United States Environmental **Protection Agency**

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 I lowe.charles@epa.gov I 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.

Abstract

The U.S. Environmental Protection Agency's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard) hosts a plethora of environmentallyrelevant 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.



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/).

Issues Discovered During Aggregation and Curation

SMILES 💌	CASRN -	NAME	ARTICLE -	PUBMED -	Water 🔄 UNIT {Wa 💌	Water 💌 UNIT {W 💌
OC(C1=CC=CC=C1)C1=CC=CC=C1	91-01-0	benzhydrol	A111278	NA	-2.5494 log(mol/L)	-2.5494 log(mol/L)
C1CCC2CCCC2C1	91-17-8	DECAHYDRONAPHTHALENE	A111278	NA	-5.1918 log(mol/L)	-5.1918 log(mol/L)
C1=C[NH0]=C2[NH0]=C[NH0]=CC2=[N	91-18-9	pteridine	A111278	NA	0.0343 log(mol/L)	0.0343 log(mol/L)
C1=CC2=[NH0]C=C[NH0]=C2C=C1	91-19-0	Quinoxaline	A111278	NA	0.7051 log(mol/L)	0.7051 log(mol/L)
C1=CC=C2[NH0]=CC=CC2=C1	91-22-5	Quinoline	A111278	NA	-1.3251 log(mol/L)	-1.3251 log(mol/L)
COC1=CC=CC=C1[N+]([O-])=O	91-23-6	2-nitroanisole	A111278	NA	-1.9572 log(mol/L)	-1.9572 log(mol/L)
CC1=CC=C2C=CC=CC2=C1	91-57-6	2-methylnaphthalene	A111278	NA	-3.762 log(mol/L)	-3.762 log(mol/L)
NC1=CC2=CC=CC=C2C=C1	91-59-8	2-NAPHTHYLAMINE	A111278	NA	-2.8795 log(mol/L)	-2.8795 log(mol/L)
O=C1OC2=CC=CC=C2C=C1	91-64-5	Coumarin	A111278	NA	-1.886 log(mol/L)	-1.886 log(mol/L)
CCN(CC)C1=CC=CC=C1	91-66-7	n,n-diethylaniline	A111278	NA	-3.0278 log(mol/L)	-3.0278 log(mol/L)
[#7]C1=[NH0]C(=[NH0]C([#7])=[NH0]:	191-76-9	BENZOGUANAMINE	A111278	NA	-2.4942 log(mol/L)	-2.4942 log(mol/L)
[#6]N([#6])CCN(CC1=CC=C[SH0]1)C1=	91-80-5	methapyrilene	A111278	NA	-2.6383 log(mol/L)	-2.6383 log(mol/L)
NC1=CC=C(C=C1Cl)C1=CC=C(N)C(Cl)=	91-24-1	3,3'-DICHLOROBENZIDINE	A111278	NA	-4.912 log(mol/L)	-4.912 log(mol/L)
OC1=CC=C(C=C1O)C1=CC=CC=C1	92-05-7	4-PHENYLCATECHOL	A111278	NA	-2.0659 log(mol/L)	-2.0659 log(mol/L)
C1=CC=C(C=C1)C1=CC(=CC=C1)C1=CC	92-06-8	-TERPHENYL	A111278	NA	-5.183 log(mol/L)	-5.183 log(mol/L)
OC1=C(O)C=C2C=CC=CC2=C1	92-44-4	2,3-INAPHTHALENEDIOL	A111278	NA	-2.7377 log(mol/L)	-2.7377 log(mol/L)
CCN(CCO)C1=CC=CC=C1	92-50-2	ETHANOL, 2-(ETHYLPHENYLAMINO)-	Δ111278	NΔ	-1 9171 log(mol/L)	-1.9171 log(mol/L)
C1=CC=C(C=C1)C1=CC=CC=C1	92-52-4	biphenyl			468 log(mol/L)	-4.3468 log(mol/L)
BrC1=CC=C(C=C1)C1=CC=CC=C1	92-66-0	4-BROMOBIPHENYL SMAR	TS re	ecorde	526 log(mol/L)	-5.5526 log(mol/L)
OC1=CC=C(C=C1)C1=CC=CC=C1	92-69-3	p-phenylphenol	01.411	=0	812 log(mol/L)	-3.4812 log(mol/L)
N1C2=CC=CC=C2SC2=CC=CC=C12	92-84-2	Phenothiazine 28	SMIL	ES	981 log(mol/L)	-5.0981 log(mol/L)
CC(=O)C1=CC=C(C=C1)C1=CC=CC=C1	92-91-1	4-ACETYLBIPHE C			104 log(mol/L)	-3.3104 log(mol/L)
[O-][N+](=O)C1=CC=C(C=C1)C1=CC=C	92-93-3	P-NITROBIE PLNYL	A111278	NA	-5.2094 log(mol/L)	-5.2094 log(mol/L)
C1=CC=C(C=C1)C1=CC=C(C=C1)C1=CC	92-94-4	1,1':4' 1 '-7 ERPHENYL	A111278	NA	-7.107 log(mol/L)	-7.107 log(mol/L)
CC(=0)C1=CC=C2C=CC=CC2=C1	93-08-3	1-(2-NAPHTHALENYL)ETHANONE	A111278	NA	-2.7964 log(mol/L)	-2.7964 log(mol/L)
OC(=O)C1=CC2=CC=CC=C2C=C1	93-09-4	2-MAPHTHOIC ACID	A111278	NA	-3.5639 log(mol/L)	-3.5639 log(mol/L)
COC1=C(OCC(O)CO)C=CC=C1	93-14-1	1,2-Propanediol, 3-(2-methoxyphenoxy)-	A111278	NA	-0.5982 log(mol/L)	-0.5982 log(mol/L)
COC1=C(OC)C=C(CC=C)C=C1	9° 15-2	Methyleugenol	A111278	NA	-2.552 log(mol/L)	-2.552 log(mol/L)
[#6]C1=CC2=[NH0]C([#6])=CC=C2C=C1	193-37-8	Quinoline, 2,7-dimethyl-	A111278	040		ol/L
COC(=0)C1=CC=CC=C1	°3-58-3	Benzoic acid, methyl ester	A111278	CAS	SRN record	
[#6]OC(=O)C1=CC=C[NH0]=C1	93-60-7	Nicotinic acid, methyl ester	A111278	0	hemical Na	ol/L
CC(OC1=CC=C(Cl)C=C1C)C(O)=O	NA	MECOPROP, MCPP, Mechanop-P	A111278		nemical Na	
CC(OC1=CC(Cl)=C(Cl)C=C1Cl)C(O)=O	NA	93-72-1, 30365-50-5	A111278	NA	-3.4667 log(mol/L)	-3.4667 log(mol/L)
OC(=O)COC1=CC(Cl)=C(Cl)C=C1Cl	93-76-5	2,4,5-trichlorophenoxyacetic acid	A111278	NA	-2.9633 log(mol/L)	-2.9633 log(mol/L)
OC(=O)CCCOC1=CC(CI)=C(CI)C=C1CI	93-80-1	2,4,5-TB	A111278	NA	-3.8294 log(mol/L)	-3.8294 log(mol/L)
CC(=O)CC(=O)C1=CC=CC=C1	93-91-4	1,3-Butanedione, 1-phenyl-	A111278	NA	-2.6268 log(mol/L)	-2.6268 log(mol/L)
O=C(OC(=O)C1=CC=CC=C1)C1=CC=CC	93-97-0	BENZOIC ACID, ANHYDRIDE	A111278	NA	-4.3546 log(mol/L)	-4.3546 log(mol/L)

SMILES Left: Issues like being represented instead by SMARTS and chemical containing names 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 truncated chemical names with significant information loss.

									_		_
SMILES	CASRN	NAME	ARTICLE	▼ PUBMED ▼	Water 💌 UNIT {Wate	er 💌 Water 💌 UNIT {W 💌 Tempe	UNIT {T	 Ionic st 	UNIT {I	💌 comme	▼ pH
CC(C)COC(=0)C1=C(C=CC=C1C(O):	=O) 6744-88-3	Phtharc acid, 3-nitro-, 2-isobutyl ester	A5643	NA	-3.0467 log(mol/L)	-3.0467 log(mol/L)	20 °C	NA	NA	NA	NA
CC(C)CCOC(=0)C1=C(C=CC=C1C(O)=0 6744-92-9	Phthalic acid, 3-nitro-, isopentyl ester	A5643	NA	-3.0689 log(mol/L)	-3.0689 log(mol/L)	20 °C	NA	NA	NA	NA
CC1=CC(C)=C(C(C)=C1)P(O)=O	6781-97-1	phosphinic acid, mesityl-	A5643	NA	-2.7896 log(mol/L)	-2.7896 log(mol/L)	25 °C	NA	NA	NA	NA
OCC(CO)OCC=C	6806-76-4	1,3-Propanediol, 2-(2-propenyloxy)-	A5643	NA	0.481 log(mol/L)	0.481 log(mol/L) NA	-	NA	NA	NA	NA
CN(C(=N)NN(=O)=O)N(=O)=O	9/9/681	0 Capidine. N-methyl-N,N'-dinitro-	A5643	NA	-0.6104 log(mol/L)	-0.6104 log(mol/L)	25 °C	NA	NA	NA	NA
CCC(C)NC(N)=S	6814-99-9	Thiourea, (1-methylpropyl)-	A5643	NA	-1.1391 log(mol/L						NA
CCOC(=0)C1=C(C=CC=C1C(0)=0)N	I(=C 6828-46-2	Phthalic acid, 3-nitro-, 2 ethyl ester	A5045	NIA	-1.9163 log(mol/L						NA
CCCCOC(=0)C1=C(C=CC=C1C(0)=C)N 6828-47-3	Phthalic acid, 3-nitro-, 2-butyl ester	A5643	NA	-2.7193 log(mer	CAS numbers	conve	erted	to c	lates	NA
CIC1=CC=C(OC2=CC(CI)=CC=C2)C=	C1 6842-62-2	3,4'-Dichlorodiphenyl ether	AJ040	NA	-4.7701 log(mol/L	one manipere	001111	SILCA	.00		NA
NC(CC1=CC(I)=C(OC2=CC(I)=C(O)C	=C <mark>2 2/3/68</mark> 9	3 C (+-ITURUXY-3-IODOPHENYL)-3,5-DIIODO-L-TYROSINE	1.5043	NA	-5.2161 log(mol/L						NA
NC(CCC(O)=O)C(O)=O	6893-26-1	D-glutamic acid	A5643	NA	-1.2194 log(mol/L)	-1.2194 log(mol/L)	25 °C	NA	NA	NA	NA
NC(CO)C(O)=O	6898-95-9	SERINE	A5643	NA	-0.3861 log(mol/L)	-0.3861 log(mol/L)	20 °C	NA	NA	NA	NA
NC(CCC(O)=O)C(O)=O	5/4/689	9 Ciutamic acid	A5643	NA	-0.9916 log(mol/L)	-0.9916 log(mol/L)	20 °C	NA	NA	NA	NA
CIC1=CC=C(OC2=C(CI)C=CC=C2)C=	C1 6903-65-7	2,4'-DICHLORODIPHENYL ETHER	A5643	NA	-5.52 log(mol/L)	-5.52 log(mol/L)	25 °C	NA	NA	NA	NA
CC(=O)NC(NC(C)=O)C1=CC=CC=C1	6907-68-2	Acetamide, N,N'-(phenylmethylene)bis-	A5643	NA	-1.3144 log(mol/L)	-1.3144 log(mol/L)	20 °C	NA	NA	NA	NA
NC1=CC=C(C=C1)S(=O)(=O)NC1=C	NC(6912-98-7	Benzenesulfonamide, 4-amino-N-(1,2,3,4-tetrahydr	A5643	NA	-2.764 log(mol/L)	-2.764 log(mol/L)	37 °C	NA	NA	NA	NA
COC1=C(Cl)C(Cl)=CC(Cl)=C1Cl	6936-40-9	BENZENE, 1,2,4,5-TETRACHLORO-3-METHOXY-	A5643	NA	-5.1307 log(mol/L)	-5.1307 log(mol/L)	25 °C	NA	NA	NA	NA
NC(=0)N\N=C\C(0)C(0)C(0)C(0)	CO 6936-69-2	D-Galactose, (aminocarbonyl)hydrazone	15643	NA	-0.4257 log(mol/L)	-0.4257 log(mol/L)	21 °C	NA	NA	NA	NA
CCCCOC(=O)C1=CC=CN=C1	6/3/693	8 NICOTINIC ACID, BUTYL ESTER	A5642	NA	-1.8367 log(mol/L)	-1.8367 log(mol/L)	25 °C	NA	NA	NA	NA
CC(SC(C)C(N)=O)C(N)=O	6944-30-5	Propanamide, 2,2'-thiobis-	A5643	NA	-0.9239 log(mol/L)	-0.9239 log(mol/L)	18 °C	NA	NA	NA	NA
OCCCOC(=O)C1=CC=CC=C1	6946-99-2	1,3-Propanediol, monobenzoate	A5643	NA	-1.2558 log(mol/L)	-1.2558 log(mol/L) NA	-	NA	NA	NA	NA
CC(=0)CC(=0)OC1CCCCC1	6947-02-0	Butanoic acid, 3-oxo-, cyclohexyl ester	A5643	NA	-2.9037 log(mol/L)	-2.9037 log(mol/L)	35 °C	NA	NA	NA	NA
CSC1=NC=NC2=C1N=CC=N2	6966-78-5	Pteridine, 4-(methylthio)-	A5643	NA	-2.365 log(mol/L)	-2.365 log(mol/L)	20 °C	NA	NA	NA	NA
CCC(C)C(N)C(O)=O	9/3/700	4 Isoleucine	A5643	NA	-0.0525 log(mol/L)	-0.6525 log(mol/L)	20 °C	NA	NA	NA	NA
CN(C)C(C)(C)CO	7005-47-2	DMAMP	A5643	NA	0.9311 log(mol/L)	0.9311 log(mol/L)	20 °C	NA	NA	NA	NA
NC(CC(N)=O)C(O)=O	7006-34-0	asparagine	A5643	NA	-0.8755 log(mol/L)	-0.8755 log(mol/L)	20 °C	NA	NA	NA	NA
CCOC(=O)CN1C=NC2=C1C(=O)N(C)C(= 7029-96-1	7H-Purine-7-acetic acid, 1,2,3,6-tetrahydro-1,3-	A5643	NA	-1.5388 log(mol/L)	-1.5388 log(mol/L)	20 °C	NA	NA	NA	NA
CC(C)C1=CC=CC=C1Br	7073-94-1	BENZENE, 1-BROMO-2-(1-METHYLETHYL)-	A5643	NA	-4.1898 log(mol/L)				NA	NA	NA
CC(OC1=C(C)C=C(Cl)C=C1)C(O)=O	7085-19-0	МСРР	A5643	NA	-2.466 log(mol/L)	Truncated	name		NA	NA	NA
CCCCP(=O)(CCCC)OCC	7100-92-7	Phosphinic acid, dibutyl-, ethyl ester	A5643	NA	-1.2005 log(mol/L)	Tuncated	name	-	NA	NA	NA
NC1=CC=C(C=C1)S(=O)(=O)C1=CC=	C(C7146-68-1	BENZENAMINE, 4-[(4-CHLOROPHENYL)SULFONYL]-	A5643	NA	-4.1267 log(mol/L)	-4.1207 10g(1107 L)	20 0		NA	NA	NA
COC1=NC(Cl)=CC(=C1)C(Cl)(Cl)Cl	7159-34-4	4-Picoline, 2-chloro-6-methoxy-alpha,alpha,alpha	A5643	NA	-4.3635 log(mol/L)	-4.3635 log(mol/L)	20 °C	NA	NA	NA	NA
CCOC1=CC=C(C=C1)N(C(C)=O)C(C)	=0 7174-45-0	Acetamide, N-acetyl-N-(4-ethoxyphenyl)-	A5643	NA	-1.947 log(mol/L)	-1.947 log(mol/L)	25 °C	NA	NA	NA	NA
CCCCOP(=O)(OCC)OCCCC	7242-58-2	Phosphoric acid, dibutyl ethyl ester	A5643	NA	-1.8456 log(mol/L)	-1.8456 log(mol/L)	25 °C	NA	NA	NA	NA
CCCCOP(=O)(OC)OCCCC	7242-59-3	Phosphoric acid, dibutyl methyl ester	A5643	NA	-1.4995 log(mol/L)	-1.4995 log(mol/L)	25 °C	NA	NA	NA	NA
OC(=O)C1=CC=CC2=C1C=CC=N2	7250-53-5	5-Quinolinecarboxylic acid	A5643	NA	-2.6774 log(mol/L)	-2.6774 log(mol/L)	20 °C	NA	NA	NA	NA
OC(CCI)C(O)C(O)C(O)CCI	7251-85-6	D-Mannitol, 1,6-dichloro-1,6-dideoxy-	A5643	NA	-0.6835 log(mol/L)	-0.6835 log(mol/L)	14 °C	NA	NA	NA	NA

SMILES	CASRN 💌	NAME	▼ ARTICLE ▼	PUBMED 💌 V	Vater 💌 UNIT {Water 💌	Water 💌 UNIT {W 💌	Tempe	r UNIT {T ▼ Ionic st ▼	UNIT {I 💌 comm	е 💌 рН	▼ UNIT {p ▼	INCHI_KEY
CC(=O)CC(=O)C(C)(C)C	NA	5,5-dimethyl-2,4-hexadione	A64	11749573	-1.63 log(mol/L)	-1.63 log(mol/L)	NA	- NA	- NA	NA	-	LCLCVVVHIPPHCG-UHFFFAOYSA
CCCC#C	627-19-0	1-pentyne	A64	11749573	-1.64 log(mol/L)	-1.64 log(mol/L)	NA	- NA	- NA	NA	-	IBXNCJKFFQIKKY-UHFFFAOYSA
COC(=O)NS(=O)(=O)C1=CC=C(N)C=C	1 3337-71-1	asulam	4.54	11740570	1.00 (1.00 (N1.0	81.0	- NA	NA	-	VGPYEHKOIGNJKV-UHFFFAOYSA
COC(=0)C1=CC=CC=C1C(=0)OC	131-11-3	dimethyl phthalate	1			11		the second sectors of	- NA	NA	-	NIQCNGHVCWTJSM-UHFFFAOYSA
OC(=O)CNC(=O)C1=CC=CC=C1	495-69-2	hippuric acid	Incons	istent	representa	ition of ste	ereoc	cnemistry	- NA	NA	-	QIAFMBKCNZACKA-UHFFFAOYSA
CCCCC(CC)CN	104-75-6	2-ethylhexylamine							- NA	NA	-	LTHNHFOGQMKPOV-UHFFFAOYSA
CCNC(=O)C(C)OC(=O)NC1=CC=CC=C	1 16118-49-3	carbetamide	A64	11749573	-1.83 log(mol/L)	-1.83 log(mol/L)	NA	- NA	NA	NA	-	AMRQXHFXNZFDCH-UHFFFAOYSA
OCC(0)C10C2OC(0C2C10)C(Cl)(Cl)	CI 15879-93-3	chloralose	1.64	11749573	-1.84 log(mol/L)	-1.84 log(mol/L)	NA	- NA	- 100	NA	-	OJYGBLRPYBAHRT-UHFFFAOYSA
CCCCCCCN	111-68-2	h-prylamine	A64	11749573	-1.85 log(mol/L)	-1.85 log(mol/L)	NA	- NA	- NA	NA	-	WJYIASZWHGOTOU-UHFFFAOYSA
COC(=0)C1=CC=CC=C1	93-58-3	meanyl benzoate	A64	11749573	-1.85 log(mol/L)	-1.85 log(mol/L)	NA	- NA	- NA	NA		QPJVMBTYPHYUOC-UHFFFAOYSA
CC1(C)C2CCC(C)(C2)C1=O	4695-62-9	d-fenchone	A64	11749573	-1.85 log(mol/L)	-1.85 log(mol/L)	NA	- NA	- NA	NA	-	LHXDLQBQYFFVNW-UHFFFAOYSA
CICCCBr	109-70-	1-bromo-3 hloropropane	A64	11749573	-1.85 log(mol/L)	-1.85 log(mol/L)	NA	- NA	- NA	NA	-	STESCIUQSIBMSM-UHFFFAOYSA
C=CC=C	106-99-0	the voladione	A64	11749573	-1.87 log(mol/L)	-1.87 log(mol/L)	NA	- NA	- NA	NA	-	KAKZBPTYRLMSJV-UHFFFAOYSA
O=C1OC2=CC=CC=C2C=C1	91-64-5	Coumarin	A64	11749573	-1.89 log(mol/L)	-1.89 log(mol/L)	NA	- NA	- NA	NA	-	ZYGHJZDHTFUPRJ-UHFFFAOYSA
OC(=O)C1=CC=CC=C1Cl	118-91-2	o-chlorobenzois ecid	A64	11740573	-1.89 log(mol/L)	-1.89 log(mol/L)	NA	- NA	- NA	NX	-	IKCLCGXPQILATA-UHFFFAOYSA
CC(=O)NC1=CC=CC=C1N(=O)=O	552-32-9	2-nitroacetanilide	A64	11749573	-1.91 log(mol/L)	-1.91 log(mol/L)	NA	- NA	- NA	NA	-	BUNFNRVLMKHKIT-UHFFFAOYSA
CICC(CI)CCI	96-18-4			49573	-1.92 log(mol/L)	-1.92				NA	-	CFXQEHVMCRXUSD-UHFFFAOYSA
NC1=CC=CC2=C1C=CC=C2	134-32-7			49573	-1.92 log(mol/L)	-1.92				NA	-	RUFPHBVGCFYCNW-UHFFFAOYSA
CC1=NC2=C(C=CC=C2)C(C)=C1	1198-37-4		CH ₃	49573	-1.94 log(mol/L)	-1.94			'3 I	NA	-	ZTNANFDSJRRZRJ-UHFFFAOYSA
NC1=NC=NC2=C1C=CN2C1OC(CO)C(O 69-33-0			49573	-1.95 log(mol/L)	-1.95	\checkmark		Ŭ	NA	-	HDZZVAMISRMYHH-UHFFFAOYSA
CC1=NC2=C(N1)C=CC=C2	615-15-6			49573	-1.96 log(mol/L)	-1.96		\mathbf{N}		NA	-	LDZYRENCLPUXAX-UHFFFAOYSA
COC1=CC=CC=C1O	90-05-1	1 [Š - ¥-	CH-	49573	-1.96 log(mol/L)	-1.96		Y(NA	-	LHGVFZTZFXWLCP-UHFFFAOYSA
CIC=C(CI)CI	79-01-6			49573	-1.96 log(mol/L)	-1.96	/		3	NA	-	XSTXAVWGXDQKEL-UHFFFAOYSA
FC1=CC=C(F)C=C1	540-36-3			49573	-1.97 log(mol/L)	-1.97				NA	-	QUGUFLIAFISSW-UHFFFAOYSA
CC(C)=CCCC(C)(O)C=C	78-70-6			49573	-1.99 log(mol/L)	-1.99	\	1		NA	-	CDOSHBSSFJOMGT-UHFFFAOYSA
NC(=N)NS(=O)(=O)C1=CC=C(N)C=C1	57-67-0	くう 人		49573	-1.99 log(mol/L)	-1.99				NA	-	BRBKOPJOKNSWSG-UHFFFAOYSA
CC1(C)C2CCC1(C)C(=O)C2	76-22-2			49573	-1.99 log(mol/L)	-1.99	\mathbf{N}	\sim		NA	-	DSSYKIVIOFKYAU-UHFFFAOYSA
CCCCCC(=O)OC	106-70-7		\mathbf{N}	49573	-2 log(mol/L)	-2	Ý			NA	-	NUKZAGXMHTUAFE-UHFFFAOYSA
CC(C)CCI	513-36-0		U	49573	-2 log(mol/L)	-2			·	NA	-	QTBFPMKWQKYFLR-UHFFFAOYSA
FC1=CC(F)=CC=C1	372-18-9			49573	-2 log(mol/L)	-2	1			NA	-	UEMGWPRHOOEKTA-UHFFFAOYSA
CC(=C)C=C	78-79-5			49573	-2.03 log(mol/L)	-2.03	1			NA	-	RRHGJUQNOFWUDK-UHFFFAOYSA
COC1=C(O)C=CC(=C1)C(O)=O	121-34-6	H ₃ C		49573	-2.05 log(mol/L)	-2.05	$_{2}C$			NA	-	WKOLLVMJNQIZCI-UHFFFAOYSA
CC(C)=CCC(C)=CC=O	5392-40-5	1 130		49573	-2.06 log(mol/L)	-2.06	3			NA	-	WTEVQBCEXWBHNA-UHFFFAOYSA
C=CCC1(CC=C)C(=O)NC(=O)NC1=O	52-43-7	Ŭ		49573	-2.06 log(mol/L)	-2.06,				NA	-	FDQGNLOWMMVRQL-UHFFFAOYS4
CC1=CC=CC2=C1N1C=NN=C1S2	41814-78-2	tricyclazole	A64	11749573	-2.07 log(mol/L)	-2.07 log(mol/L)		- NA	- NA	NA	-	DQJCHOQLCLEDLL-UHFFFAOYSA
BrCCCBr	109-64-8	1,3-dibromopropane	A64	11749573	-2.08 log(mol/L)	-2.08 log(mol/L)	NA	- NA	- NA	NA	-	VEFLKXRACNJHOV-UHFFFAOYSA
CNC(=0)ON=C1C(Cl)C2CC(C#N)C1C2	2 15271-41-7	tranid	A64	11749573	-2.08 log(mol/L)	-2.08 log(mol/L)	NA	- NA	- NA	NA	-	QCQPGRMMDFIQMB-UHFFFAOYSA

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.

		NAME			ater 💌 UNIT {Water 💌 \					{I(_ comme		UNIT {p	INCHI_KEY	
0000	71-36-3	1-Butanol	A171	10850781	0 log(mol/L)	0 log(mol/l		- NA	-	NA	NA	-	LRHPLDYGYMQRH	N-UHFFFAOYSA
DC1=CC=CC=C1	108-95-2	Phenol	A171	10850781	0 log(mol/L)	0 log(mol/l		- NA	-	NA	NA	-	ISWSIDIOOBJBQZ-	UHFFFAOYSA
DC1=CC=CC=C1	NA	Phenol	A5370	19226181	0 log(mol/L)	0 log(mol/l	L) NA	- NA	-	NA	NA	-	ISWSIDIOOBJBQZ-	UHFFFAOYSA
0000	NA	1-Butanol	A5370	19226181	0 log(mol/L)	0 log(mol/l	L) NA	- NA	-	NA	NA	-	LRHPLDYGYMQRH	N-UHFFFAOYSA
C\C=C\C(O)=O	3724-65-0	crotonic acid	A5643	NA	0 log(mol/L)	0 log(mol/l	L) NA	°C NA	-	NA	NA	-	LDHQCZJRKDOVO	X-NSCUHMNNS/
D[C@H]([C@@H](O)C(O)=O)C(O)=O	. 51-42-3	NA	A5643	NA	0 log(mol/L)	0 log(mol/l	L) NA	- NA	-	NA	NA	-	YLXIPWWIOISBDD	-KZKMUFAMSA
0000	NA	1-Butanol	A80390	NA	0 log(mol/L)	0 log(mol/l	L) NA	- NA	NA	NA	NA	-	LRHPLDYGYMQRH	N-UHFFFAOYSA
CN[Pt](Cl)(Cl)NCC	NA	NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA	NA	NA	NA	-	USSSQXLMZPPDSJ	-UHFFFAOYSA
CCN[Pt](O)(O)(O)(O)NCC	NA	NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA						IHFFFAOYSA
D[Pt]1(O)(O)(O)NC2CCCC2N1	NA	NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA	1					P-UHFFFAOYSA
N[Pt]1(N)(OC(=O)CCC(O)=O)(OC(=O	NA	NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA		In the lite	, data	UHFFFAOYSA		
COC(=O)CCC(=O)O[Pt]1(N)(N)(OC(=	0 NA	NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA		Invali	data	UHFFFAOYSA		
COC(=O)		<u></u>	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA	in		-			UHFFFAOYSA
COC(=O)(A103509	NA	0 #NAME?	0 log(mol/l	L) NA 🦯	- NA	111	conec	uy c	onvei	ted to a	UHFFFAOYSA
Leading ze	ros or	n	A103509	NA	0 #NAME?	0 log(mol/l	L) N"	- NA		M	bile	forma	t	-UHFFFAOYSA
IPtI1(N			A103509	NA	0 #NAME?	0 log(mol/l		- NA	1	V	anu	TOTTIC		C-UHFFFAOYS
coc(=o CAS nun	nbers		A103509	NA	0 #NAME?	0 log(mol/l		- NA	1					JHFFFAOYSA
COC(=0			A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	CGTXEOBYNSQVL	-UHFFFAOYSA
COC(=0,			A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	HVFPKBDKSKFHG	I-UHFFFAOYSA
N[Pt]1(N)(OC(=O)CCC(=O)NC2CCCC		NA	A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	CXGWWVPBZJUN	N-UHFFFAOYS
N[Pt]123(N)OC(=O)CC(=O)O1.O=C(C		NA	A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	KIAMCWXPDFXUN	W-UHFFFAOY
D=C1CC(=O)O[Pt]23(NCCN2)(O1)OC		NA	A103509	NA	0 #NAME?	0 log(mol/l	L) NA	- NA	NA	NA	NA	-	DLQXGXILLTTYBT-	JHFFFAOYSA
D=C10[Pt]23(NCCN2)(OC(=O)C11CC		NA	A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	DPPBVBJQAOMG	A-UHFFFAOYS
N[Pt]123(NC4CCCCC4)OC(=0)CC(=0)		NA	A103509	NA	0 #NAME?	0 log(mol/l		- NA	NA	NA	NA	-	OFPACIQPTZBGQ	-UHFFFAOYSA
C\C=C\C(O)=O	3724-65 0	crotonic acid	A108291	NA	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	-	LDHQCZJRKDOVO	x-NSCUHMNNS
CNCC(O)C1=CC=C(O)C(O)=C1	51-42-3	ADRENALINE TARTRATE (1:1)	A111278	NA	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	-	UCTWMZQNUQW	SLP-UHFFFAOY
DC1=CC=CC=C1	NA	Phenol		NA	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	-	ISWSIDIOOBJBQZ-	
0000	NA 🚽	1-Butanol		NA	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	-	LRHPLDYGYMQRH	
NA	000115-10-6	NA	A111695	27463195	3 log(mmol/L)			NA	NIA	NA	NA	NA	98dd04fe45ae608	
NA	000536-75-4		A111695	27463195	3 log(mmol/L)	0 1				A	NA	N	98dd04fe45ae608	
CCCO	NA	NA	A112073	NA	0 log(mol/L)	0.1				IA	NA.	NA	LRHPLDYGYMQRH	
IC(CCCNC(N)=N)C(O)=O	NA	NA	A112076	24456022	0 log(mol/L)	0 1	nvalio	d InChiKe	evs	-4	NA	NA	ODKSFYDXXFIFQN	
CC(C)(O)C#C	NA	NA	A112076	24456022	0 log(mol/L)	0 1			-,-	A	NA	NA	QXLPXWSKPNOQL	
	71-36-3	NA	A112473	12146628	0 log(mol/L)	0 1				A	NA	NA	LRHPLDYGYMQRH	
	NA	NA	A112523	26457702	0 log(mol/L)	0 log(mol/l) NA	- NA	NA	NA	NA		LRHPLDYGYMQRH	
V[C@H](CCCn:[c+](:[n]):[n])C([O-])=		NA	A112523	26457702	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	NA	63b8601e5597f6d0	
C1=CC=CC=C1	NA	NA	A112523	26457702	0 log(mol/L)	0 log(mol/l		- NA	NA	NA	NA	NA	ISWSIDIOOBJBQZ-	

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



Copy 🔻 Share 👻 Submit Comment 🔍 🔍 Search all data

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.

DETAILS		Property Water Solubility							
EXECUTIVE SUMMARY		- Water Soldbirdy		Water S	alubility				
PROPERTIES				Water So	Jubility				
ENV. FATE/TRANSPORT		Lownload Summary 🔻							
HAZARD		Туре	Average	\$ Median	₽ Rang	ge		♥ Unit	
ADME		Experimental	8.55e-4	5.26e-4	5.25	e-4 to 1.51e-3		mol/L	
		Predicted	8.78e-4	7.56e-4	5.35	e-4 to 1.31e-3		mol/L	
EXPOSURE									
BIOACTIVITY									
SIMILAR COMPOUNDS				Experir	nental				
GENRA (BETA)		🕹 Download Experimental Data 🔻							
RELATED SUBSTANCES	-					Result	Experimental Details		
SYNONYMS		r V., et al. "Estimation of aqueous ate indices." . J. Chem. Inf. and Co				5.26e-4			
LITERATURE		Tetko et al. J. Chem. Inf. and Comp.	Sci. 41.6 (2001): 1488-1493			1.51e-3			
		Kovdienko, et. al. Molecular informa	atics 29.5 (2010): 394-406.			5.25e-4			
LINKS									
COMMENTS									

- 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

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.

www.epa.gov/research

Innovative Research for a Sustainable Future