Systematizing Data Gathering for PFAS Toxicity and Property Modeling

G. Sinclair[1]; C. Ramsland[1]; T. Martin[2]; A. Williams[3]

U.S. Environmental Protection Agency, Office of Research and Development

[1]Oak Ridge Associated Universities (contractors), RTP, NC

[2]Center for Computational Toxicology and Exposure, Cincinnati, OH

[3]Center for Computational Toxicology and Exposure, RTP, NC

Dr. Todd Martin | martin.todd@epa.gov | 000-000-0000

Objective

A vast amount of chemical toxicology and property data is publicly accessible via the Internet. However, these data are often uncurated, unreferenced, and distributed across many data sources. Despite the proliferation of data, certain classes of chemicals remain poorly characterized experimentally, notably including per- and polyfluoroalkyl substances (PFAS). This project seeks to develop a systematic approach to consolidate existing chemical data for use in quantitative structure-activity relationship (QSAR) modeling. This approach will increase the quality and quantity of data available to model the toxicology and properties of PFAS, as well as identifying present gaps and problems in data collection.

Method

Data sources for this project included academic (e.g. OCHEM), governmental (e.g. PubChem and eChemPortal), and commercial vendors (e.g. LookChem). Initially, all data were collected and stored in their original format. Using tailored processing tools developed in Java for the project, these data were translated to a structured intermediate JSON format retaining all original information. They were then standardized to a final JSON format highlighting specific properties of interest for QSAR modeling with normalized units, measurement methods, and remarks for each property. These standardized data points were integrated into a single SQL database. The combination of chemical identifiers for each data point was mapped to a single substance ID in the EPA's Distributed Structure-Searchable Toxicology Database (DTXSIDs).

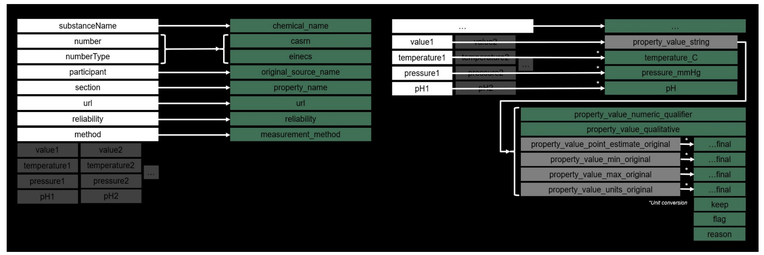

Figure 1: Data conversion and reformatting process for eChemPortal. White fields represent original data as stored in the intermediate format; gray fields are used for processing but do not appear in final database; green fields are final.

Data Summary

Approximately 2.25M candidate experimental data points were obtained across fourteen sources. 390k of these data points were eliminated as invalid. Major reasons for elimination are shown in **Table 1**.

Reason for Elimination	Data Points
Duplicate data points within the same data source	201k
Implausible data, missing or incorrect units	112k
Insufficient or incorrect identifying information	34k
Calculated, estimated, modeled, or extrapolated data	24k
Single batch of bad data in OCHEM	5k

Table 1: Top five categories of eliminated data points.

The remaining valid data points cover a total of 370k substances (as defined by an available and distinct CAS RN) for nine properties of interest (**Table 2**).

Property	Substances
Boiling point	318k
Flash point	315k
Density	311k
Melting point	56k
Octanol-water partition coefficient	21k
Water solubility	20k
Vapor pressure	10k
pKa	4.7k
Henry's law constant	2.8k

Table 2: Substance coverage over valid data points with available CAS RN, by property.

Results

A significant driver of this research is the need to identify the publicly-available data of greatest utility for future research in every area, be it for further data collection and analysis, laboratory investigation, or regulatory decision-making.

One major concern in establishing data source utility is the coverage of substances of interest (i.e. how many substances have data points included in a particular source). A related concern is uniqueness: a data source that contains only a few data points not found in other sources is of lower utility. In **Table 3**, we analyze the uniqueness of coverage of eleven data sources for physicochemical property data:

Overlap % (by CAS RN)	ADDoPT	ChemID Plus	eChem Portal	LookChem	OCHEM	OFMPub	OPERA	PubChem	QSARDB	Sander	EPI Suite	Unique %
ADDoPT		52.97%	40.00%	87.57%	97.48%	6.31%	88.11%	60.18%	51.89%	52.97%	93.87%	1.62%
ChemIDPlus	5.93%		28.94%	93.53%	97.76%	3.99%	88.10%	53.21%	12.65%	21.42%	44.29%	0.38%
eChemPortal	2.05%	13.24%		58.23%	27.89%	3.63%	20.78%	16.03%	3.21%	6.29%	10.79%	39.36%
LookChem	0.14%	1.33%	1.81%		7.35%	0.10%	4.57%	2.64%	0.30%	0.51%	1.29%	90.40%
OCHEM	1.59%	14.24%	8.88%	75.43%		0.91%	56.61%	17.55%	3.20%	5.24%	15.41%	9.25%
OFMPub	5.98%	33.85%	67.18%	58.63%	52.65%		43.08%	25.98%	9.74%	21.37%	27.52%	16.75%
OPERA	2.23%	19.96%	10.29%	72.93%	88.04%	1.15%		23.89%	4.79%	7.47%	20.67%	6.01%
PubChem	2.42%	19.13%	12.60%	66.93%	43.32%	1.10%	37.91%		4.90%	8.21%	24.20%	29.47%
QSARDB	26.40%	57.47%	31.90%	94.87%	99.91%	5.22%	96.06%	61.87%		59.85%	88.08%	0.09%
Sander	15.08%	54.49%	34.94%	92.30%	91.43%	6.41%	83.84%	58.08%	33.50%		73.42%	3.49%
EPI Suite	9.03%	38.07%	20.26%	78.37%	90.88%	2.79%	78.40%	57.84%	16.66%	24.80%		3.12%

Table 3: Data coverage overlap by source. Percentages in the table represent the percentage of compounds in the row source also covered by the column source. Percentages in the final column represent the percentage of compounds in the row source not covered by any other source. Three data sources that did not include CAS RN along with the data were not included in this analysis.

The data set from LookChem dominates in terms of uniqueness, with significant contributions also from PubChem and eChemPortal. In raw numbers, LookChem also provides the greatest coverage, with 349k distinct substances: more than ten times the second-largest source (OCHEM).

A further dimension of coverage and uniqueness—the similarity of specific properties and data points across sources—overlaps with the removal of duplicate data. In this project, exact duplicate data was removed when present in a single source. However, it is possible that exact duplicates exist across sources, or the same data points exist in different formats across sources. Developing tools to identify these cases and a strategy to productively address them, without discarding useful data, when constructing modeling data sets will be a further avenue of investigation in the future.

Conclusions

The increasing public availability of detailed chemical data is potentially a great boon for all fields of chemistry, especially in fields such as toxicology where modeling is a growing component of new research. However, the utility and reliability of these data may vary greatly, and specialized tools are needed to optimally access, understand, and make use of it.

This project generated a very large QSAR-usable data set of experimental values for physicochemical properties in addition to a systematic, extensible framework for incorporating other sources and types of data in the future. The creation of this data set allowed for the assessment of existing data sources from a bulk statistical viewpoint as well as identifying specific patterns and defects in each source, which led to correction of raw data in one source and development of programmatic data validation tools for several others.

An important aspect for the future of this project is the consideration of PFAS in particular. Sources with good overall coverage may lack PFAS coverage and vice versa. Additionally, certain property measurements are more apt to be inaccurate for PFAS compounds than for others. The validation and modeling of PFAS data will be an area of focus for the project moving forwards.

Additionally, coverage and uniqueness are far from the sole determiners of data set utility. Clearly, the accuracy of the provided data is of paramount importance. By collating records by substance and property and comparing them across sources, some obvious patterns of flaws begin to appear.

Consider **Figure 2**, which was generated by matching CAS RNs across water solubility records in OPERA and OCHEM and plotting the respective values. Simply by viewing the plot, it is obvious that a large number of records in OCHEM contain sign errors.

Ultimately, it was determined that a single user had uploaded a batch of 5,000 records with incorrect signs in April 2018. It seems this error went undetected for nearly three years. The database owner was informed and the records at issue were corrected immediately. This clearly demonstrates the importance of a project that incorporates data from as many sources as possible in order to discover and correct such errors in a programmatic fashion.

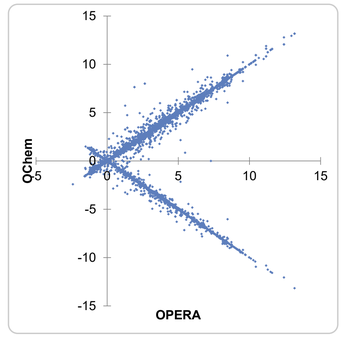

Figure 2: Substance-by-substance comparison of water solubility records in OPERA and OCHEM as -log10(mol/L).

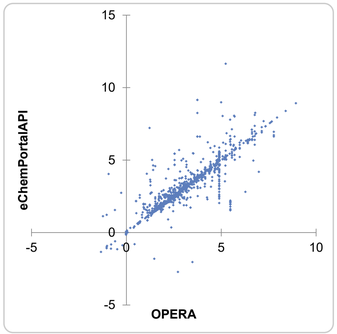

Figure 3: Substance-by-substance comparison of water solubility records in OPERA and eChemPortal as -log10(mol/L).

Figure 3 shows the same comparison made between OPERA and eChemPortal. Note the instances of vertical banding, indicating a spread of experimental values in eChemPortal where a single value is present in OPERA.

This was determined to be the result of eChemPortal data points wherein water solubility was measured at a range of ambient pH. These data points were detected by computing the standard deviation of experimental values; for the purposes of this project, they were then eliminated, since they do not represent pure water solubility measurements usable for QSAR modeling.

Data Sources Referenced

ADDoPT: https://www.ncbi.nlm.nih.gov/pubmed/27338156 (Table S1)
ChemIDPlus: https://chem.nlm.nih.gov/chemidplus/
eChemPortal: https://www.echemportal.org/echemportal/property-search
OFMPub: https://ofmpub.epa.gov/oppthpv/hpv_hc_characterization.get_report_by_cas?doctype=2
OPERA: Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform. 2018 Mar 8;10(1):10. doi: 10.1186/s13321-018-0263-1.
LookChem: https://www.lookchem.com/ **OCHEM**: https://ochem.eu/home/show.do
PubChem: https://pubchem.ncbi.nlm.nih.gov/ **QSARDB**: https://qsardb.org/repository/
Sander: http://satellite.mpic.de/henry/ **EPI Suite**: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm