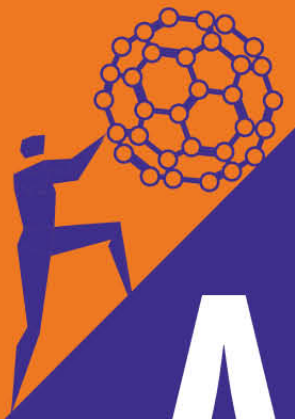
ACS
Chemistry for Life®

AMERICAN CHEMICAL SOCIETY

MEETINGS & EVENTS



ACS FALL 2021

August 22–26

RESILIENCE OF CHEMISTRY

#ACSFall2021

ACS Chemistry for Life®
American Chemical Society
MEETINGS & EVENTS

ACS FALL 2021
RESILIENCE OF CHEMISTRY

Systematic development of QSAR data sets from online data

#3586580

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

[1]EPA-ORD-CCTE-CCED-CCCB, Cincinnati, OH

[2]Oak Ridge Associated Universities, RTP, NC

[3]EPA-ORD-CCTE-CCED-CCCB, RTP, NC

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

(ORCID: 0000-0003-0802-2282)

What?

A systematic methodology for…

- Aggregation of raw data from extant compilations & literature
- Standardization of raw data format
- Validation of identifiers
- Preparation of QSAR data sets

Why?

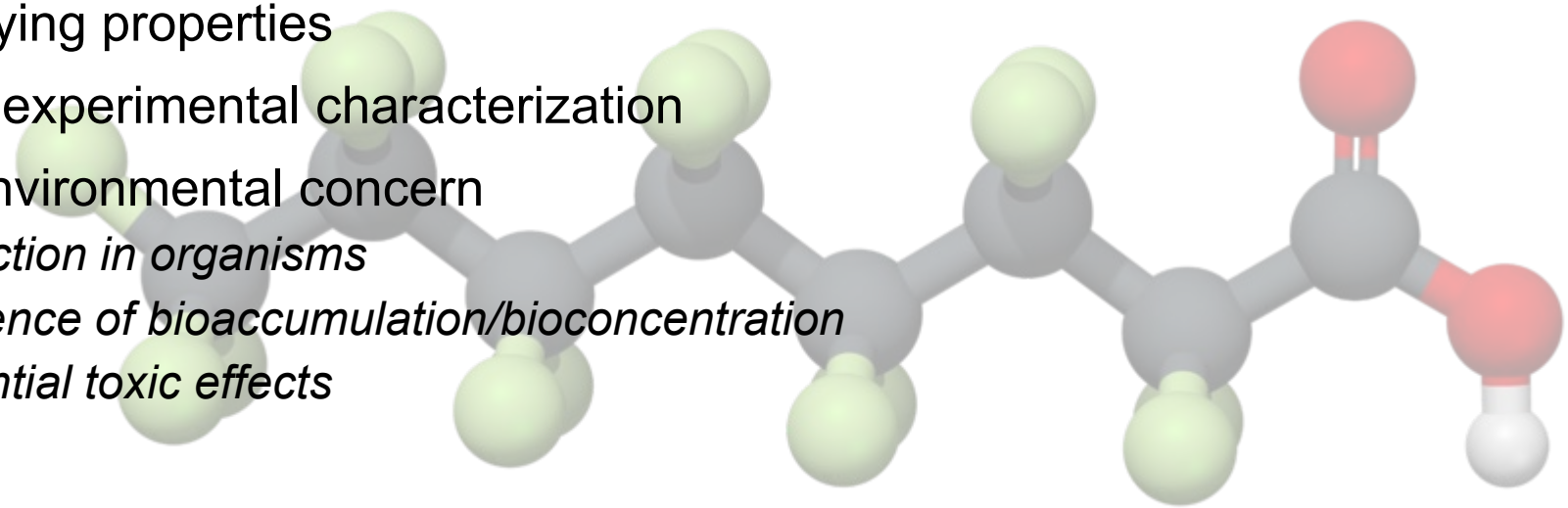
To advance PFAS modeling capabilities by…

- Expanding available PFAS experimental data sets
- Facilitating comparison of local vs. global modeling
- Facilitating comparison of applicability domains
- Enabling application of novel machine learning methods

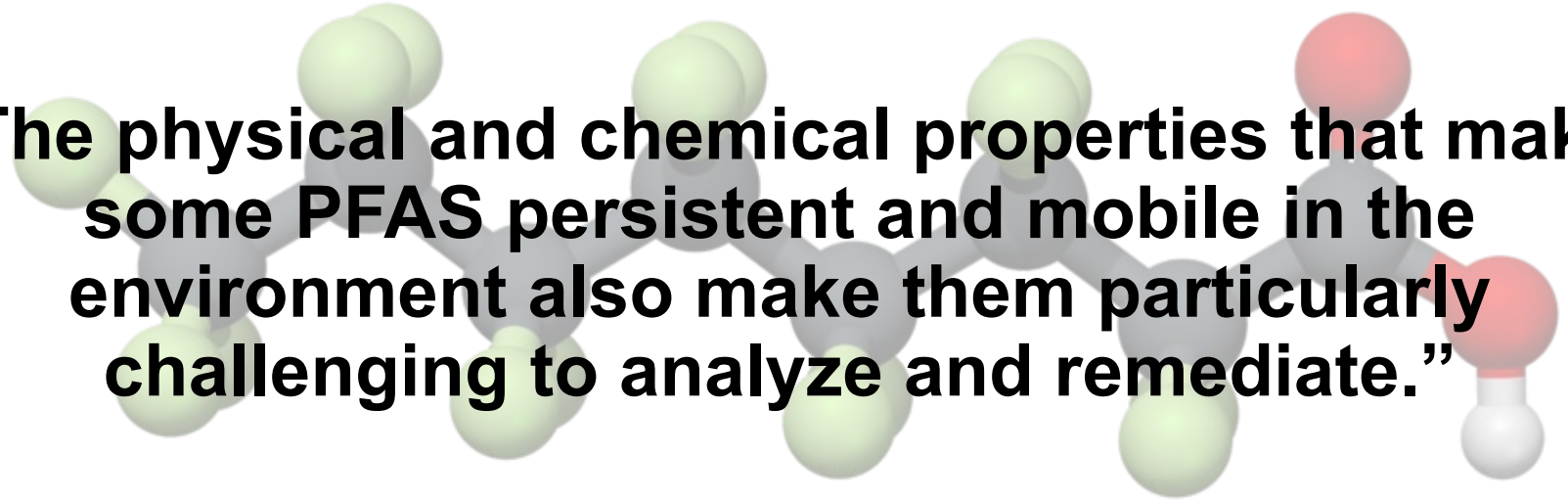
Per- & Polyfluoroalkyl Substances

A large, structurally-diverse family of fluorinated chemicals with…

- Widely varying properties
- Barriers to experimental characterization
- Growing environmental concern
    - *Detection in organisms*
    - *Evidence of bioaccumulation/bioconcentration*
    - *Potential toxic effects*



("Introduction" 2021, Rivas 2016)



“The physical and chemical properties that make some PFAS persistent and mobile in the environment also make them particularly challenging to analyze and remediate.”

(“Introduction” 2021, Rivas 2016)

Quantitative Structure-Activity Relationship

Structural descriptor-based prediction of chemical data by a process of…

- *Experimental data collection*
- Molecular representation
- Model training
- Model validation
- Applicability domain analysis
- *Interpretation*

(Gramatica 2020)

“The first point, and sometimes the most challenging in QSAR, is that QSAR modellers need experimental data as input for their models.”

(Gramatica 2020)

Method

Experimental data compilation:

- Extract raw physicochemical property data from public compilations
- Parse data into machine-readable intermediate format (JSON)
- Translate intermediate format into unified final format (JSON)
- Merge data into single database (SQL)

QSAR data set creation:

- Filter data for experimental validity & QSAR relevance
- Obtain molecular structural data (DSSTox)
- Re-filter data for structural QSAR relevance

Raw Data Extraction

Properties of interest:

- Melting point
- Boiling point
- Density
- Flash point
- Water solubility
- Octanol-water partition coefficient
- pKa
- Vapor pressure
- Henry's law constant

- All code in Java
- Different interfaces
  - *Single or batch file downloads*
  - *Native APIs*
  - *API wrappers*
  - *HTML scraping*
- Different data formats
  - *HTML*
  - *JSON*
  - *Excel*

ADDoPT Original Records.json - Notepad

File Edit Format View Help

Intermediate Data Parsing


```
[
  {
    "cas": "100-00-5",
    "solubility": "-3.01",
    "temp": "10"
  },
  {
    "cas": "100-00-5",
    "solubility": "-2.88",
    "temp": "20"
  },
  {
    "cas": "100-00-5",
    "solubility": "-2.76",
    "temp": "30"
  },
  {
    "cas": "100-00-5",
    "solubility": "-2.63",
    "temp": "40"
  },
  {
    "cas": "100-01-6",
    "solubility": "-2.5499999999999998",
    "temp": "20"
  },
```

OFMPub Original Records.json - Notepad

File Edit Format View Help

Intermediate Data Parsing


```
[
  {
    "endpoint": "Melting Point",
    "categoryChemicalCAS": "88-73-3",
    "categoryChemicalName": "Benzene, 1-chloro-2-nitro-",
    "testSubstanceCAS": "88-73-3",
    "testSubstanceName": "Benzene, 1-chloro-2-nitro-",
    "testSubstanceComments": "o-Chloronitrobenzene CAS No 88-73-3",
    "categoryChemicalResultType": "",
    "testSubstanceResultType": "",
    "indicator": "",
    "value": "32.5 °C",
    "resultRemarks": "",
    "reference": "Verschueren, K. 1996. Handbook of environmental data on organic chemicals. 3rd Ed. New York, NY. Van Nostrand Reinholc
    "reliability": "Valid with Restrictions",
    "reliabilityRemarks": "Obtained from accepted reference text and value cited as Peer reviewed in HSDB (2002) for o-chloronitrobenzer
    "url": "https://ofmpub.epa.gov/oppthpv/Public_Search.PublicTabs?section=1&SubmissionId=24966103&epcount=1&epname=Melting+Point&epdis
    "date_accessed": "11/27/2020"
  },
  {
    "endpoint": "Melting Point",
    "categoryChemicalCAS": "100-00-5",
    "categoryChemicalName": "Benzene, 1-chloro-4-nitro-",
    "testSubstanceCAS": "100-00-5",
    "testSubstanceName": "Benzene, 1-chloro-4-nitro-",
    "testSubstanceComments": "p-Nitrochlorobenzene CAS NO 100-00-5",
    "categoryChemicalResultType": "",
```

Final Data Translation

100000	100000	100000	100000	100000	100000	100000	100000	100000	##	83508	6473	6473	89981	25732	1690	9980	6009
id_physche	keep	casrn	einecs	chemic	synony	source_n	property_nam	property_value_string	p	proper	pro	pro	proper	pressu	temper	proper	note
LookChem1	true	10-00-4		4-[[5,7-dih	C28H34O8	LookChem	Density	1.284g/cm3		1.284			g/cm3				
LookChem2	true	10-00-4		4-[[5,7-dih	C28H34O8	LookChem	Boi	764.5°C at 760 mmHg		764.5			C	760			
LookChem3	true	10-00-4		4-[[5,7-dih	C28H34O8	LookChem	Flas			3.2			C				
LookChem4	true	11-30-3		rel-(2R,3a		LookChem	Den			.43			g/cm3				
LookChem5	true	11-30-3		rel-(2R,3a		LookChem	Boiling point	971.4°Cat760mmHg		971.4			C	760			
LookChem6	true	11-30-3		rel-(2R,3a		LookChem	Flash point	288.1°C		288.1			C				
LookChem7	true	19-15-8		8-[[3-[[3-[	8-[[3-[[3-[	LookChem	Density	1.783g/cm3		1.783			g/cm3				
LookChem8	true	23-13-2		(hydroxy-	(hydroxy-	LookChem	Density	1.01g/cm3		1.01			g/cm3				
LookChem9	true	23-13-2		(hydroxy-	(hydroxy-	LookChem	Boiling point	891.4°C at 760 mmHg		891.4			C	760			
LookChem10	true	23-13-2		(hydroxy-	(hydroxy-	LookChem	Flash point	492.9°C		492.9			C				
LookChem11	true	37-87-6		b-D-Gluco	Glucopyra	LookChem	Density	1.695g/cm3		1.695			g/cm3				
LookChem12	true	37-87-6		b-D-Gluco	Glucopyra	LookChem	Boiling point	736.3°Cat760mmHg		736.3			C	760			
LookChem13	true	37-87-6		b-D-Gluco	Glucopyra	LookChem	Flash point	399.1°C		399.1			C				
LookChem14	true	38-84-6		(2S,3R)-2-		LookChem	Density	1.201g/cm3		1.201			g/cm3				
LookChem15	true	38-84-6		(2S,3R)-2-		LookChem	Boiling point	1112.5°Cat760mmHg		1112.5			C	760			
LookChem16	true	38-84-6		(2S,3R)-2-		LookChem	Flash point	626.6°C		626.6			C				
LookChem17	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Density	1.09 g/mL at 25 °C		1.09			g/cm3		25		
LookChem18	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Melting point	-15 °C		-15			C				
LookChem19	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Boiling point	97 °C		97			C				
LookChem20	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Flash point	133 °F		56.11111			C				
LookChem21	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Water solubility	soluble in water								soluble	
LookChem22	true	50-00-0	200-001-8	Formaldel	FM 282|Fa	LookChem	Appearance	Clear liquid								clear liquid	
LookChem23	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Density	1.18 g/mL at 25 °C(lit.)		1.18			g/cm3		25		literature
LookChem24	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Melting point	180-185 °C(lit.)			180	185	C				literature
LookChem25	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Boiling point	132.9 °C at 760 mmHg		132.9			C	760			
LookChem26	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Flash point	34.2 °C		34.2			C				
LookChem27	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Water solubility	2280 g/L (20 °C) in water		2280			g/L		20		
LookChem28	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Appearance	White cryst. powder								white cryst. powder	
LookChem29	true	50-02-2	200-003-9	Dexameth	Diodex|Et	LookChem	Density	1.32 g/cm3		1.32			g/cm3				
LookChem30	true	50-02-2	200-003-9	Dexameth	Diodex|Et	LookChem	Melting point	262-264 °C(lit.)			262	264	C				literature
LookChem31	true	50-02-2	200-003-9	Dexameth	Diodex|Et	LookChem	Boiling point	568.2 °C at 760 mmHg		568.2			C	760			
LookChem32	true	50-02-2	200-003-9	Dexameth	Diodex|Et	LookChem	Flash point	297.5 °C		297.5			C				

Final Data Translation

Identifies and reformats numbers and ranges

Identifies and reformats equivalent units

Converts units

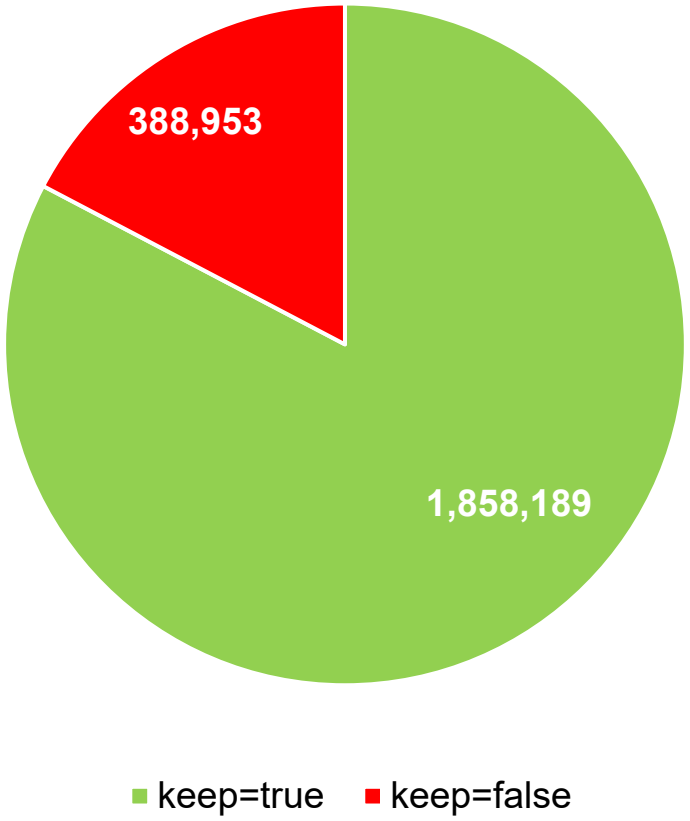
Identifies ambient conditions

Identifies qualitative entries and other notes

LookChem17	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Density	1.09 g/mL at 25 °C		1.09		g/cm3		25		
LookChem18	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Melting point	-15 °C		-15		C				
LookChem19	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Boiling point	97 °C		97		C				
LookChem20	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Flash point	133 °F		56.11111		C				
LookChem21	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Water solubility	soluble in water							soluble	
LookChem22	true	50-00-0	200-001-8	Formaldel	FM 282 | Fa	LookChem	Appearance	Clear liquid							clear liquid	
LookChem23	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Density	1.18 g/mL at 25 °C (lit.)		1.18		g/cm3		25		literature
LookChem24	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Melting point	180-185 °C(lit.)			180 185	C				literature
LookChem25	true	50-01-1	200-002-3	Guanidine	Guanidine	LookChem	Boiling point	132.9 °C at 760 mmHg		132.9		C	760			

Experimental Data

- 368,992 distinct CAS RNs



Experimental Data

- 368,992 distinct CAS
- 16 sources

Source	Records	Distinct CAS
LookChem	996,515	349,882
OChem	522,726	34,034
PubChem	164,848	14,436
eChemPortalAPI	89,995	10,838
OPERA	37,147	21,881
ChemIDplus	11,140	4,958
AqSolDB	9,981	9,890*
Sander	6,818	1,949
EpisuiteISIS	5,779	5,779
+6 others	*13,240*	

Experimental Data

- 368,992 distinct CAS RNs
- 16 sources
- 9 physicochemical properties

Property	Records	Distinct CAS
Melting point	422,763	55,586
Boiling point	364,301	318,422
Density	361,078	310,584
Flash point	335,460	314,739
Water solubility	160,296	25,940
Octanol water partition coefficient	98,420	20,912
Vapor pressure	41,438	10,459
pKA	14,978	2,043
Henry's law constant	11,213	2,818

Data Set Creation

- Query for relevant data points (property, source) using SQL
- Filter on experimental data:
  - *Exclude implausible data*
  - *Exclude or average ranges*
  - *Exclude data qualified by ~, <, >*
  - *Select ambient pressure, temperature, pH of interest*
- Match molecular structures in DSSTox
- Filter on structural data:
  - *Merge duplicates & isomers on connectivity (median or 80% consensus value)*
  - *Excessively high stdev*
  - *Missing structures*
  - *Salts, inorganics, non-QSAR-compatible elements, etc.*

Data Set Creation

Henry's law constant:

- 11,213 total records gathered
- 10,261 records after filtering for appropriate experimental conditions & data
- 2,170 records after mapping in DSSTox
- 1,032 records after merging & filtering structures
- *Outlier detection, data set splitting, & modeling!*

Related Presentations

- *Development of models to predict physicochemical properties of PFAS*, presented by Dr. Todd Martin
- *Development of skin sensitization, skin irritation, and eye irritation models using online data sources and Python-based machine learning*, presented by Christian Ramsland

CompTox Chemicals Dashboard

883 Thousand Chemicals

Chemicals | Product/Use Categories | Assay/Gene

Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey

Identifier substring search

See what people are saying, read the dashboard comments!
Cite the Dashboard Publication click here

Latest News

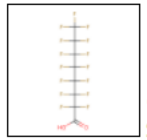
Read more news

New article to help understand the Batch Search functionality published

March 22nd, 2021 at 4:21:30 PM

A new article regarding the batch search on the Dashboard is described in a recent article in the Journal of Chemical Information and Modeling: Enabling High-Throughput Searches for Multiple Chemical Data Using the U.S.-EPA CompTox Chemicals Dashboard.

CompTox.epa.gov/dashboard



Perfluorooctanoic acid

335-67-1 | DTXSID8031865

Searched by DSSTox Substance Id.

Property

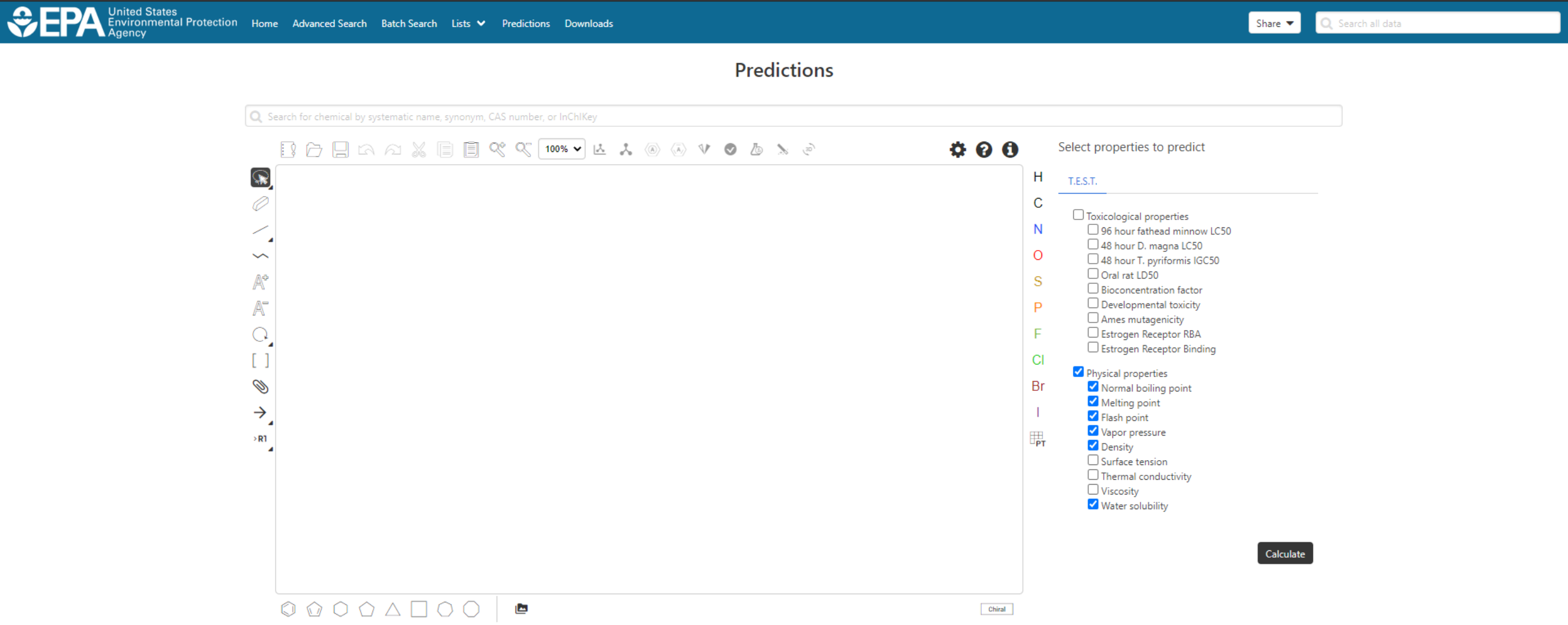
Summary

Summary

Download | Columns | Search query

Property	Experimental average	Predicted average	Experimental median	Predicted median	Experimental range	Predicted range	Unit
LogKow: Octanol-Water	3.10 (5)	5.68	3.60	5.94	1.92 to 3.60	3.11 to 7.75	-
Melting Point	56.1 (20)	24.3	55.5	27.3	47.5 to 59.5	-8.69 to 54.2	°C
Boiling Point	190 (17)	193	189	191	188 to 199	188 to 204	°C
Water Solubility	1.37e-2 (15)	0.753	1.00e-2	1.66e-2	8.21e-3 to 2.50e-2	6.27e-8 to 2.98	mol/L
Thermal Conductivity	-	65.3			-	65.3	mW/(m*K)
Flash Point	-	68.0		68.0	-	62.1 to 73.9	°C
Vapor Pressure	0.952 (13)	0.243	3.90e-2	0.274	1.65e-2 to 10.0	0.111 to 0.345	mmHg
Density	1.80 (1)	1.72		1.72	1.80	1.70 to 1.75	g/cm^3
Surface Tension	-	16.8			-	16.8	dyn/cm
Index of Refraction	-	1.29			-	1.29	-
Molar Refractivity	-	42.9			-	42.9	cm^3
Polarizability	-	17.0			-	17.0	Å^3
Molar Volume	-	237			-	237	cm^3
LogKoa: Octanol-Air	-	4.16			-	4.16	-
Henry's Law	-	1.92e-10			-	1.92e-10	atm-m3/mole
pKa Acidic Apparent	3.15 (2)		3.15		2.50 to 3.80	-	-

CompTox.epa.gov/dashboard



CompTox.epa.gov/dashboard

Acknowledgments

EPA-ORD-CCTE-CCED-CCCB

- Dr. Todd Martin
- Dr. Antony Williams
- Dr. Charles Lowe
- Dr. Ann Richard, Dr. Chris Grulke, & ChemReg Project
- CompTox Chemicals Dashboard Project

Oak Ridge Associated Universities

- Christian Ramsland

Thank you! Questions?

Presented by Gabriel Sinclair
sinclair.gabriel@epa.gov

Related Presentations

- *Development of models to predict physicochemical properties of PFAS*, presented by Dr. Todd Martin
- *Development of skin sensitization, skin irritation, and eye irritation models using online data sources and Python-based machine learning*, presented by Christian Ramsland

Abbreviations & Acronyms

- **API**: Application Programming Interface
- **DSSTox**: Distributed Structure-Searchable Toxicology Database
- **DTXSID**: DSSTox Substance ID
- **JSON**: JavaScript Object Notation
- **PFAS**: Per- & Polyfluoroalkyl Substances
- **QSAR**: Quantitative Structure-Activity Relationship
- **SQL**: Structured Query Language

Full Abstract

“A vast amount of chemical toxicology and property data is publicly accessible via the Internet. However, these data are often uncurated, unreferenced, distributed across many data sources, and can contain a myriad of data quality issues. This project sought to develop a systematic approach to consolidate existing chemical data for use in quantitative structure-activity relationship (QSAR) modeling. A large compilation of physicochemical data (>2 million data points) was collected from 16 publicly available sources using automated tools built in Java. These data were converted to a consistent machine-readable format, and stored in an SQLite database. The use of SQL queries allowed for the convenient assembly of data subsets by characteristics such as experimental property, conditions, and test methods. The experimental data were filtered for QSAR validity (e.g. eliminating implausible property values and constraining experimental ambient conditions), and substances were mapped to unique substance identifiers (DTXSIDs) using the EPA’s Distributed Structure-Searchable Toxicology (DSSTox) Database to obtain structural data. The structural data were filtered again for QSAR validity (e.g. removing salts and metallic atoms) and used to generate molecular descriptor values. Finally, records were stored in “QSAR-ready” form (i.e. desalted non-stereoforms with isotopes removed) for use as input to a variety of existing and newly developed QSAR models. The development of automated data collection tools, as well as web services called via an application programming interface (API) for individual steps of data preparation and modeling, created a generalizable workflow. This workflow could be applied to any type of experimental data; for any set or subset of substances of interest; with any desired constraints, descriptors, and methods for modeling. The effectiveness and generality of this system was demonstrated through data gathering and modeling efforts on water solubility, skin sensitization, skin irritation, and eye irritation. The views expressed here are those of the authors and do not necessarily represent the views or the policies of the U.S. Environmental Protection Agency.”

References

1. *CAS No. 307-24-4*. LookChem.com. (2021). Retrieved 5 August 2021, from https://www.lookchem.com/cas-307/307-24-4.html.
2. Gramatica, P. (2020). Principles of QSAR Modeling. *International Journal Of Quantitative Structure-Property Relationships*, *5*(3), 61-97. https://doi.org/10.4018/ijqspr.20200701.oa1
3. *Introduction*. PFAS — Per- and Polyfluoroalkyl Substances. (2021). Retrieved 4 August 2021, from https://pfas-1.itrcweb.org/1-introduction/.
4. *PFASSTRUCT Chemicals*. CompTox Chemicals Dashboard. (2021). Retrieved 4 August 2021, from https://comptox.epa.gov/dashboard/chemical_lists/PFASSTRUCT.
5. Rivas, M. (2016). *3D model of a PFOA (perfluorooctanoic acid) molecule, in its acid form* [Image]. Retrieved 4 August 2021, from https://commons.wikimedia.org/wiki/File:PFOA-3D.png.

ACS
Chemistry for Life®

AMERICAN CHEMICAL SOCIETY
MEETINGS & EVENTS

REPORT INSTANCES OF HARASSMENT

Contact ACS Secretary and General Counsel,
Flint Lewis at f_lewis@acs.org or

call the ACS ANONYMOUS HOTLINE:

toll-free at **855-710-0009 (English)**
or **800-216-1288 (Spanish)**