Development of a Water Solubility Dataset to Establish Best Practices for Curating New Datasets for QSAR Modeling

Charles Lowe[1], Chris Grulke[1] and Antony Williams[1]

[1]U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, Research Triangle Park, NC

ORCID: 0000-0001-9151-6157

Charles Lowe I lowe.charles@epa.gov I 919-541-5618

ACS Fall 2020
VIRTUAL
August 17-20, 2020

Problem Definition and Goals

Problem: There are numerous peer-reviewed publications and public websites that contain experimental data that could be used to improve existing QSAR/QSPR models. Commonly these data are not available in an ideal form: often limited to PDF supplementary info files for publications (with names or CASRNs and no electronic structure formant). However, when aggregation of these data has been attempted curation has been necessary.

Goals: Provide a *de facto* dataset for water solubility data that can be used to build multiple models and eventually a consensus model. Identify specific sets of chemicals that can improve existing models. Curate these data to ensure chemical identifiers represent the same chemical structure, physicochemical property data has consistent units, etc. Make these data available as downloadable data for use in QSAR/QSPR models and reuse in other databases.

Simplified Workflow for Dataset Assembly

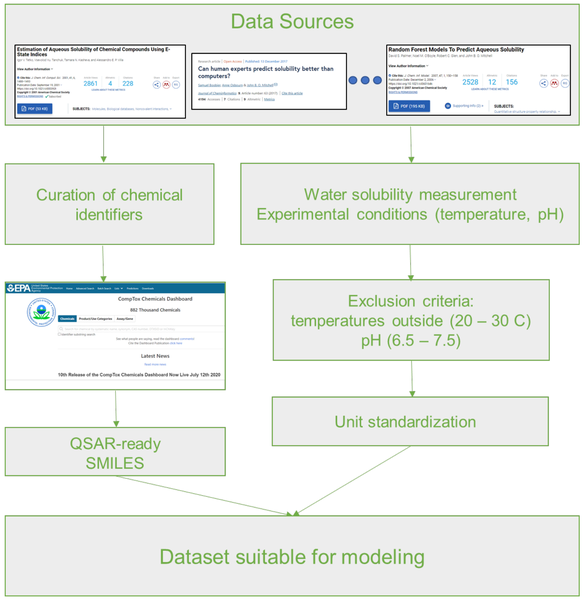

Figure 1: This diagram shows the simplified workflow used in the assembly of the water solubility dataset. Note that chemical structure is represented using QSAR-ready SMILES – a SMILES representation of the desalted, de-isotoped, stereo-neutral forms of chemical structures associated with particular chemical substances.

Article Identifier	Original No. of Chemicals	No. of QSAR-ready SMILES
https://doi.org/10.1021/ci700307p	287	286
https://doi.org/10.1002/minf.201000001	2810	2596
https://doi.org/10.1016/S0045-6535(02)00118-2	1719	1530
https://doi.org/10.1080/10807039.2015.1133242	1190	1155
https://doi.org/10.1186/s13321-017-0250-y	100	99
https://doi.org/10.6084/m9.figshare.1514952.v1	3315	1836
https://www.ebi.ac.uk/chembl/	326	323

Table 1: Current articles and databases assembled using the workflow shown in **Figure 1**. The number of QSAR-ready SMILES denotes the unique chemical structures available for modeling after curation and solubility measurement cleaning.

Simplified Modeling Workflow

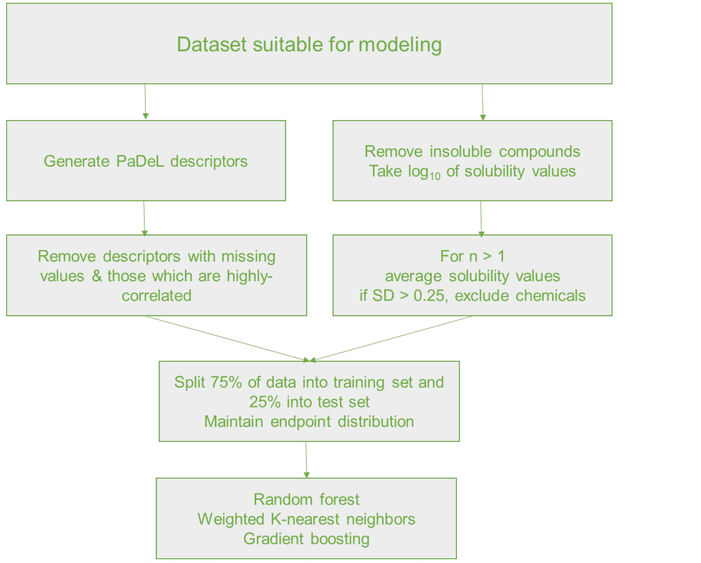

Figure 2 (above): This diagram shows the simplified workflow used in the modeling of the water solubility dataset.

Table 2 (below): Modeling approaches with performance values for cross-validated training and both internal and external test datasets[1].

Model Name	Training Dataset (5-fold CV) Dataset Size: 3153 RMSE	R^2	Test Dataset Dataset Size: 1049 RMSE	R^2	External Test Dataset Dataset Size: 4224 RMSE	R^2
Weighted K-Nearest Neighbors	0.95	0.82	0.98	0.81	0.76	0.89
Gradient Boosting	0.84	0.86	0.90	0.84	0.68	0.91
Random Forest	0.89	0.85	0.92	0.84	0.69	0.91

Dataset and Model Performance Metrics

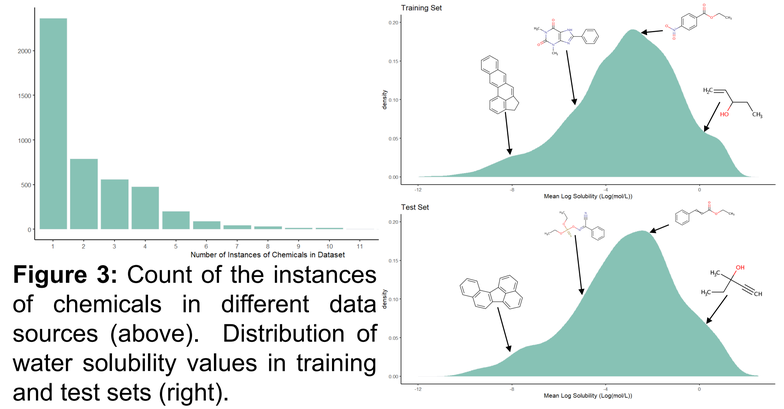

Figure 3: Count of the instances of chemicals in different data sources (above). Distribution of water solubility values in training and test sets (right).

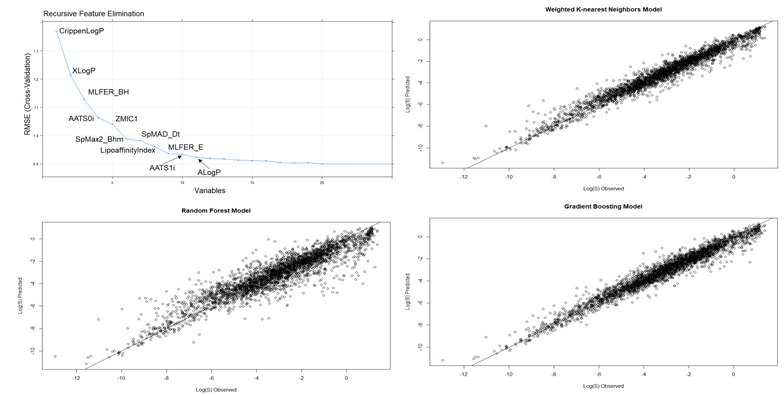

Figure 4: Selection of PaDeL descriptors via recursive feature elimination for K-NN model (top left). Correlation of predicted and experimental values in training set for models (remaining plots).

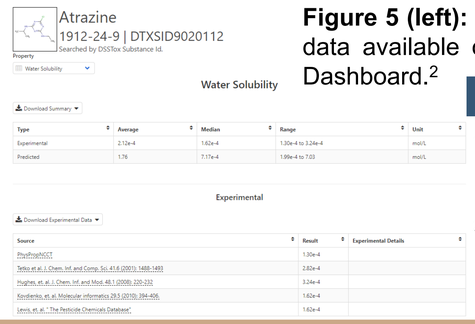

Figure 5 (left): Experimental water solubility data available on the CompTox Chemicals Dashboard.[2]

Supplemental Information

1. External test set is a curated version of the PhysProp dataset developed for EPI Suite, https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface.
2. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID9020112#properties and select "Water Solubility" from the dropdown menu.