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

Charles Lowe, and Antony Williams

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

ChemCuration 2019
December 3rd, 2019

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

Problem Definition and Goals

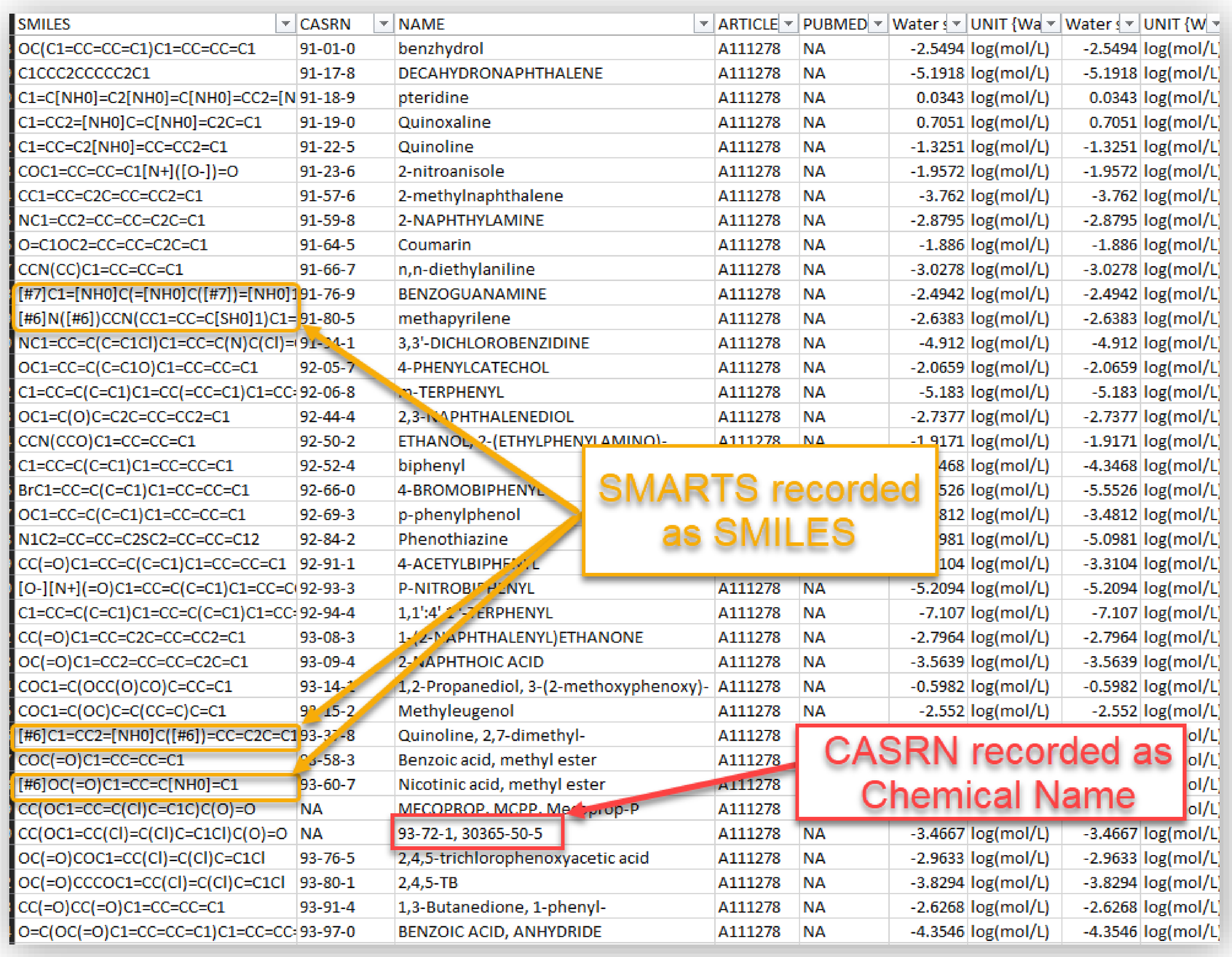
Problem: There are numerous peer-reviewed publications and public websites that contain experimental data that could be used to improve existing QSAR/QSPR models. Commonly these data are not available in an ideal form: often limited to PDF supplementary info files for publications (with names or CASRNs and no electronic structure 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. The project was started using aqueous solubility data available from the OCHEM database (https://ochem.eu/).

Abstract

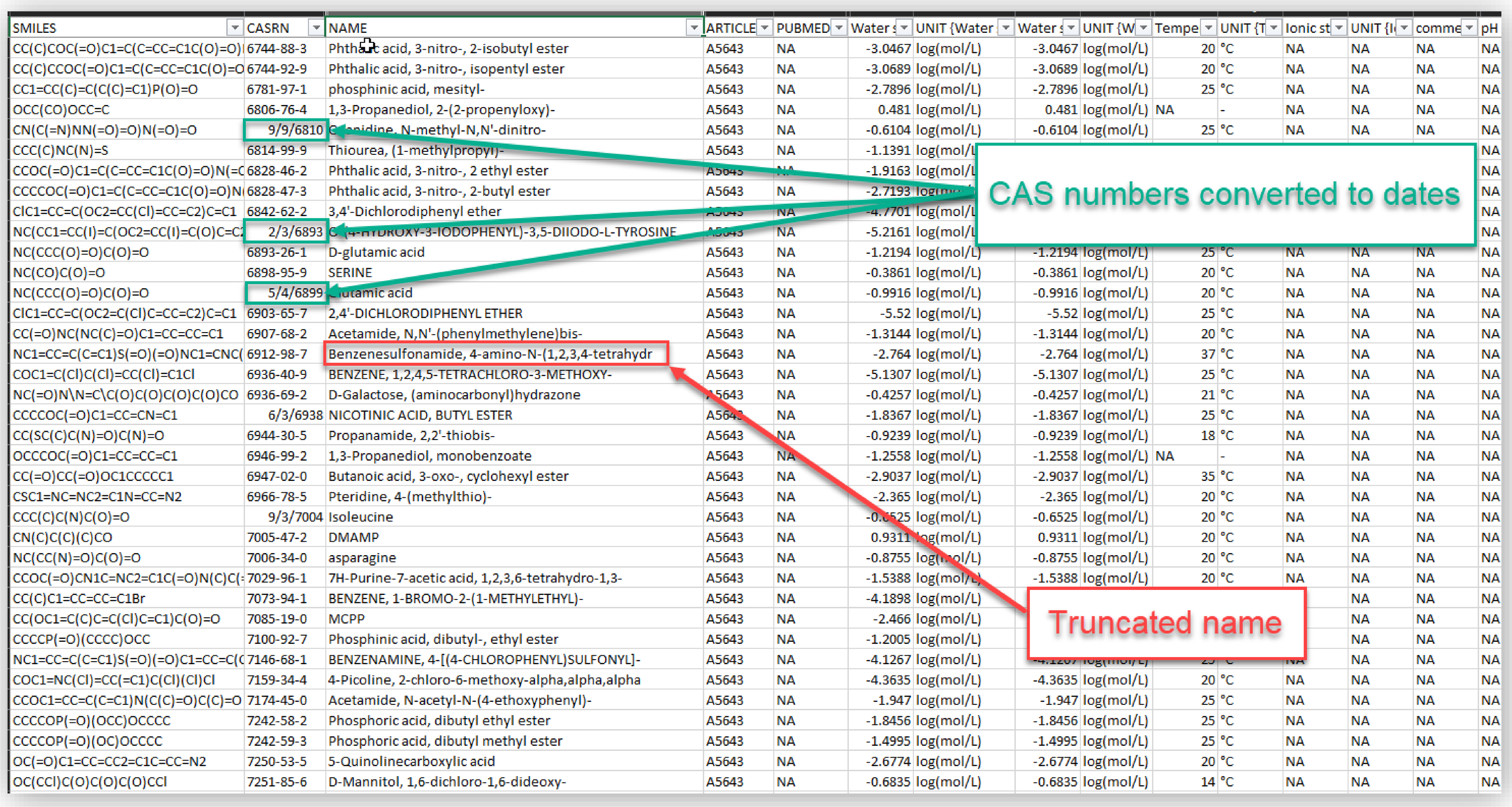
The U.S. Environmental Protection Agency's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) hosts a plethora of environmentally-relevant chemical information, including physical property data suitable for QSAR/QSPR modeling. The development of these physical property datasets has generally involved the curation of publicly-available experimental data. The ease of accessing this data, along with the overall quality of the dataset (i.e. machine-readable formatting, inclusion of experimental conditions, etc) is highly variable. This purpose of this work is to identify the challenges associated with acquiring physical property datasets, with a focus on obtaining water solubility values for organic compounds. Common issues discovered in this data will be presented, along with solutions that can be easily implemented in a high-throughput manner. The end result will be a standard workflow a researcher can follow when curating physical property datasets. *This abstract does not necessarily represent the views or policies of the U.S. Environmental Protection Agency.*

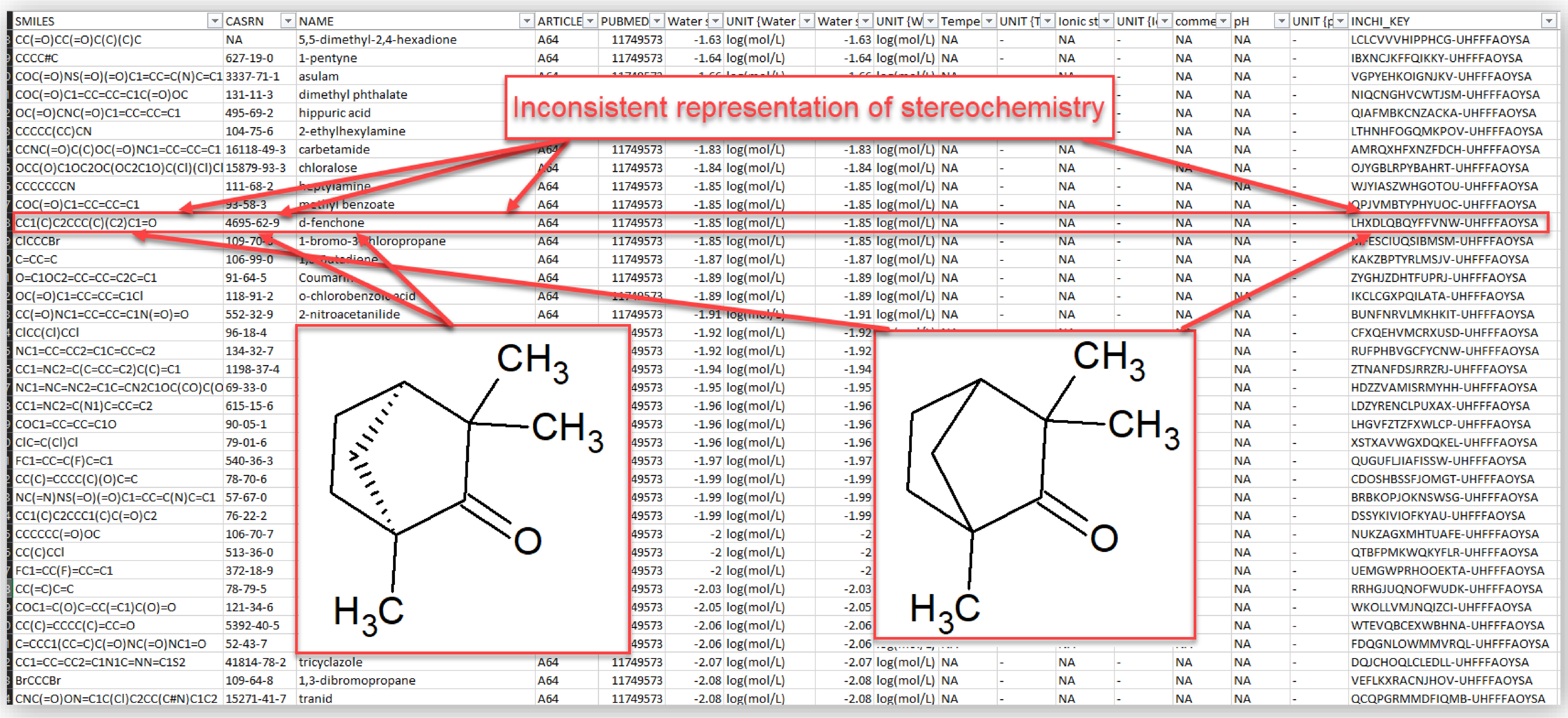
Issues Discovered During Aggregation and Curation

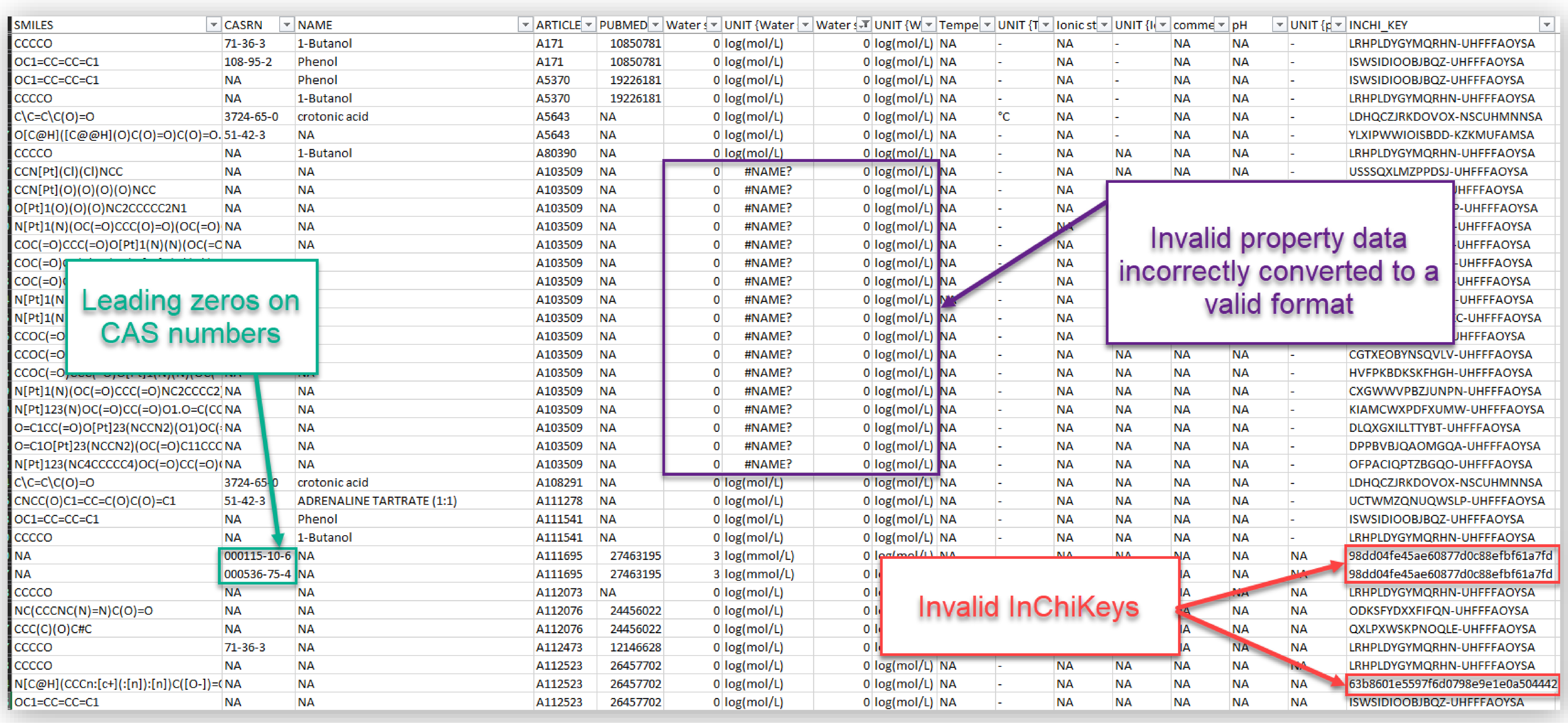


Left: Issues like SMILES being represented instead by SMARTS and chemical names containing other identifiers like CASRN are numerous.

Right: A well-known issue where Excel converts to CASRN to dates can be solved by using the function *=TEXT(CASRN value,"yyyy-mm-d")* in an adjacent cell. It's also common to see chemical names truncated with significant information loss.



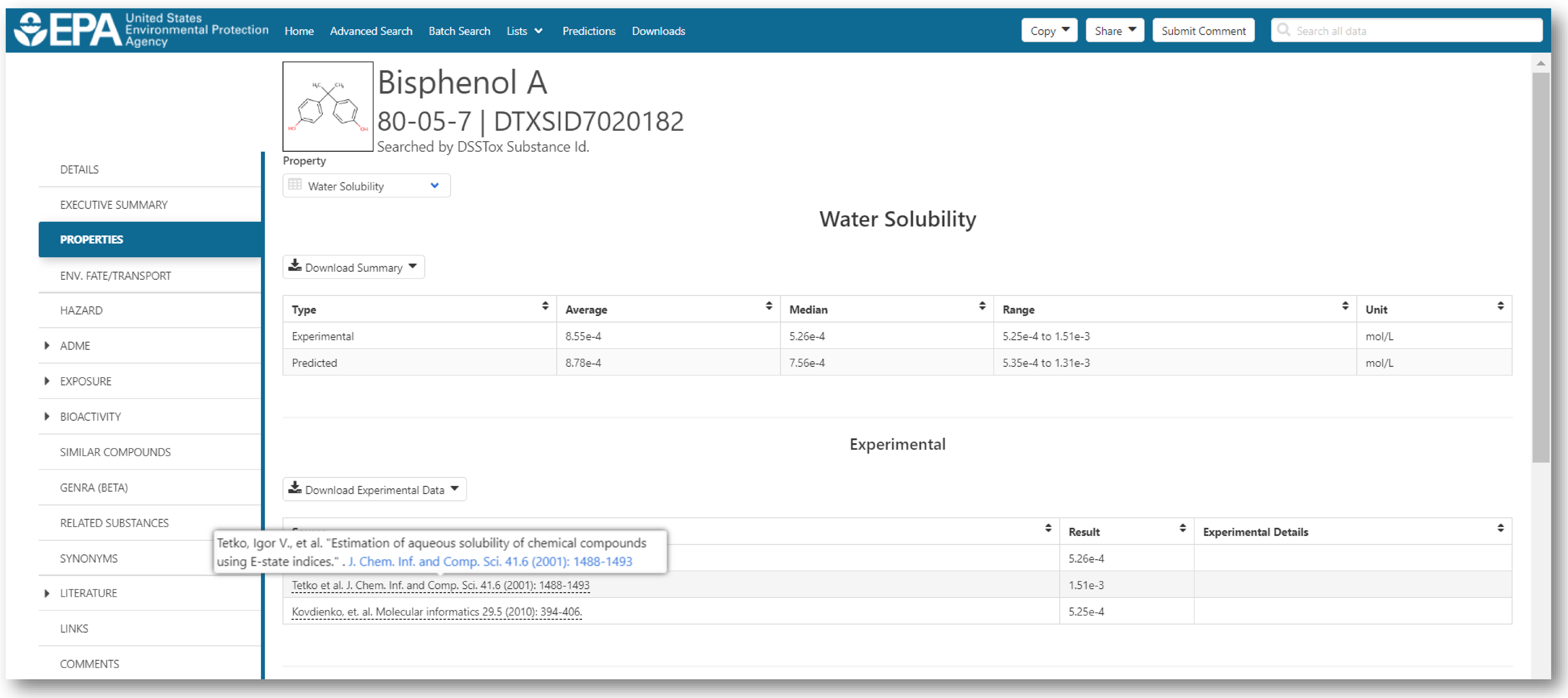




Above, Left: There are issues where stereochemical information is present in some chemical identifiers and not in others for a certain value. This can lead to issues for specific endpoints that are dependent on differences in stereochemistry. QSAR/QSPR models may not specifically take stereochemistry into account but registration of experimental data would.

Above, Right: Other problems include the inclusion of superfluous text such as leading zeros on CAS numbers and incorrect data types in certain data columns (e.g. invalid InChIKeys). Property data that has undergone a unit conversion can have problems such as too many significant digits and incorrect values due to improper entry of the original dataset.

Future Work



Left: Curated physicochemical property data in the CompTox Chemicals Dashboard appears in the *Experimental* Section under the *Properties* Tab for each property endpoint associated with that chemical.

- We continue to harvest and curate experimental data for display of multiple physicochemical and fate and transport properties on the Dashboard.
- Retrain existing QSAR models using this newly harvested property data and note any improvements this approach offers.
- New data harvested to date will be displayed in the March 2020 release of the dashboard.

References

1. Sushko, Iurii, et al. "Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information." Journal of computer-aided molecular design 25.6 (2011): 533-554.

Innovative Research for a Sustainable Future