The EPA Online Database of Experimental and Predicted Data to Support Environmental Scientists

Antony Williams[1,*], Christopher Grulke[1], Kamel Mansouri[2], Jennifer Smith[1], Jordan Foster[1], Michelle Krzyzanowski and Jeff Edwards[1]

1. U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC
2. Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC

2627/P125
American Chemical Society Fall Meeting
Philadelphia, PA.
August 21-25, 2016

ORCID: 0000-0002-2668-4821
Antony Williams I williams.antony@epa.gov I 919-541-1033

Problem Definition and Goals

Problem: There is limited access to freely available high quality experimental and predicted chemical data online.

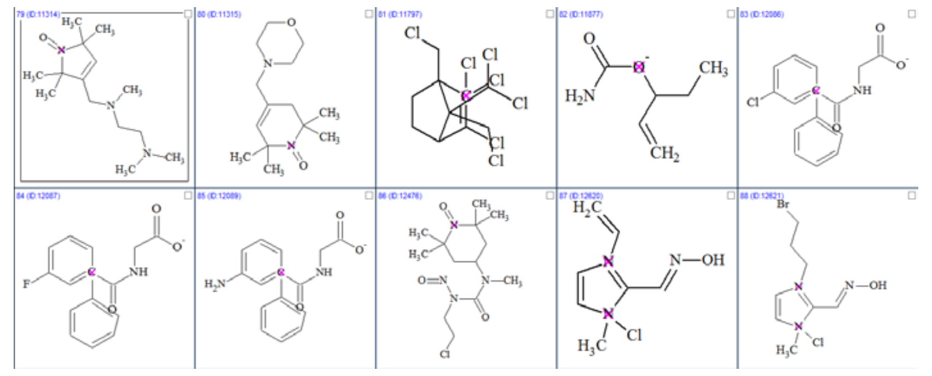
Goals: To deliver access to curated datasets of experimental and physicochemical data associated with chemicals of interest to environmental scientists. Make the data available as downloadable Open data. To develop predictive models from the curated data and use to predict properties for >700,000 chemicals and make available online. To provide details regarding the performance of the models.

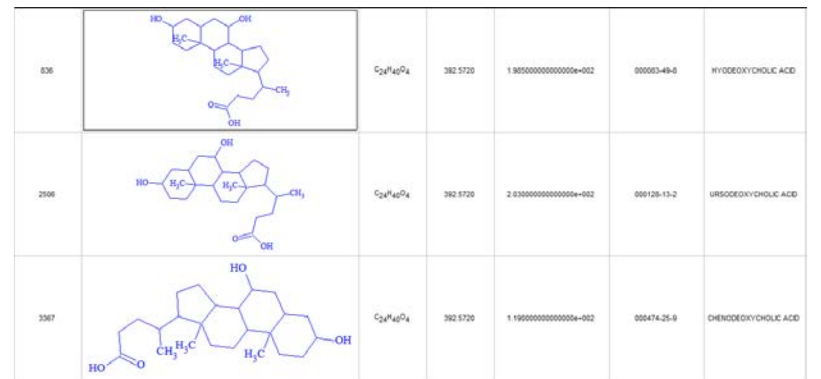
Abstract

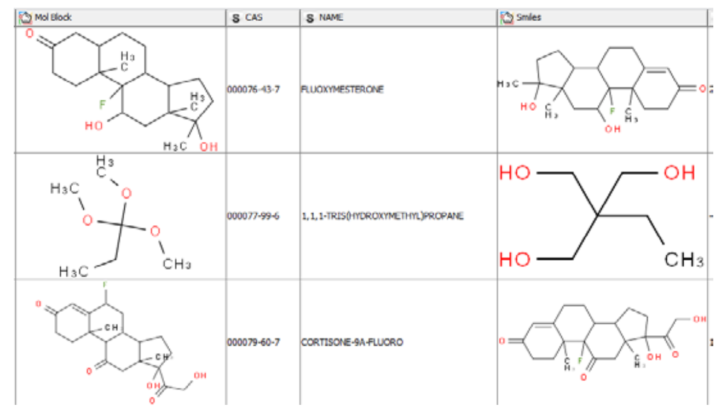
As part of our efforts to develop a public platform to provide access to experimental data and predictive models for physicochemical data we have delivered a web-based database, the EPA Chemistry Dashboard (https://comptox.epa.gov). In order to develop the prediction models we used publicly available data and a combination of manual and automated review processes to produce highly curated datasets. The automated curation processes for the validation of the data used a KNIME workflow and included approaches to validate different chemical structure representations (e.g. molfile and SMILES), identifiers (chemical names and registry numbers), and methods to standardize the data into QSAR-consumable formats for modeling. Machine-learning approaches were applied to create a series of models that have been used to generate predicted physicochemical and environmental parameters for over 700,000 chemicals. The data have been made available online via the EPA's iCSS Chemistry Dashboard and includes detailed model reports regarding whether or not a chemical is contained within a particular applicability domain, provides access to nearest neighbors within the training set as well as detailed QMRF (QSAR Modeling Report Format) reports for each of the models.

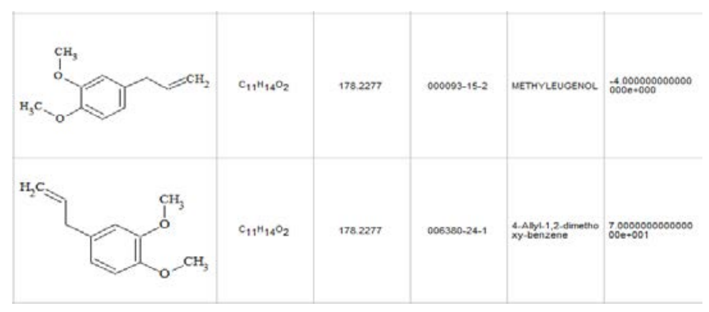
Source Data and Example Errors

- The PHYSPROP data were sourced online[1] as SDF files. A total of 13 endpoints were represented, including LogKow, water solubility, melting point, boiling point, and others. The largest dataset, LogKow, contained over 15,800 individual chemicals. Each data point included the Mol-Block, SMILES, CASRN, Name, LogKow value and, where available, a reference. This dataset was chosen as representative for examining the quality of data.
- A manual examination of the data revealed a number of issues: e.g., SMILES and Mol-Block did not agree; CASRN did not match the correct structure; SMILES could not be converted; a single chemical structure would be listed multiple times with different property values. Example errors across various PHYSPROP datasets are shown below.


Examples of hypervalency in LogKow dataset


Equivalent structures, different CASRN, names, and values in MP dataset

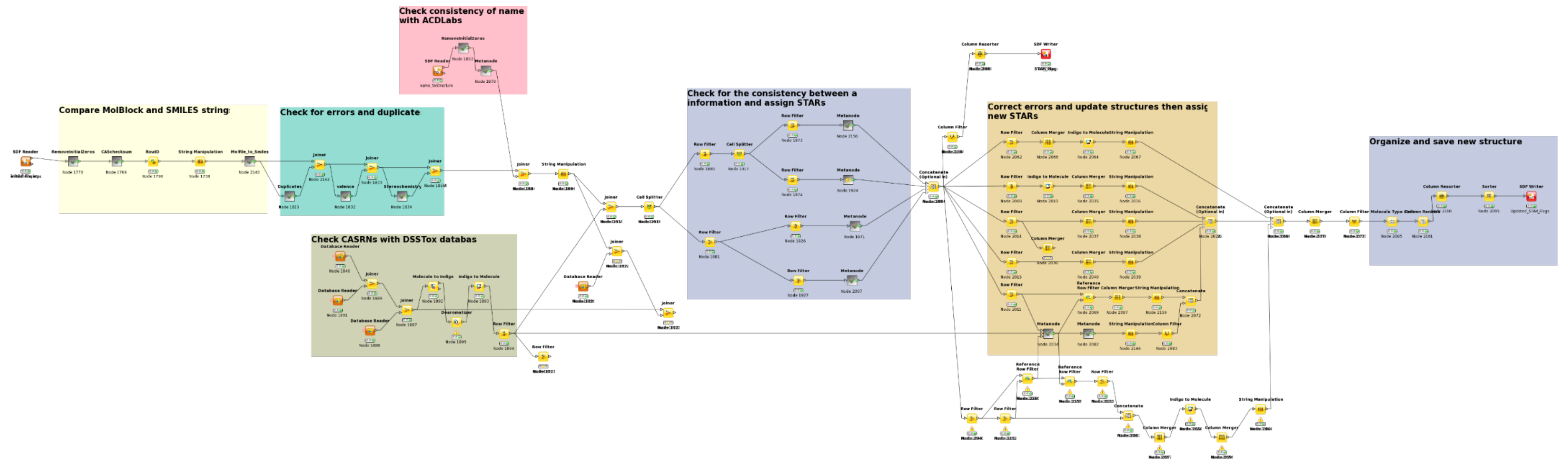

Different structures in Mol-Block and SMILES


Equivalent structures but with different CASRN, names and values in MP dataset: -4 and + 70ºC

Automated Analysis Using KNIME

The 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 representing, for example:

- Compare Mol-Block and SMILES (2268 different)
- Check for duplicates (657 structures, 531 names)
- Check CASRN Numbers (3646 invalid CASRN)
- Check names against dictionary (555 invalid)
- Assign Quality flags based on consistency among data fields


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

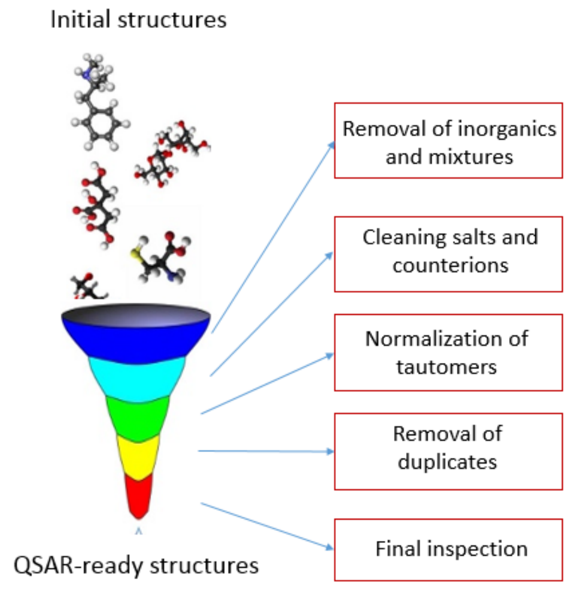
The KNIME workflow for automated processing of PHYSPROP data

The KNIME workflow was used to insert various levels of Quality Flags indicating consistency between chemical structure formats and identifiers. The consistency flag definitions and distribution are summarized below for the >15k chemicals.

4 STAR ENHANCED:	Name/CASRN/Mol/SMILES added Stereo:	550
4 STAR:	3 of 4 Name/CASRN/Mol/SMILES:	5967
3 STAR ENHANCED:	3 of 4 Name/CASRN/Mol/SMILES added Stereo:	177
3 STAR:	3 of 4 Name/CASRN/Mol/SMILES:	7910
2 STAR PLUS:	2 of 4 Name/CASRN/Mol/SMILES/Tautomer:	133
2 STAR:	2 of 4 Name/CASRN/Mol/SMILES:	1003
1 STAR:	No two fields consistent	379

Predictive models were developed using only the 3 and 4 star curated data.

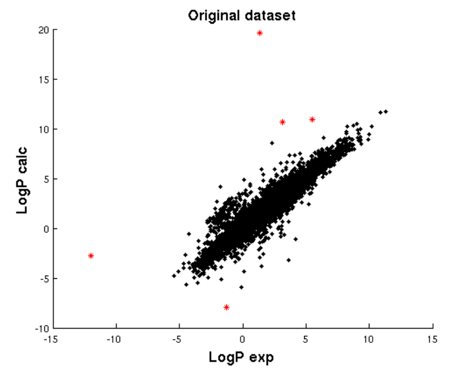
Preparing data for QSAR Modeling


Initial structures
Removal of inorganics and mixtures
Cleaning salts and counterions
Normalization of tautomers
Removal of duplicates
Final inspection
QSAR-ready structures

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.

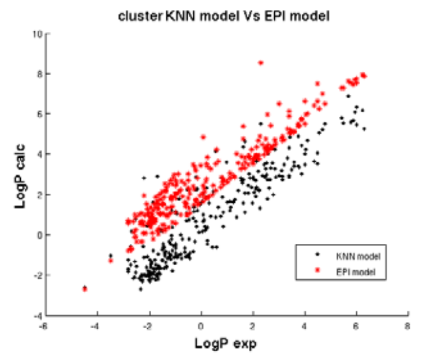
Model Performance

The curation of the available data, utilization of large datasets for training and validation, and application of innovative machine-learning approaches produced high performance, yet simple models with a minimum number of descriptors. The logP model was built on a dataset of more than 14k chemicals using only 10 descriptors. With a wider applicability domain this model outperformed EPI Suite's predictions (left side plot) especially for the over-estimated cluster extracted and plotted together with the new model's predictions (right side plot).



EPI Suite's predictions.

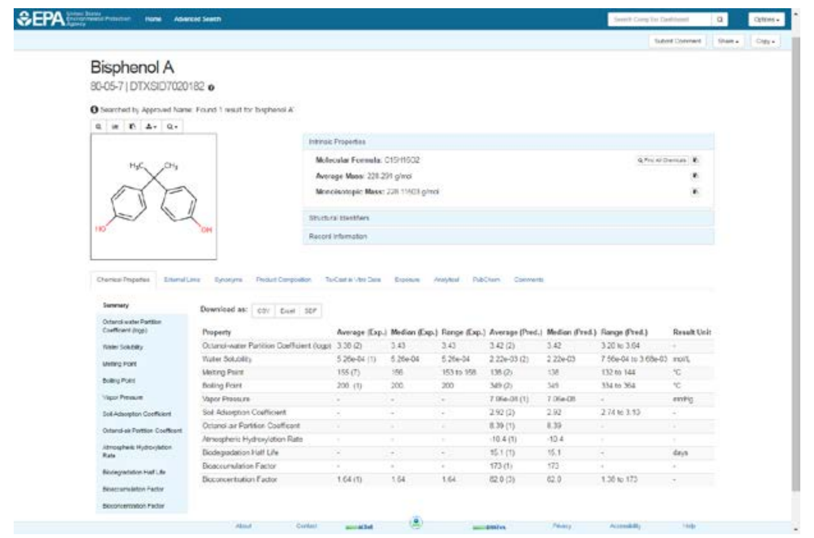
Statistics of the new model
5-fold cross-validation:
Q2: 0.87 RMSE: 0.67
Fitting 11247:
R2: 0.87 RMSE: 0.66
Test set prediction :
R2: 0.84 RMSE: 0.65

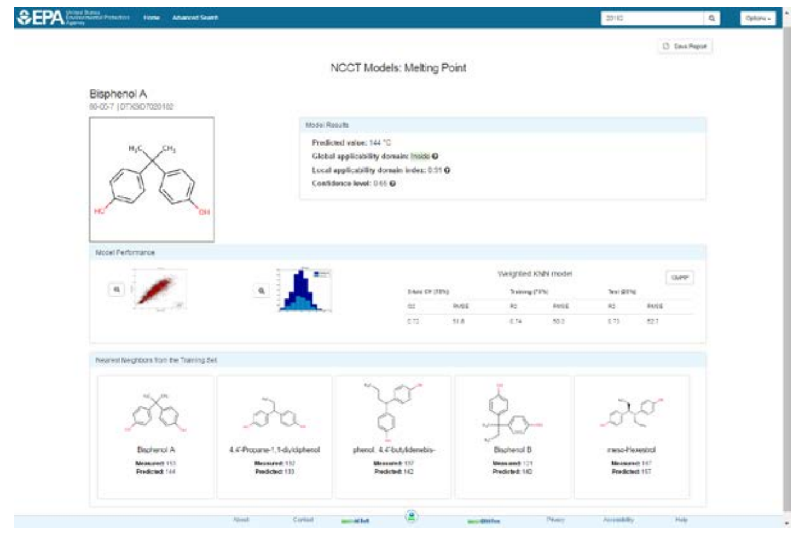


EPI Suite's (red) Vs NCCT model's predictions (black)

Accessing ~1 Million Predicted Properties Online

The iCSS Chemistry Dashboard (ICD) is a public EPA-hosted web application providing access to over 700,000 chemicals from EPA's DSSTox database[5]. It integrates experimental and predicted data and is a hub to other NCCT apps and web-based resources. The 3 and 4 STAR curated experimental data are accessible via the ICD application. All chemicals were also passed through the prediction models and detailed model reports and results are freely available at http://comptox.epa.gov/dashboard.


A summary of available chemical properties for Bisphenol A


The model report for the melting point prediction for Bisphenol A – including nearest neighbors.

Future Work

- Release NCCT models as interactive online prediction tools in the near future via the ICD.
- Integrate the suite of EPA T.E.S.T[6] physchem and toxicity prediction models to expand the collection of available models.
- Source additional data to expand the training sets underpinning the prediction algorithms.

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
6. EPA Toxicity Estimation Software Tool (T.E.S.T.) software http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

Innovative Research for a Sustainable Future

This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.