United States
Environmental
Protection Agency

# Establishing Best Practices for Water Solubility Dataset Curation

Charles Lowe[1], Gabriel Sinclair[1,2], Christian Ramsland[1,2], Todd Martin[1], Christopher Grulke[1], and Antony J. Williams[1]

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency (U.S. EPA), Research Triangle Park, North Carolina, USA
[2]Oak Ridge Associated Universities (ORAU), Oak Ridge, Tennessee, USA.

ACS Fall 2021
August 22-26, 2021

ORCID: 0000-0001-9151-6157
Charles Lowe l lowe.charles@epa.gov l 919-541-5618

## Problem Definition and Goals

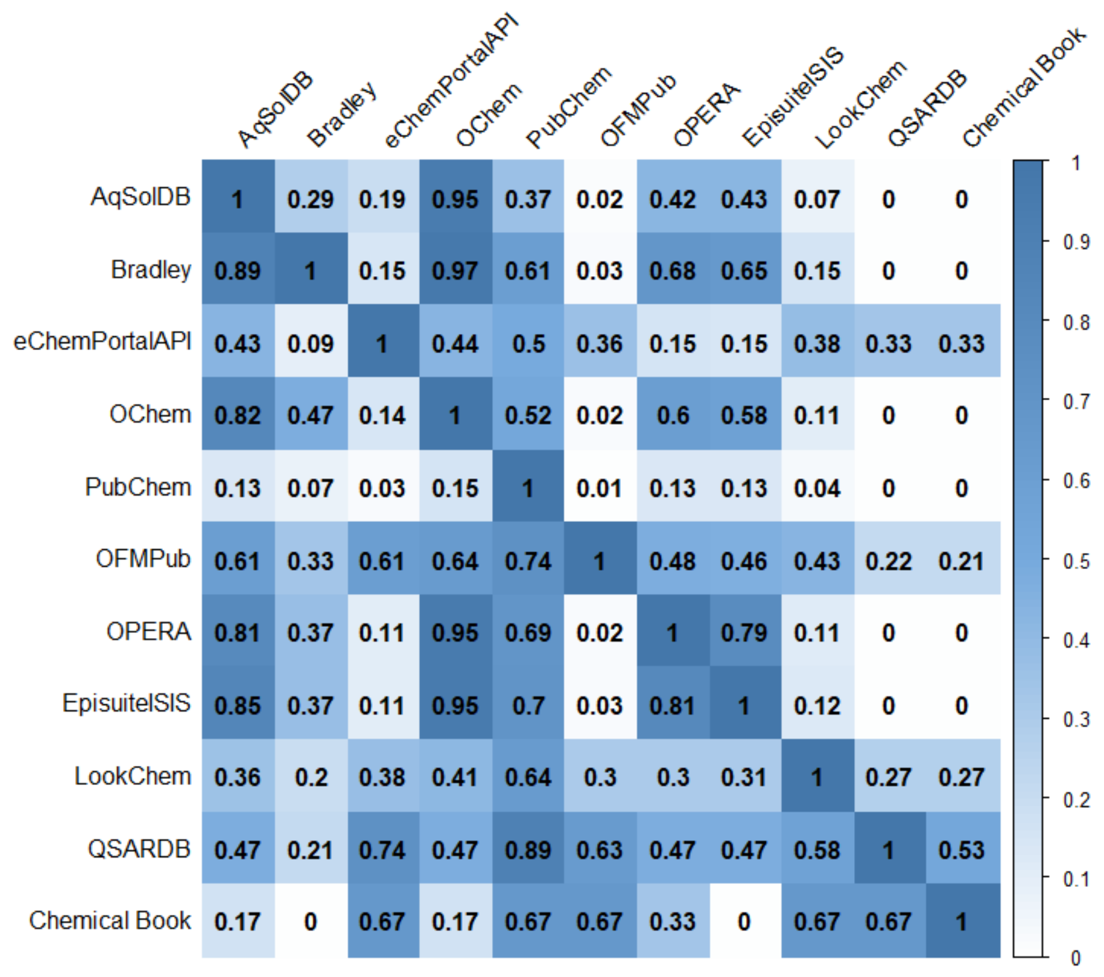
**Problem**: There are numerous peer-reviewed publications and public websites that contain experimental data that could be used to improve existing QSAR/QSPR models. Commonly these data are not available in an ideal form: often limited to PDF supplementary info files for publications (with names or CASRNs and no electronic structure format). 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. Determine erroneous records through curation and validation of chemical identifiers, as well as standardize solubility values and exclude outlying values using statistical approaches and cutoff values. Make these data available as downloadable data for use in QSAR/QSPR models and reuse in other databases.

## Dataset Assembly

| Database | URL | Original No. of Records | Curated No. of Records |
|---|---|---|---|
| AqSolDB | https://doi.org/10.1038/s41597-019-0151-1 | 9559 | 5657 |
| Bradley Dataset | http://dx.doi.org/10.1021/ci800406y | 3892 | 2909 |
| eChemPortalAPI | https://echa.europa.eu/ | 7037 | 2722 |
| Ochem | https://ochem.eu/ | 27928 | 18499 |
| PubChem | https://pubchem.ncbi.nlm.nih.gov/ | 45920 | 20701 |
| OFMPub | https://ofmpub.epa.gov/oppthpv/ | 386 | 213 |
| OPERA | ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/PHYSPROP_Analysis/ | 4839 | 4068 |
| EPISuiteISIS | http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm | 4755 | 3689 |
| LookChem | https://www.lookchem.com/ | 1820 | 907 |
| QSARDB | https://qsardb.org/ | 19 | 8 |
| Chemical Book | https://www.chemicalbook.com/ | 6 | 0 |

**Table 1:** The nine databases (and two journal articles), their corresponding URLs, and the original number of records found. The curated number of records was obtained using the workflow shown in **Figure 3**.



|  | AqSolDB | Bradley | eChemPortalAPI | OChem | PubChem | OFMPub | OPERA | EpisuiteISIS | LookChem | QSARDB | Chemical Book |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AqSolDB | 1 | 0.29 | 0.19 | 0.95 | 0.37 | 0.02 | 0.42 | 0.43 | 0.07 | 0 | 0 |
| Bradley | 0.89 | 1 | 0.15 | 0.97 | 0.61 | 0.03 | 0.68 | 0.65 | 0.15 | 0 | 0 |
| eChemPortalAPI | 0.43 | 0.09 | 1 | 0.44 | 0.5 | 0.36 | 0.15 | 0.15 | 0.38 | 0.33 | 0.33 |
| OChem | 0.82 | 0.47 | 0.14 | 1 | 0.52 | 0.02 | 0.6 | 0.58 | 0.11 | 0 | 0 |
| PubChem | 0.13 | 0.07 | 0.03 | 0.15 | 1 | 0.01 | 0.13 | 0.13 | 0.04 | 0 | 0 |
| OFMPub | 0.61 | 0.33 | 0.61 | 0.64 | 0.74 | 1 | 0.48 | 0.46 | 0.43 | 0.22 | 0.21 |
| OPERA | 0.81 | 0.37 | 0.11 | 0.95 | 0.69 | 0.02 | 1 | 0.79 | 0.11 | 0 | 0 |
| EpisuiteISIS | 0.85 | 0.37 | 0.11 | 0.95 | 0.7 | 0.03 | 0.81 | 1 | 0.12 | 0 | 0 |
| LookChem | 0.36 | 0.2 | 0.38 | 0.41 | 0.64 | 0.3 | 0.3 | 0.31 | 1 | 0.27 | 0.27 |
| QSARDB | 0.47 | 0.21 | 0.74 | 0.47 | 0.89 | 0.63 | 0.47 | 0.47 | 0.58 | 1 | 0.53 |
| Chemical Book | 0.17 | 0 | 0.67 | 0.17 | 0.67 | 0.67 | 0.33 | 0 | 0.67 | 0.67 | 1 |

**Figure 1:** Redundancy matrix showing the intersection of chemicals between datasets as a fractional value. While some of the databases have significant overlap (i.e., OPERA with OCHEM, EPI Suite with AqSolDB), no database perfectly overlaps with another. The significant overlap between databases allows for checks of parity, where ambiguously-represented chemicals may be corrected or removed.

## Dataset Curation

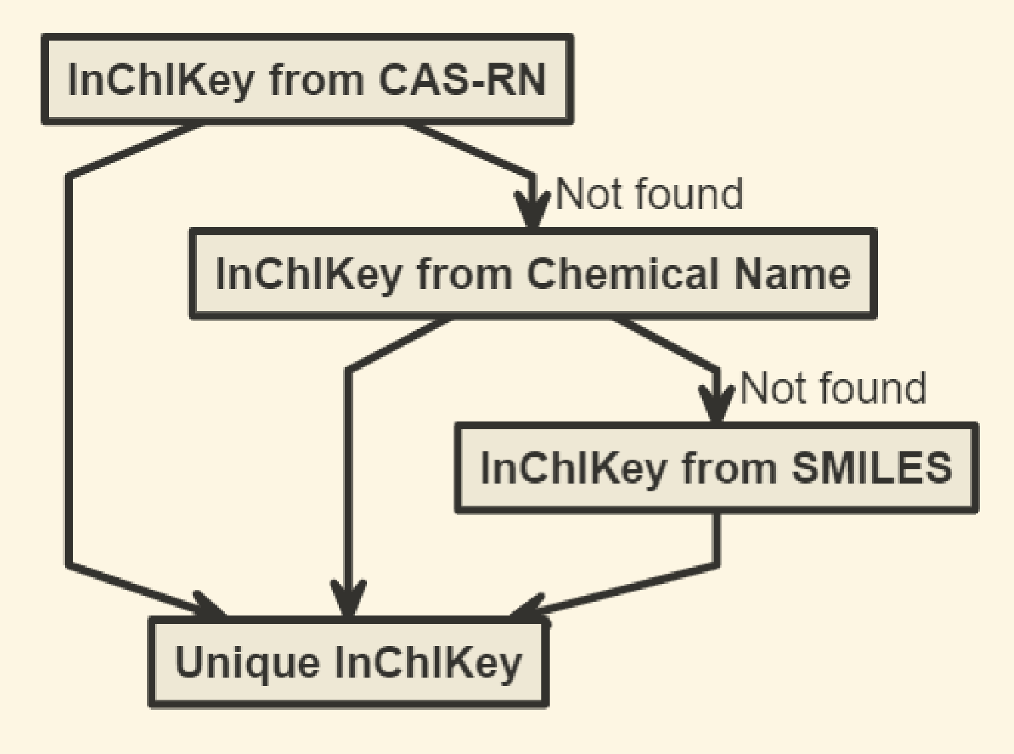
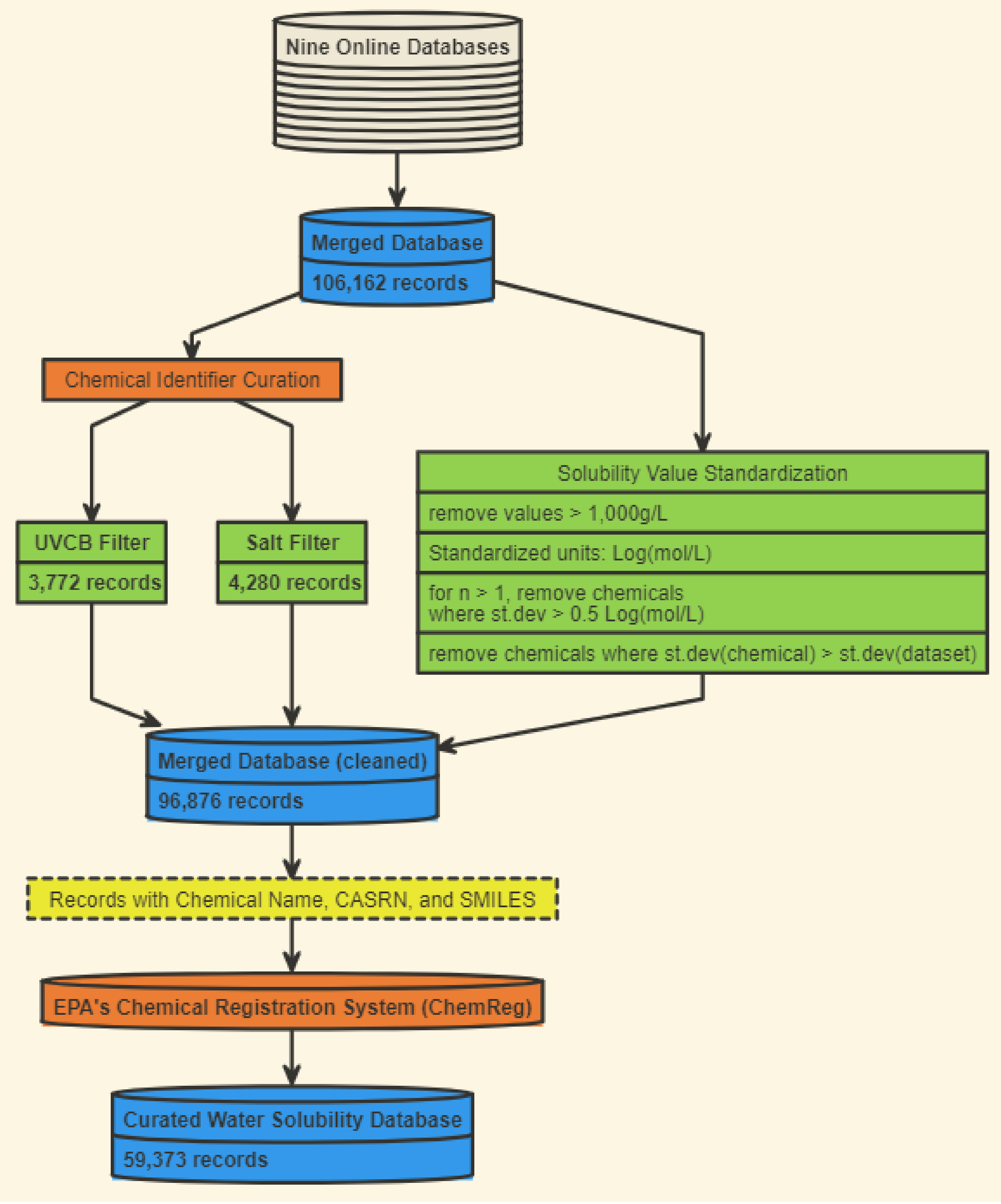


**Figure 2:** This diagram shows the workflow for selecting a unique chemical identifier for each dataset entry. (above) InChIKeys are determined via a search of OPSIN or ACD/labs software.



Nine Online Databases
Merged Database
106,162 records
Chemical Identifier Curation
UVCB Filter
3,772 records
Salt Filter
4,280 records
Solubility Value Standardization
remove values > 1,000g/L
Standardized units: Log(mol/L)
for n > 1, remove chemicals where st.dev > 0.5 Log(mol/L)
remove chemicals where st.dev(chemical) > st.dev(dataset)
Merged Database (cleaned)
96,876 records
Records with Chemical Name, CASRN, and SMILES
EPA's Chemical Registration System (ChemReg)
Curated Water Solubility Database
59,373 records

**Figure 3:** This diagram shows the workflow for curation and standardization of the dataset. Chemical identifier curation is achieved as shown in **Figure 2.**

## Main Results

- In total, 49,804 unique chemicals mapped to 47,121 QSAR-ready structures*.
- 22,675 unique chemical structures mapped to 22,365 QSAR-ready structures* in EPA's chemical registration system.
- Examples of curation issues: multiple CAS-RNs or names per record (not UVCBs), truncated chemical names, UVCBs given a single chemical structure, inverted signs (negative instead of positive) that are correctable based on database parity.

*desalted, de-isotoped, stereo-neutral forms of chemical structures

## Preliminary Modeling of Curated Dataset

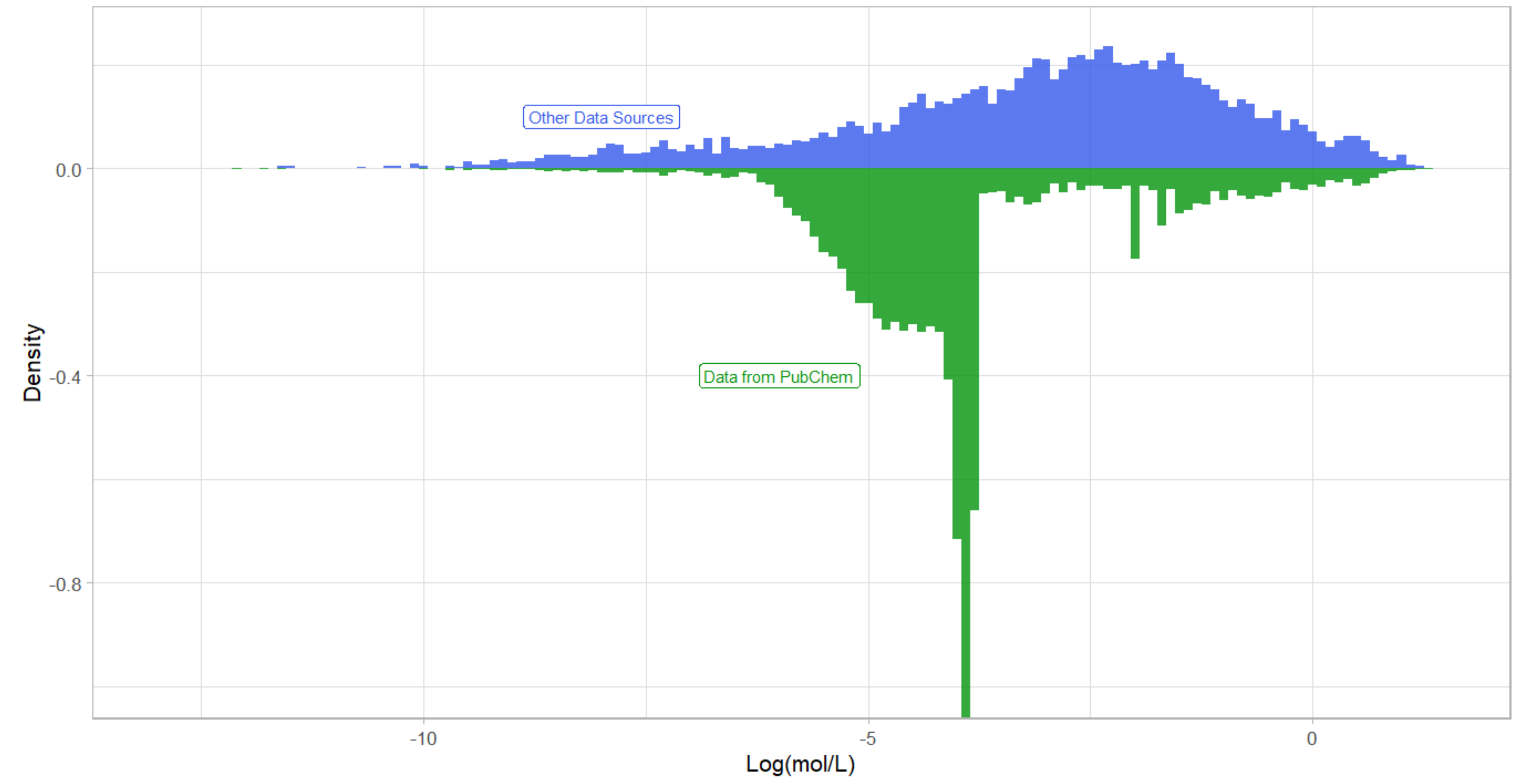
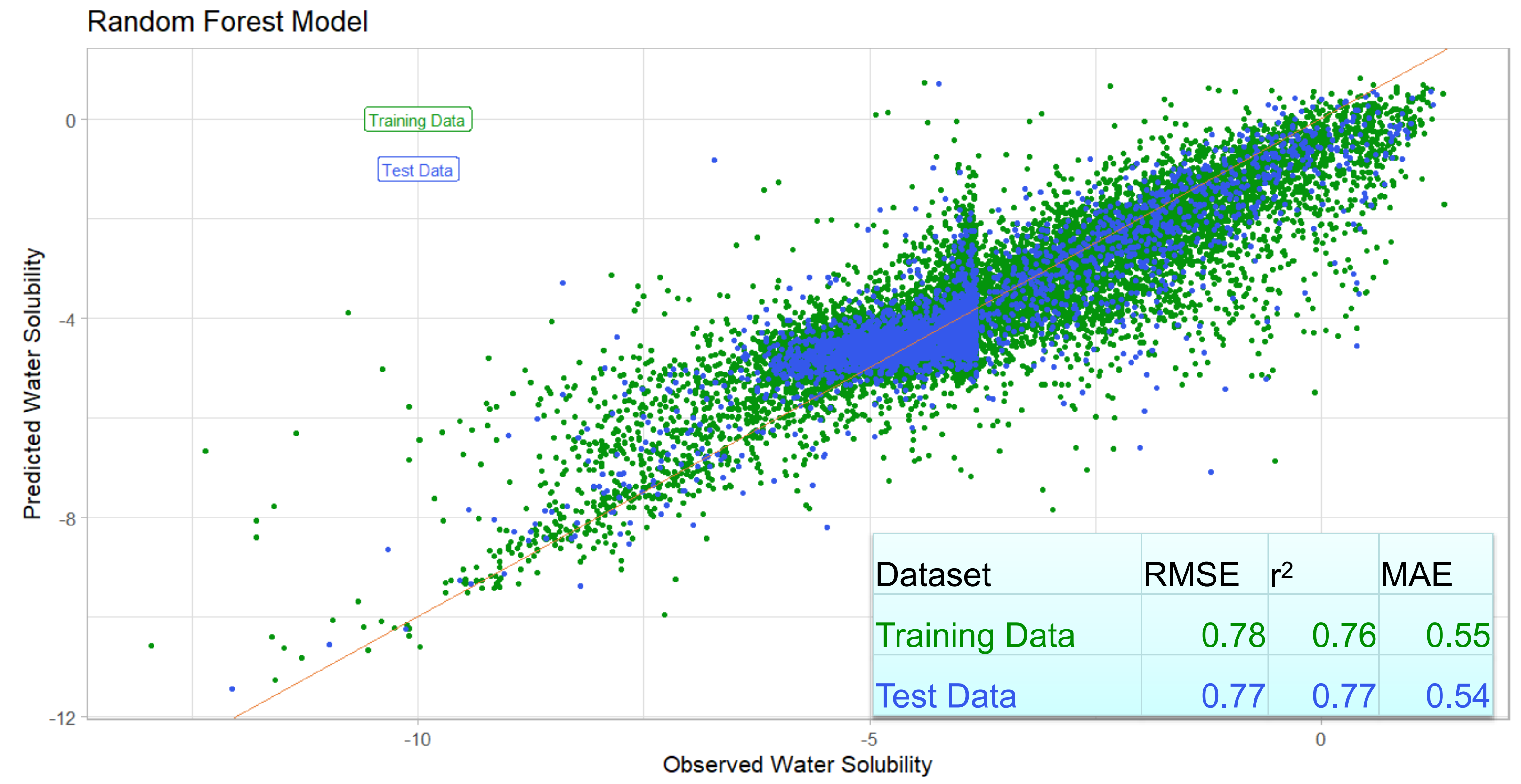


**Figure 4:** A density plot showing the different distributions of solubility values for the data from PubChem versus the data from other sources. The presence of numerous drugs in PubChem has imbalanced the overall dataset toward the fit-for-purpose region, [3 Log(mol/L) to 5 Log(mol/L)].



| Dataset | RMSE | $r^2$ | MAE |
|---|---|---|---|
| Training Data | 0.78 | 0.76 | 0.55 |
| Test Data | 0.77 | 0.77 | 0.54 |

**Figure 5:** This plot shows the correlation of experimental values to prediction results obtained from a random forest model constructed using 85% of the 22,365 QSAR-ready structures/solubility value pairs and PaDEL descriptors.

*Innovative Research for a Sustainable Future*

Disclaimer: The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA.