Establishing Best Practices for Water Solubility Dataset Curation

#198

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. ORCID: 0000-0001-9151-6157

OBJECTIVES

- The solubility of chemical compounds in water is important in most scientific disciplines, especially in the fields of toxicology and pharmacology.
- We will provide a *de facto* dataset for water solubility data that can be used to build multiple models and eventually a consensus model.
- Current water solubility models available in the CompTox Chemicals Dashboard (OPERA and TEST) are composed of approximately 4-5k unique chemicals.

APPROACH

- Gather water solubility data (between 20° – 30° C) from 9 large online databases and merge into one database.
- Determine erroneous records through curation and validation of chemical identifiers.
- Standardize solubility values and exclude outlying values using statistical approaches and cutoff values.
- Produce QSAR-ready structures (desalted, de-isotoped, stereo-neutral forms of chemical structures) and identifiers for future modeling work.

MAIN RESULTS

- 84,206 records are identifiable by name, 19,021 records were identifiable by CAS-RN, and 96,872 were identifiable by structure (SMILES).
- Currently 49,804 unique chemicals mapped to 47,121 QSAR-ready structures.
- Examples of curation issues: multiple CAS-RNs or names per record (not UVCBs), truncated chemical names, UVCB names given a single chemical structure, inverted signs

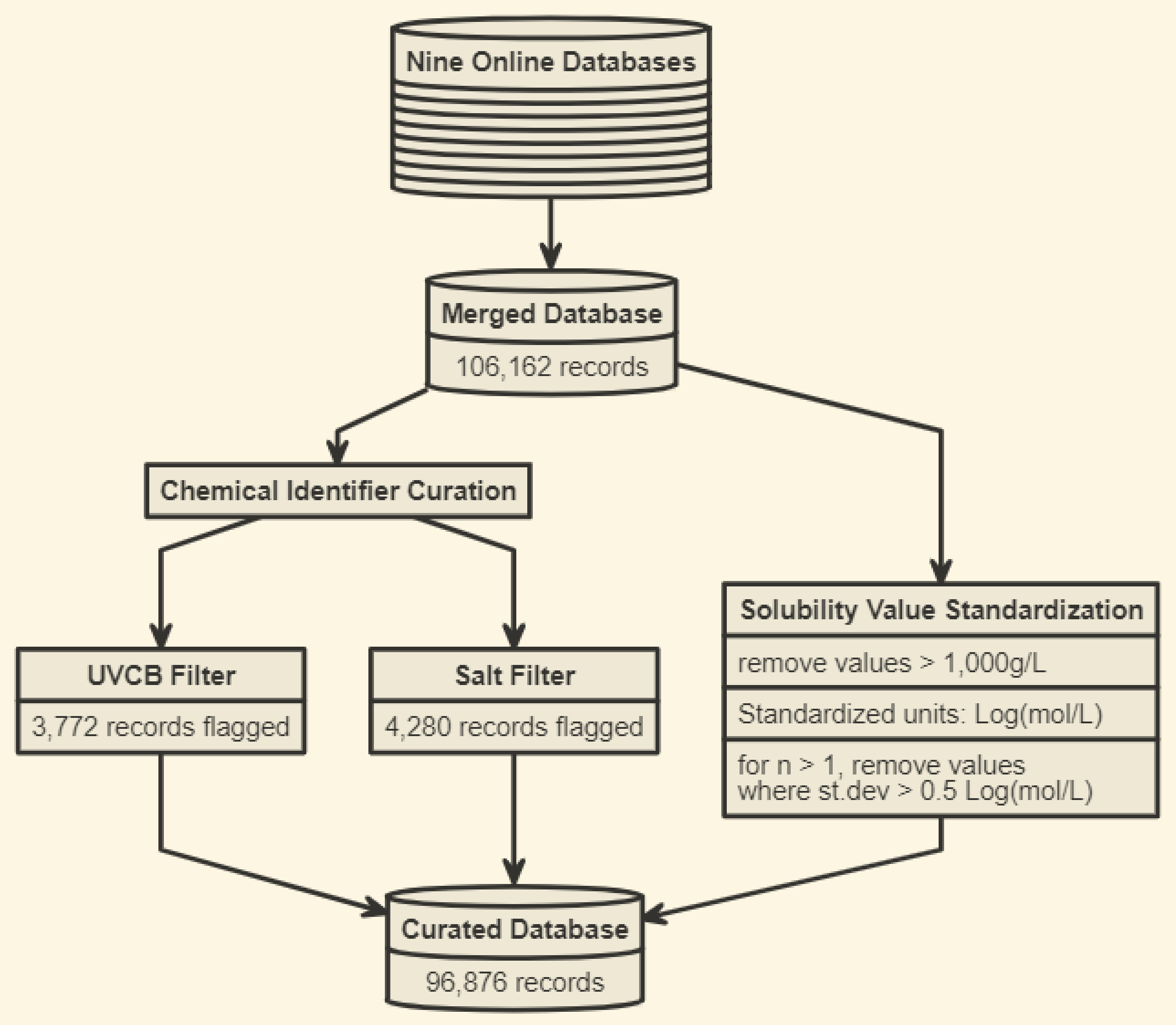
IMPACT

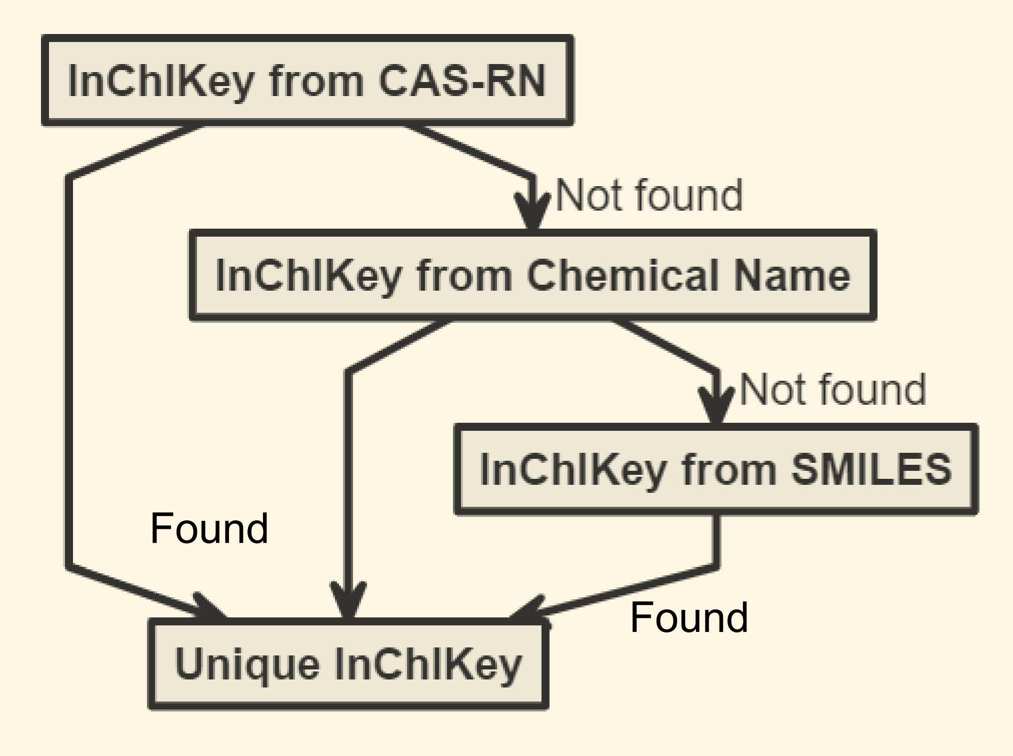
- The main result of this work is the creation of the largest assembled publicly-available water solubility dataset.
- The registration of this dataset in EPA's Distributed Structure Toxicity Database (DSSTox) is in progress.
- This dataset should support multiple EPA research projects with improved water solubility predictions in the future.

- **For more information, contact:** Charles Lowe,

Establishing Best Practices for Water Solubility Dataset Curation

APPROACH



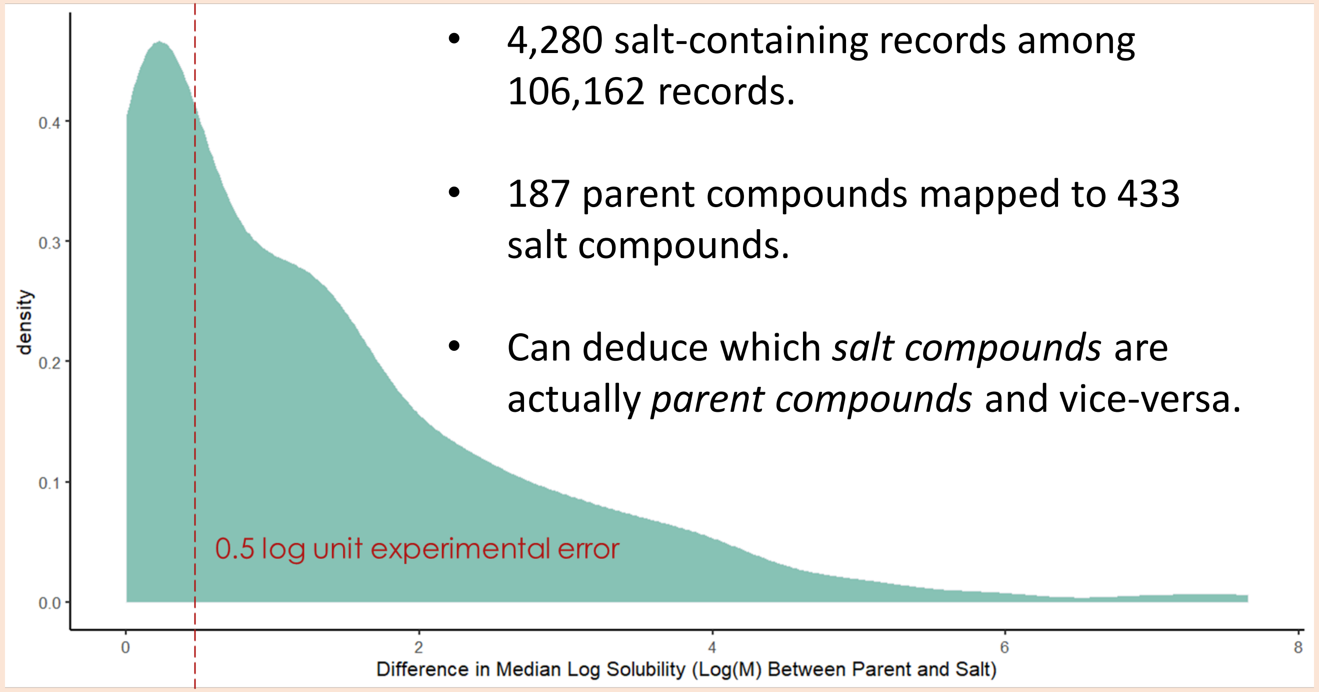


Workflow for selecting a unique chemical identifier for each dataset entry. (above) InChIKeys are determined via a search of OPSIN or ACD/labs software.

Workflow for curation and standardization of dataset. (left)

Establishing Best Practices for Water Solubility Dataset Curation

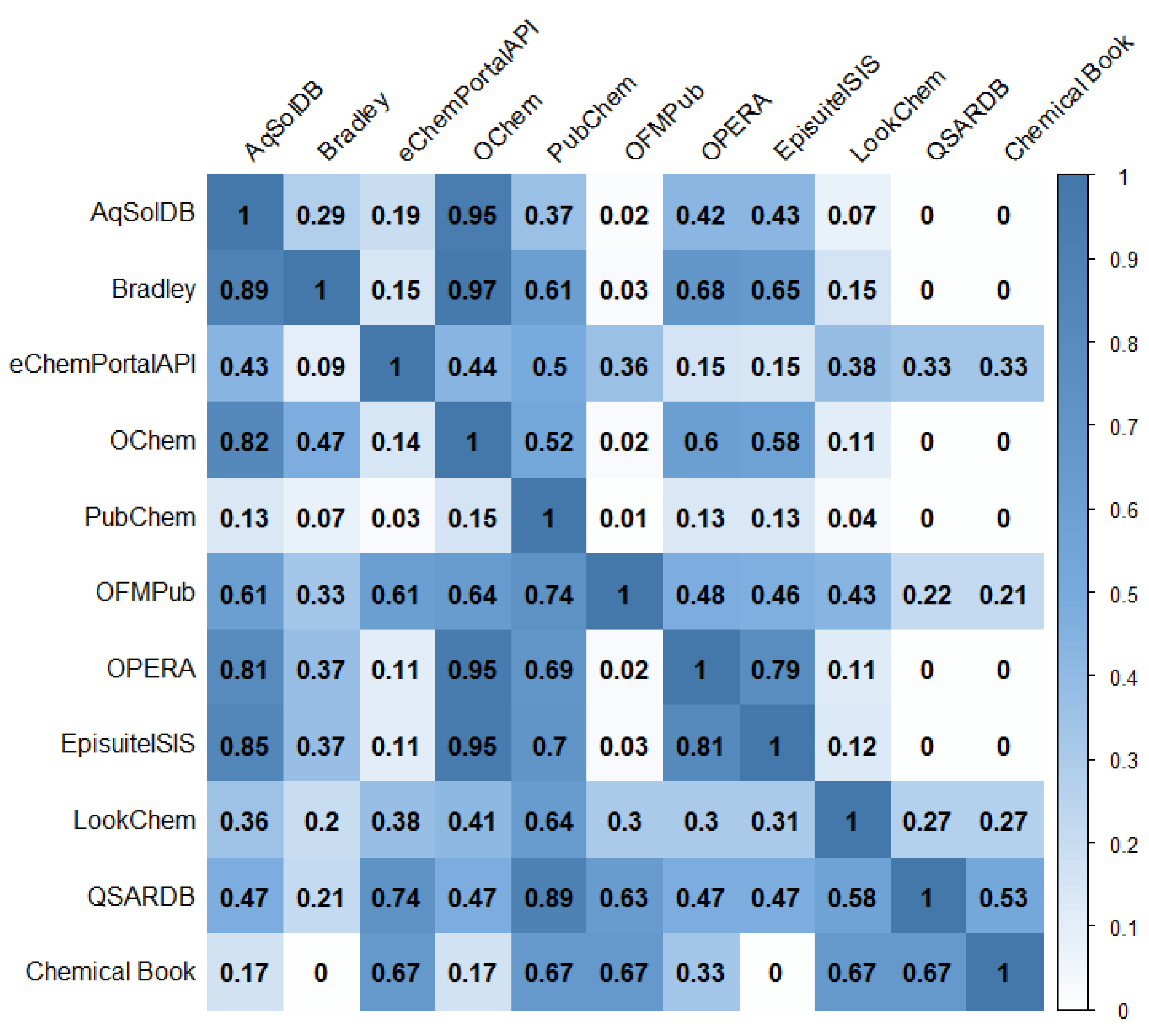
MAIN RESULTS



- 4,280 salt-containing records among 106,162 records.
- 187 parent compounds mapped to 433 salt compounds.
- Can deduce which *salt compounds* are actually *parent compounds* and vice-versa.

A density plot showing the difference in solubility between parent compounds and salt compounds

- Redundancy matrix shows that, 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.



	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

Redundancy matrix showing the intersection of chemicals between datasets as a fractional value.

Establishing Best Practices for Water Solubility Dataset Curation

MAIN RESULTS

Database	URL
AqSolDB	https://doi.org/10.1038/s41597-019-0151-1
Bradley Dataset	http://dx.doi.org/10.1021/ci800406y
eChemPortalAPI	https://echa.europa.eu/registration-dossier/
Ochem	https://ochem.eu/
PubChem	https://pubchem.ncbi.nlm.nih.gov/
OFMPub	https://ofmpub.epa.gov/oppthpv/
OPERA	ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/PHYSPROP_Analysis/
EPISuiteISIS	http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm
LookChem	https://www.lookchem.com/
QSARDB	https://qsardb.org/repository/explorer/
Chemical Book	https://www.chemicalbook.com/

A list of the nine databases (and two journal articles) and the corresponding URLs.