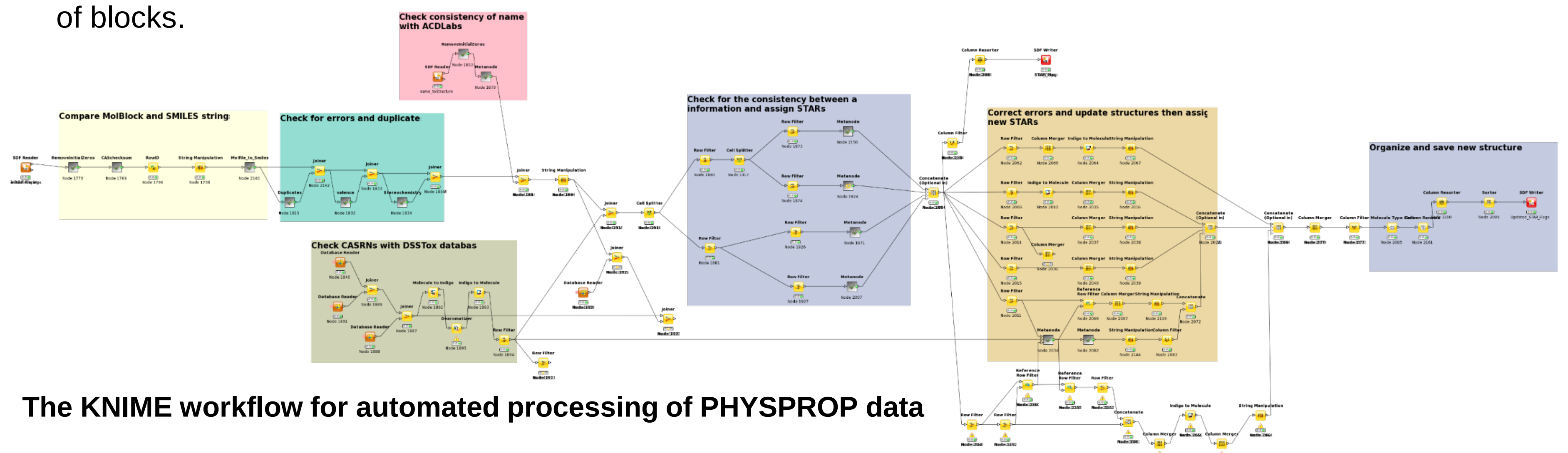
ORCID: 0000-0002-6426-8036
Kamel Mansouri I Mansouri.kamel@epa.gov

Abstract

The increasing number and size of public databases is facilitating the collection of chemical structures and associated experimental data for QSAR modeling. However, the performance of QSAR models is highly dependent not only on the modeling methodology, but also on the quality of the data used. In this study we developed robust QSAR models for endpoints of environmental interest with the aim of helping the regulatory process. We used the publicly available PHYSPROP database that includes a set of thirteen common physicochemical and environmental fate properties, including logP, melting point, Henry's coefficient, and biodegradability among others. Curation and standardization workflows have been applied to use the highest quality data and generate QSAR-ready structures. The developed models are in agreement with the five OECD principles that requires QSARs to be simple and reliable. These models were implemented in a command line application called OPERA (**Ope**n structure-activity **R**elationships **A**pplication) and applied to a set of ~700k chemicals to produce predictions for display on the EPA CompTox Chemistry Dashboard. In addition to the predictions, this free web application provides access to the experimental data used for training as well as detailed reports including general model performances, specific applicability domain and prediction accuracy, and the nearest neighboring structures used for prediction. The dashboard also provides access to model QMRFs (QSAR modeling report format) which is a downloadable pdf containing additional details about the modeling approaches, the data, and molecular descriptor interpretation.

Source Data and curation

- The PHYSPROP data were sourced online[1] as SDF files. A total of 13 endpoints were represented, including logP, water solubility, melting point, boiling point, and others. Each data point included the Mol-Block, SMILES, CASRN, Name, property value and, where available, a reference.
- A manual investigation of the data allowed us to develop a KNIME[2] workflow for automated processing. This workflow was derived from earlier work by Mansouri *et. al.*[3] and is represented in the figure below as a series of blocks.



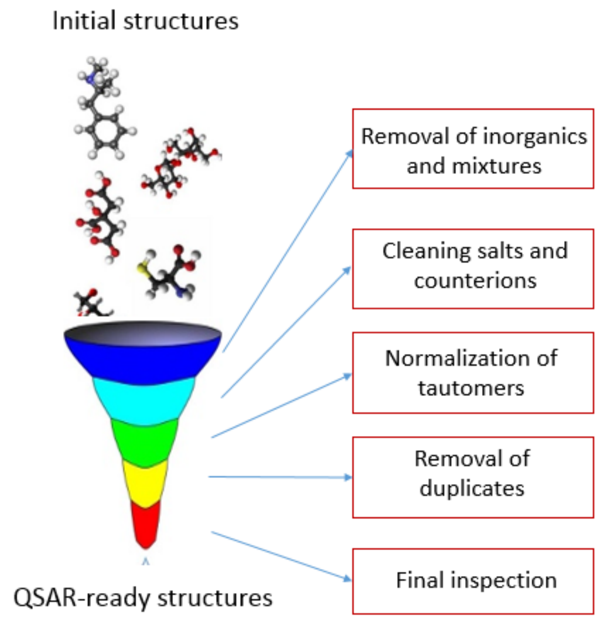
Check consistency of name with ACDLabs
Compare MolBlock and SMILES string
Check for errors and duplicate
Check CASRN with DSSTox databse
Check for the consistency between a information and assign STARs
Correct errors and update structures then assig new STARs
Organize and save new structure

The KNIME workflow for automated processing of PHYSPROP data

The KNIME workflow was used to insert various levels of Quality Flags (1 to 4 STARS) indicating consistency between chemical structure formats and identifiers after performing different checks.

QSAR-ready standardization procedure:

For the purposes of QSAR modeling, the 3 and 4 STAR datasets were processed through a KNIME workflow. This processing removed inorganics and mixtures, processed salts into neutral forms (except for melting point data), normalized tautomers, and removed duplicates. The resulting "QSAR-ready" file(s) were modeled using Genetic Algorithm-Partial Least Squares with 5-fold cross validation and utilizing 2D PaDEL[4] molecular descriptors. Multiple modeling runs (100) produced the best models using a minimum number of descriptors. The models for all 13 endpoints are available as both Windows and Linux executable binaries and as a C++ library that can be called by a separate application.



Removal of inorganics and mixtures

Cleaning salts and counterions

Normalization of tautomers

Removal of duplicates

Final inspection

QSAR-ready data used for modeling

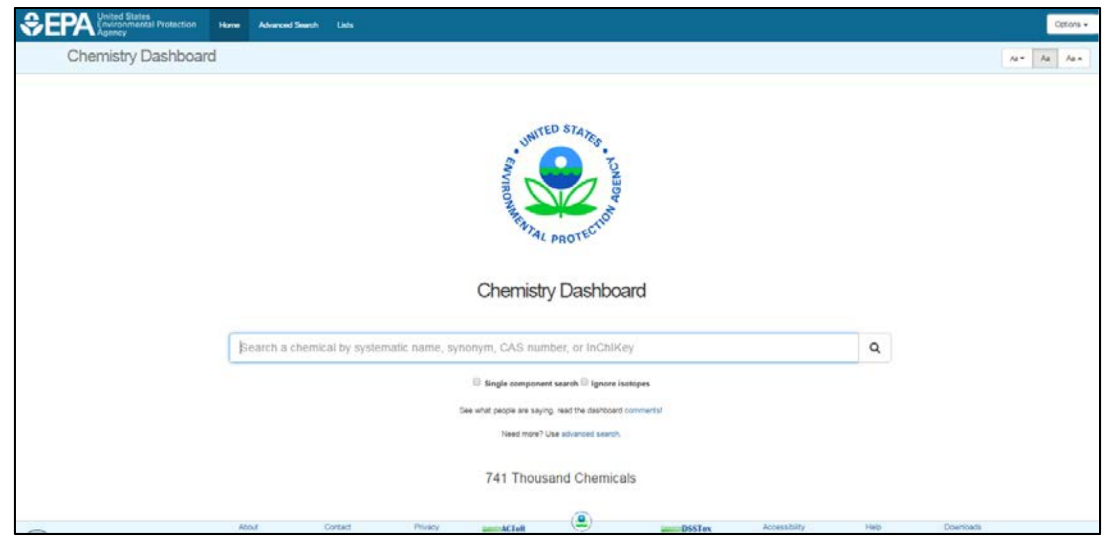
Abbreviation	Property	Initial file flagged	Updated 3-4 STAR	QSAR-ready
AOH	Atmospheric Hydroxylation Rate	818	818	745
BCF	Bioconcentration Factor	685	618	608
BioHL	Biodegradation Half-life	175	151	150
RB	Ready Biodegradability	1265	1196	1171
BP	Boiling Point	5890	5591	5436
HL	Henry's Law Constant	1829	1758	1711
KM	Fish Biotransformation Half-life	631	548	541
KOA	Octanol/Air Partition Coefficient	308	277	270
LogP	Octanol-water Partition Coefficient	15809	14544	14041
MP	Melting Point	10051	9120	8656
KOC	Soil Adsorption Coefficient	788	750	735
VP	Vapor Pressure	3037	2840	2716
WS	Water solubility	2348	2046	2010

Model validation and performances

Properties	Descriptors	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
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

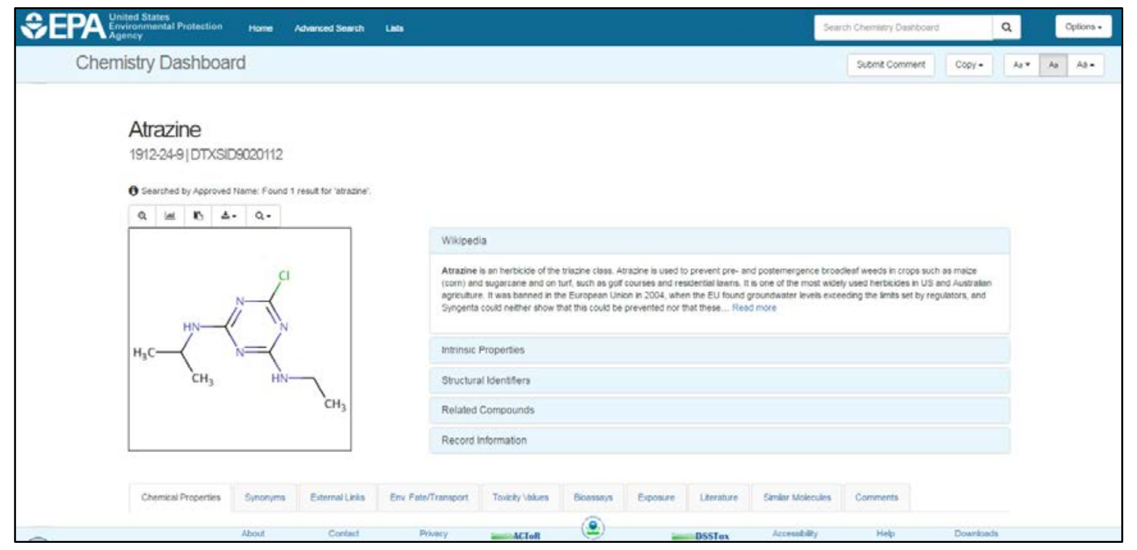
CV: cross-validation; Q2: coefficient of determination in CV; RMSE: root mean square error; N: number of chemicals; R2: coefficient of determination in training/test; BA: balanced accuracy; Sn: sensitivity; Sp: specificity.

CompTox Chemistry Dashboard

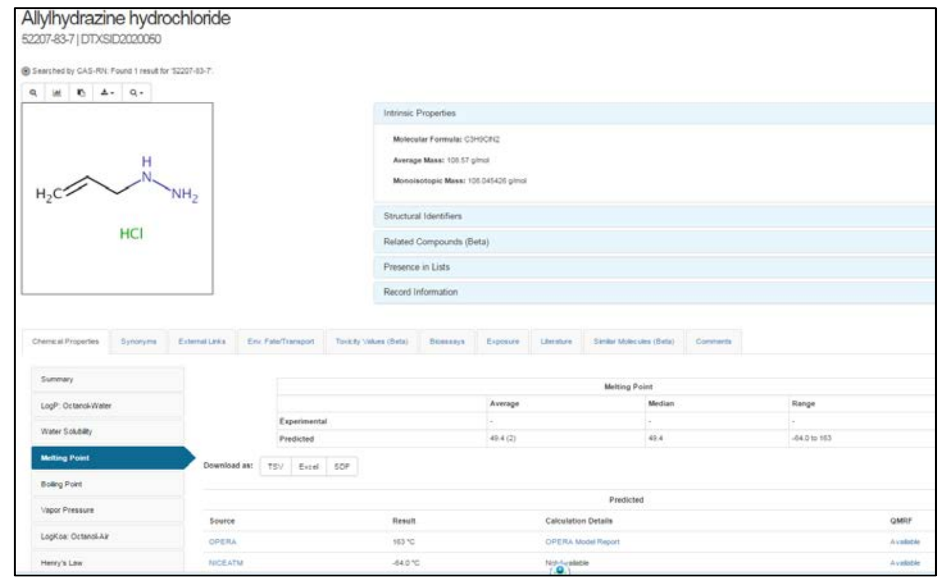


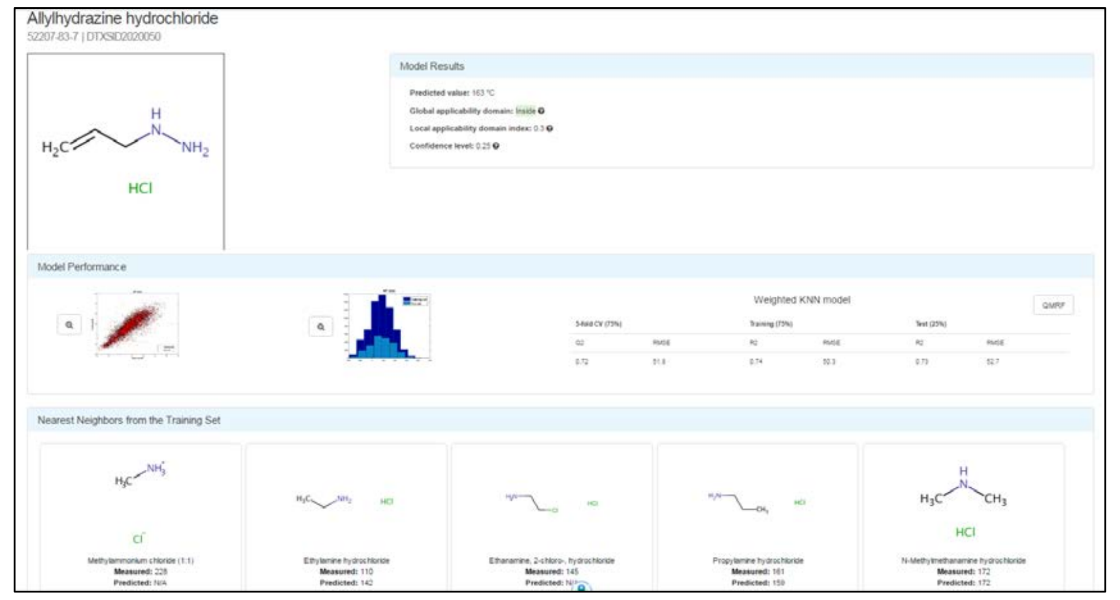
The CompTox Chemistry Dashboard is a public EPA-hosted web application providing access to ~750,000 chemicals from EPA's DSSTox database[5]. The entry to the dashboard is a simple text box allowing a type-ahead search for systematic, trade and trivial names, CAS Registry Numbers and InChI identifiers. An advanced search allows for searching based on molecular mass or molecular formula. http://comptox.epa.gov/dashboard.

For those records with associated chemical structure representations various inherent properties (molecular formula, mass, systematic name) and predicted physicochemical properties (logP, water solubility etc.) are provided. Where possible links are provided to related Wikipedia articles. An associated molfile is available for download to the desktop, and a summary report containing record data can be generated as a PDF file.



The dashboard shows chemical property predictions based on the **Ope**n structure-activity **R**elationships **A**pplication (OPERA) models developed from the curated datasets. QSAR modeling reports are available for all OPERA models and detailed model reports including global and local applicability domains, prediction reliability index, QMRF report, model statistics and nearest neighbors in the training set.





References

1. PHYSPROP Data: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm
2. KNIME: https://www.knime.org/
3. Mansouri *et al.* CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ Health Perspect;* DOI:10.1289/ehp.1510267
4. PaDEL descriptors, http://padel.nus.edu.sg/software/padeldescriptor/
5. EPA Distributed Structure-Searchable Toxicity (DSSTox) Database, http://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database