Influence of data curation on QSAR Modeling – examining issues of quality versus quantity of data

Antony Williams*, Kamel Mansouri, Ann M. Richard and Chris Grulke

U.S. Environmental Protection Agency, Office of Research and Development, National Center for Computational Toxicology (NCCT), Research Triangle Park, NC

Oak Ridge Institute for Science and Education (ORISE) Participant, Research Triangle Park, NC

2627/P125

Society of Toxicology Annual Meeting

New Orleans, LA

March 13-17, 2016

ORCID: 0000-0002-2668-4821

Antony Williams | williams.antony@epa.gov | 919-541-1033

Problem Definition and Goals

Problem: The performance of QSAR models is hampered by the *quality* of the underlying data.

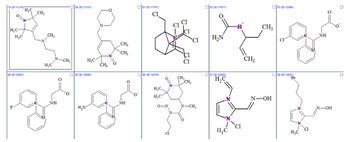
Goals: To examine the quality of the data underlying the EPI Suite prediction models using both manual and automated methods. To use the curated data to develop new prediction models and examine the influence of data quality and quantity, on the resulting QSAR predictions. To make the data and models publically available.

Abstract

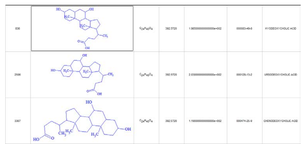
The accuracy of QSAR models is critically dependent on the quality of available data. As part of our efforts to develop public platforms to provide access to predictive models, we have attempted to discriminate the influence of the quality versus quantity of data available to develop and validate QSAR models. We have focused our efforts on the widely used EPI Suite[1] data (PHYSPROP database) that was initially developed over two decades ago. Specific examples of quality issues for PHYSPROP data include multiple records for the same chemical structure with different measured property values, inconsistency between the structure, chemical name and Chemical Abstracts Service Registry Number (CASRN) for single records, the inability to convert the SMILES strings into chemical structures, hypervalency in the chemical structures and the absence of stereochemistry for thousands of data records. Relative to the era of EPI Suite development, modern cheminformatics tools allow for more advanced capabilities in terms of chemical structure representation and storage, as well as enabling automated data validation and standardization approaches to examine data quality. This poster reviews both our manual and automated approaches to examining key datasets related to PHYSPROP data. This includes approaches to validate between chemical structure representations (e.g., Mol-Block and SMILES) and identifiers (chemical names and CASRN), as well as approaches to standardize the data into QSAR-ready formats for modeling. We have quantified and segregated the data into various quality categories to allow us to thoroughly investigate the resulting models that can be developed from these data slices and to examine to what extent efforts into the development of large high-quality datasets have the expected pay-off in terms of prediction performance.

Source Data and Example Errors

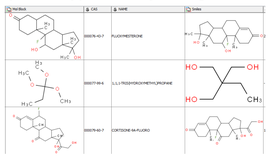
- The PHYSPROP data were sourced online[2] as SDF files. A total of 13 endpoints were represented, including LogKow, water solubility, melting point, boiling point, and others. The largest dataset, LogKow, contained over 15,800 individual chemicals. Each data point included the Mol-Block, SMILES, CASRN, 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.



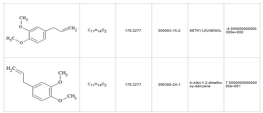
Examples of hypervalency in LogKow dataset



Equivalent structures, different CASRN, and values in MP dataset



Different structures in Mol-Block and SMILES

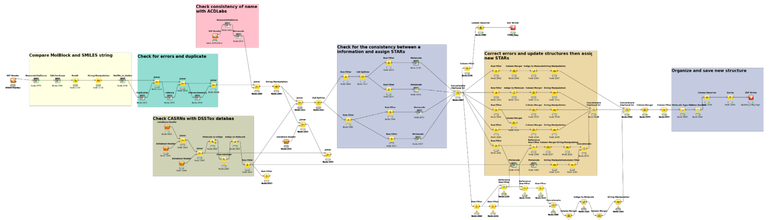


Equivalent structures but with different CASRN, names and values in MP dataset: -4 and + 70ºC

Automated Analysis Using KNIME

The manual investigation of the data allowed us to develop a KNIME[3] workflow for automated processing. This workflow was derived from earlier work by Mansouri *et. al.*[4] 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



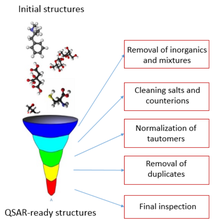
The KNIME workflow for automated processing of PHYSPROP data

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.

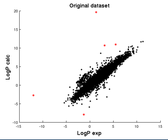
Preparing data for QSAR Modeling



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[5] 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 LogKow prediction model delivered by EPI Suite used a smaller dataset (of 2700 chemicals). The curation of the available data, utilization of a larger dataset (>14,000 chemicals) and application of novel machine-learning approaches produced a better and simpler model with only 10 descriptors. The figures below illustrate the difference between the original EPI Suite model and the newly derived predictive model. The red data points indicate the outliers from the original modeling approach, the majority not included in the original training set.



Statistics of the new model

5-fold cross-validation:

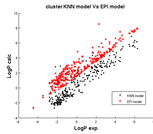
Q2: 0.87 RMSE: 0.67

Fitting:

R2: 0.87 RMSE: 0.66

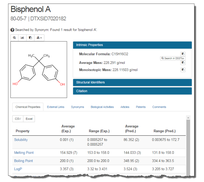
Test set Prediction:

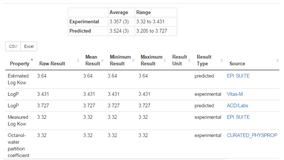
R2: 0.84 RMSE: 0.65



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[6]. 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 collection of NCCT prediction models and results will be freely available at http://comptox.epa.gov/dashboard in April 2016.





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[7] 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. EPI Suite: http://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface
2. PHYSPROP Data: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm
3. KNIME: https://www.knime.org/
4. Mansouri *et al.* CERAPP: Collaborative Estrogen Receptor Activity Prediction Project, *Environ Health Perspect;* DOI:10.1289/ehp.1510267
5. PaDEL descriptors, http://padel.nus.edu.sg/software/padeldescriptor/
6. EPA Distributed Structure-Searchable Toxicity (DSSTox) Database, http://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database
7. EPA Toxicity Estimation Software Tool (T.E.S.T.) software http://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test