EPA
United States
Environmental Protection
Agency

Benchmarking novel curation strategies for large publicly available water solubility compilations

Gabriel Sinclair, **Charles Lowe**, Nathaniel Charest, Christian Ramsland, Todd Martin, Ann Richard, and Antony Williams

Center for Computational Toxicology and Exposure, U.S. EPA, Research Triangle Park, NC

United States Environmental Protection Agency

Disclaimer: The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Motivation

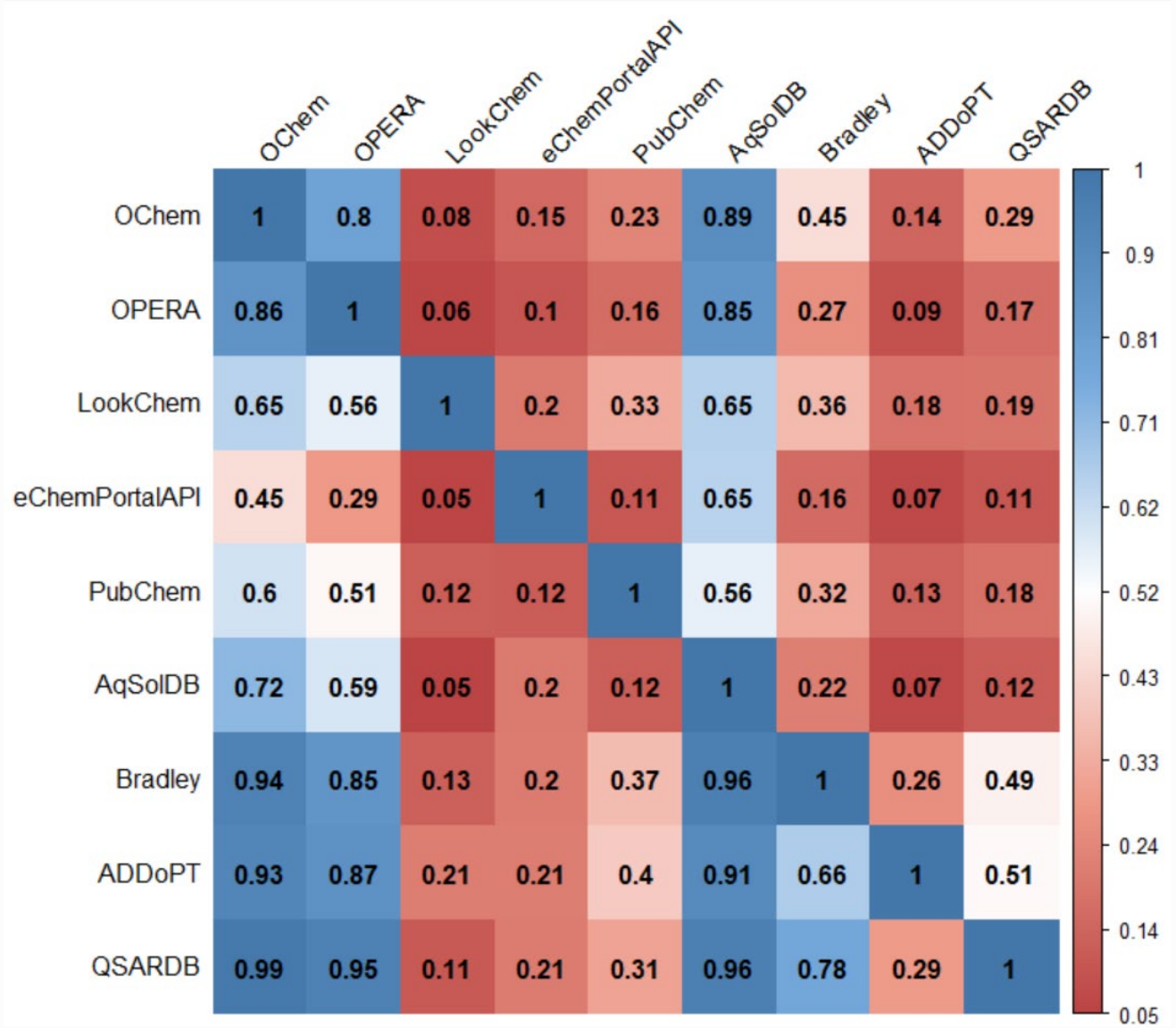
- Accurate measurement of the solubility of chemicals in water is important!
  - Pharmaceuticals
  - Toxicology and Exposure
  - Synthesis
  - Instrumental analysis
- Numerous models already exist in the literature or are available within commercial tools or open-source tools
  - OPERA, TEST, ACD/Percepta, etc.
- Goals
  - Provide a larger, more diverse dataset
  - Ensure curation procedure leads to better modelability
  - Compare results of curation against another curated dataset

https://www.sciencephoto.com/media/1062490/view/glass-of-water-and-effervescent-tablet-illustration

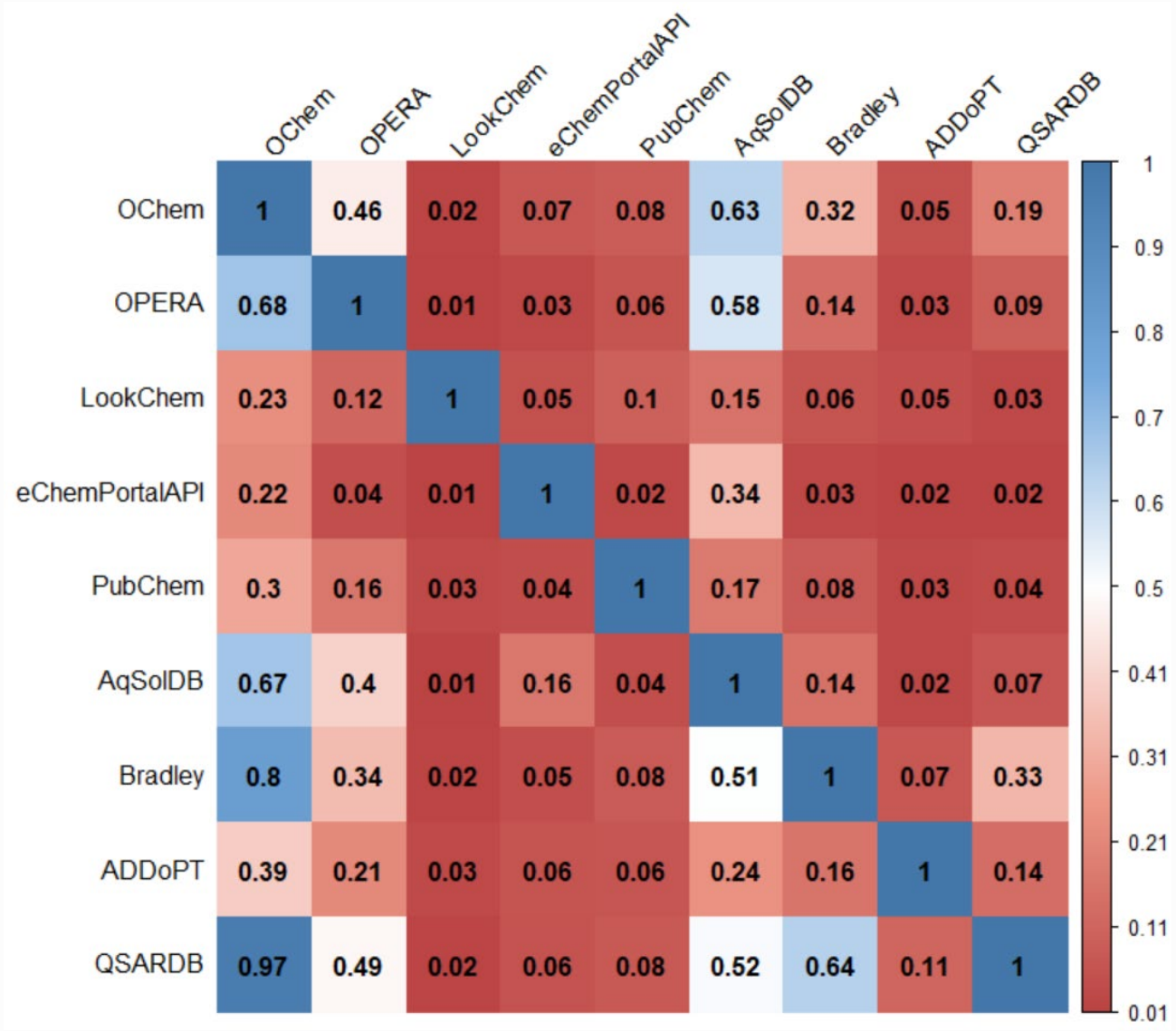
Records in dataset before and after curation

Source Abbreviation	Short Description	Original No. Records	Curated No. Records
ADDoPT	Advanced Digital Design of Pharmaceutical Therapeutics	1484	761
AqSolDB	Aqueous Solubility Database	9959	7408
Bradley	Dataset curated by Bradley et al.	3948	2493
eChemPortalAPI	OECD chemical substance database	8040	3752
LookChem	Chemical trading platform	6035	532
OChem	Online chemical modeling environment	28,683	16,864
OPERA	PhysProp dataset curated for OPERA	5267	5084
PubChem	NLM's chemical database	10,800	2031
QsarDB	FAIR repository of (Q)SAR/QSPR models	1103	928
Total Records:		75,319	39,853

Redundancy by QSAR-ready Structure



Redundancy by QSAR-ready structure & log(M)



Curation criteria

Bins for matching identifiers within DSSTox
Ambiguous Synonym matched SOURCE_CHEMICAL_NAME
CAS-RN matched other record: SOURCE_CASRN
CAS-RN matched SOURCE_CASRN
Conflict rejected: Agreement does not surpass threshold 1.0
Conflict rejected: No authoritative match found
Conflict resolution failed: QSAR-ready SMILES or standardization unavailable for best record
Conflict resolution failed: The indicated alternative records were not found in DSSTox
DTXSID matched other record: SOURCE_DTXSID
Mapped Identifier matched SOURCE_CHEMICAL_NAME

Curation criteria

Bins for matching identifiers within DSSTox
Name2Structure matched SOURCE_CHEMICAL_NAME
OPSIN ambiguous name
Other CAS-RN matched other record: SOURCE_CASRN
Preferred Name matched SOURCE_CHEMICAL_NAME
Structure matched SOURCE_SMILES
Unique Synonym matched SOURCE_CHEMICAL_NAME
UVCB keywords in name
Valid Synonym matched SOURCE_CHEMICAL_NAME

Curation criteria

Reason for discarding record based on measurement
No valid DSSTox record found
No numerical data
Unrealistic value for property
Unit conversion failed
Range width outside tolerance

Discarded records

Reason for discarding record based on identifiers	COUNT
Structure matched SOURCE_SMILES	3068
CAS-RN matched SOURCE_CASRN	2711
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME	2436
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME	1448
Structure matched SOURCE_SMILES, CAS-RN matched other record: SOURCE_CASRN	950
Conflict rejected: No authoritative match found	560
Conflict rejected: Agreement does not surpass threshold 1.0	352
Conflict resolution failed: The indicated alternative records were not found in DSSTox	317
Structure matched SOURCE_SMILES, CAS-RN matched SOURCE_CASRN	312
OPSIN ambiguous name	265
Mapped Identifier matched SOURCE_CHEMICAL_NAME	249
Name2Structure matched SOURCE_CHEMICAL_NAME	104
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	88
Conflict resolution failed: QSAR-ready SMILES or standardization unavailable for best record	43
Preferred Name matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	38
Structure matched SOURCE_SMILES, Preferred Name matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	24
UVCB keywords in name	16
Structure matched SOURCE_SMILES, Ambiguous Synonym matched SOURCE_CHEMICAL_NAME	16
Structure matched SOURCE_SMILES, Other CAS-RN matched other record: SOURCE_CASRN	12
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME, DTXSID matched other record: SOURCE_DTXSID	10
Structure matched SOURCE_SMILES, Valid Synonym matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	8
Structure matched SOURCE_SMILES, Preferred Name matched SOURCE_CHEMICAL_NAME	6
Preferred Name matched SOURCE_CHEMICAL_NAME	6
Structure matched SOURCE_SMILES, DTXSID matched other record: SOURCE_DTXSID	5
Ambiguous Synonym matched SOURCE_CHEMICAL_NAME	3
Structure matched SOURCE_SMILES, Unique Synonym matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1
Structure matched SOURCE_SMILES, Other CAS-RN matched SOURCE_CASRN	1
Valid Synonym matched SOURCE_CHEMICAL_NAME	1
Unique Synonym matched SOURCE_CHEMICAL_NAME	1
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, DTXSID matched other record: SOURCE_DTXSID	1
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1

Discarded records

Reason for discarding record based on identifiers	COUNT
Structure matched SOURCE_SMILES	3068
CAS-RN matched SOURCE_CASRN	2711
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME	2436
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME	1448
Structure matched SOURCE_SMILES, CAS-RN matched other record: SOURCE_CASRN	950
Conflict rejected: No authoritative match found	560
Conflict rejected: Agreement does not surpass threshold 1.0	352
Conflict resolution failed: The indicated alternative records were not found in DSSTox	317
Structure matched SOURCE_SMILES, CAS-RN matched SOURCE_CASRN	312
OPSIN ambiguous name	265
Mapped Identifier matched SOURCE_CHEMICAL_NAME	249
Name2Structure matched SOURCE_CHEMICAL_NAME	104
Structure matched SOURCE_SMILES, Unique Synonym matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1
Structure matched SOURCE_SMILES, Other CAS-RN matched SOURCE_CASRN	1
Valid Synonym matched SOURCE_CHEMICAL_NAME	1
Unique Synonym matched SOURCE_CHEMICAL_NAME	1
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, DTXSID matched other record: SOURCE_DTXSID	1
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1

Reason for discarding record based on identifiers

SOURCE_CASRN	SOURCE_CHEMICAL_NAME	SOURCE_SMILES	REASON
319-86-8		ClC1C(Cl)C(Cl)C(Cl)(	Structure matched SOURCE_SMILES, CAS-RN matched other recor
201743-52-4	1,3-Dioxane,4-(bromomethyl)-2		CAS-RN matched SOURCE_CASRN
319-85-7		ClC1C(Cl)C(Cl)C(Cl)(	Structure matched SOURCE_SMILES, CAS-RN matched other recor
496-46-8	octahydroimidazo[4,5-d]imidazo		CAS-RN matched SOURCE_CASRN
180516-87-4	Benzoic acid,4-(4,4,5,5-tetrame		CAS-RN matched SOURCE_CASRN
65277-42-1	1-[4-[4-[[(2R)-2-(2,4-dichloroph	CC(=O)N1CCN(CC1	Structure matched SOURCE_SMILES, CAS-RN matched other recor
185996-33-2	1,3-Dioxolane-4-acetamide,2,2-		CAS-RN matched SOURCE_CASRN
186537-58-6	L-Cystine,N,N'-bis(phenoxycarb		CAS-RN matched SOURCE_CASRN
175463-32-8	1-Boc-3-cyano-4-oxopyrrolidine		CAS-RN matched SOURCE_CASRN
2190	BENZAMIDE, 2-IODO-N-PHEN	IC1=CC=CC=C1C(=(	Structure matched SOURCE_SMILES
2127	1-chloro-2-methylpropane	CC(C)CCl	Structure matched SOURCE_SMILES, Mapped Identifier matched SO
7774-96-1	[2-methoxy-4-[(E)-prop-1-enyl]p	CC=CC1=CC(=C(C=	Structure matched SOURCE_SMILES, CAS-RN matched SOURCE_
0200	1,2-dibromopropane	CC(Br)CBr	Structure matched SOURCE_SMILES, Mapped Identifier matched SO

Discarded records

Reason for discarding record based on indentifers	COUNT
Structure matched SOURCE_SMILES	3068
CAS-RN matched SOURCE_CASRN	2711
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME	2436
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME	1448
Structure matched SOURCE_SMILES, CAS-RN matched other record: SOURCE_CASRN	950
Conflict rejected: No authoritative match found	560
Conflict reject	352
Conflict resolu	317
Structure mat	312
OPSIN ambig	265
Mapped Iden	249
Name2Struct	104
Structure mat	88
Conflict resol	43
Preferred Nar	38
Structure mat	24
UVCB keywo	16
Structure mat	16
Structure mat	12
Structure mat	10
Structure matched SOURCE_SMILES, Valid Synonym matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	8
Structure matched SOURCE_SMILES, Preferred Name matched SOURCE_CHEMICAL_NAME	6
Preferred Name matched SOURCE_CHEMICAL_NAME	6
Structure matched SOURCE_SMILES, DTXSID matched other record: SOURCE_DTXSID	5
Ambiguous Synonym matched SOURCE_CHEMICAL_NAME	3
Structure matched SOURCE_SMILES, Unique Synonym matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, CAS-RN matched other record: SOURCE_CASRN	2
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1
Structure matched SOURCE_SMILES, Other CAS-RN matched SOURCE_CASRN	1
Valid Synonym matched SOURCE_CHEMICAL_NAME	1
Unique Synonym matched SOURCE_CHEMICAL_NAME	1
Structure matched SOURCE_SMILES, Mapped Identifier matched SOURCE_CHEMICAL_NAME, DTXSID matched other record: SOURCE_DTXSID	1
Structure matched SOURCE_SMILES, Name2Structure matched SOURCE_CHEMICAL_NAME, Other CAS-RN matched other record: SOURCE_CASRN	1

Reason for discarding record based on measurement	COUNT
No valid DSSTox record found	19098
No numerical data	6277
Unrealistic value for property	705
Unit conversion failed	501
Range width outside tolerance	42

“No numerical data”

EXP_PROP_ID	SRC_CHEM_ID	PROPERTY_VALUE	PARAMETER_VALUES	REASON
EXP000001103062	SCH000000595291	insoluble		No numerical data
EXP000001103055	SCH000000595284	insoluble		No numerical data
EXP000001103035	SCH000000595263	insoluble		No numerical data
EXP000001103041	SCH000000595269	slightly soluble		No numerical data
EXP000001103026	SCH000000595254	soluble		No numerical data
EXP000001103014	SCH000000595242	insoluble		No numerical data
EXP000001103008	SCH000000595236	insoluble		No numerical data

“Unrealistic value for property”

EXP_PROP_ID	SRC_CHEM_ID	PROPERTY_VALUE	PARAMETER_VALUES	REASON
EXP000000981828	SCH000000149283	1000		Unrealistic value for property
EXP000000962955	SCH000000150276	1020		Unrealistic value for property
EXP000000981606	SCH000000282118	1000		Unrealistic value for property
EXP000000976688	SCH000000571298	1000		Unrealistic value for property
EXP000000973446	SCH000000571297	999		Unrealistic value for property
EXP000000973321	SCH000000571263	999		Unrealistic value for property
EXP000000973589	SCH000000571328	1000		Unrealistic value for property
EXP000000973625	SCH000000571334	1000		Unrealistic value for property

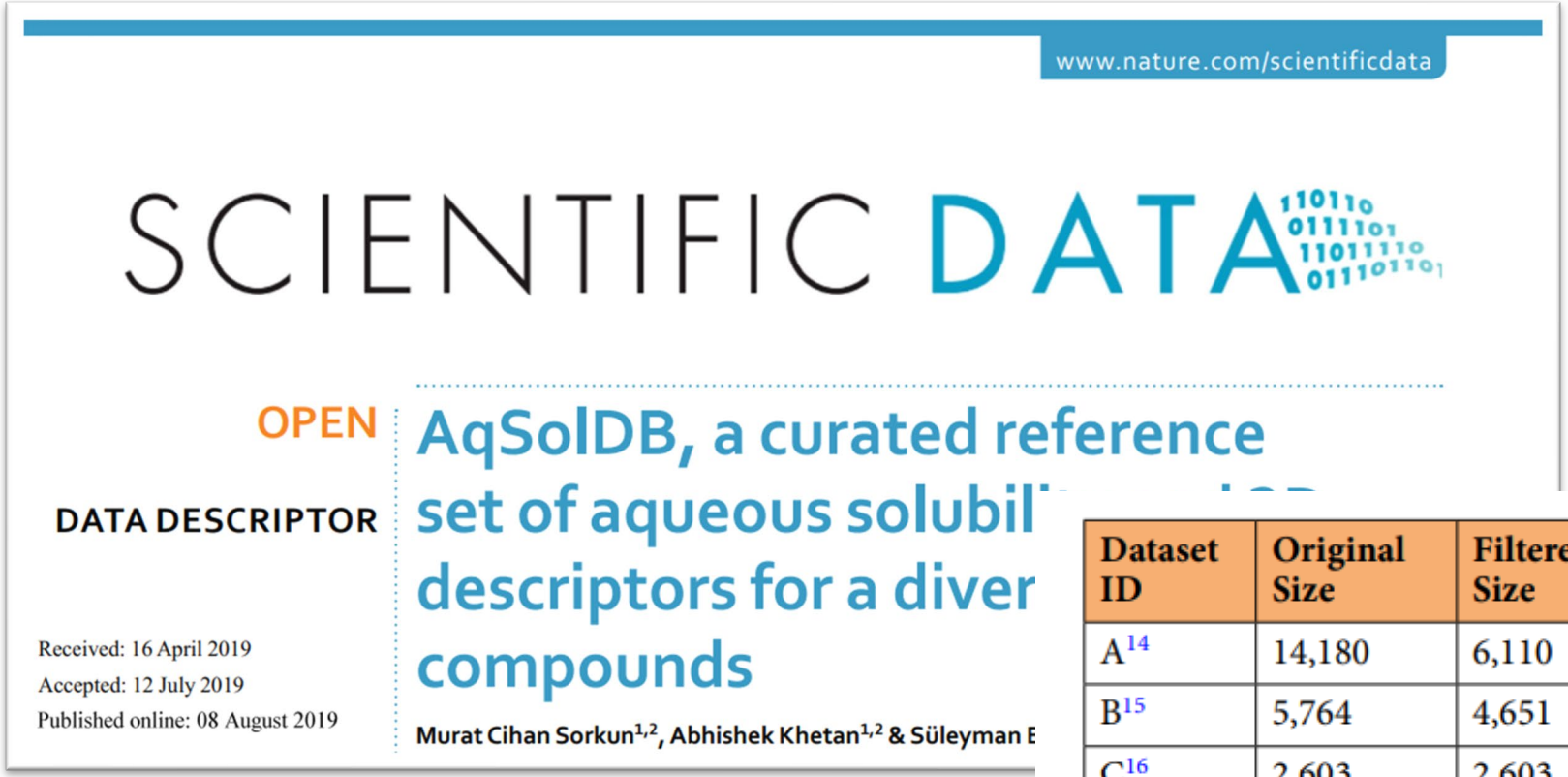
“Unit conversion failed”

EXP_PROP_ID	SRC_CHEM_ID	PROPERTY_VALUE	PARAMETER_VALUES	REASON
EXP000000454026	SCH000000140604	2.2	Reliability: 2 (reliable with restrictions); Temperature: 22.0	Unit conversion failed
EXP000000452904	SCH000000140548	2.83	Reliability: 2 (reliable with restrictions); Temperature: 19.6	Unit conversion failed
EXP000000458498	SCH000000139533	~1623.0	Reliability: 2 (reliable with restrictions); pH: ~7.0; Temperature: 25.0	Unit conversion failed
EXP000000458586	SCH000000139557	~700.0	Reliability: 2 (reliable with restrictions)	Unit conversion failed

“Range width outside tolerance”

EXP_PROP_ID	SRC_CHEM_ID	PROPERTY_VALUE	PARAMETER_VALUES	REASON
EXP000000454011	SCH000000140846	0.0-100.0	Reliability: 2 (reliable with restrictions); Temperature: 20.0	Range width outside tolerance
EXP000000454011	SCH000000140846	0.0-100.0	Reliability: 2 (reliable with restrictions); Temperature: 20.0	Range width outside tolerance
EXP000000454022	SCH000000140847	0.0-15.0	Reliability: 1 (reliable without restriction); pH: 0.0-2.5; Temperature: 20.0	Range width outside tolerance
EXP000000454022	SCH000000140847	0.0-15.0	Reliability: 1 (reliable without restriction); pH: 0.0-2.5; Temperature: 20.0	Range width outside tolerance
EXP000000454022	SCH000000140847	0.0-15.0	Reliability: 1 (reliable without restriction); pH: 0.0-2.5; Temperature: 20.0	Range width outside tolerance
EXP000000454324	SCH000000145151	9.999999999999999E-6-5.0E-4	Reliability: 1 (reliable without restriction); pH: 6.3; Temperature: 20.0	Range width outside tolerance
EXP000000454324	SCH000000145151	9.999999999999999E-6-5.0E-4	Reliability: 1 (reliable without restriction); pH: 6.3; Temperature: 20.0	Range width outside tolerance

www.nature.com/scientificdata



Dataset ID	Original Size	Filtered Size	Compound Representations	Solubility Units
A[14]	14,180	6,110	name, CAS	g/L, mg/L, μg/L
B[15]	5,764	4,651	name, CAS	LogS
C[16]	2,603	2,603	name, SMILES	LogS
D[17]	2,267	2,115	name, CAS	LogS
E[1]	1,291	1,291	name, SMILES, CAS	LogS
F[8]	1,210	1,210	SLN	LogS
G[2]	1,144	1,144	name, SMILES	LogS
H[8]	578	578	SLN	LogS
I[20]	105	94	name, SMILES, InChI	μM

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds

Received: 16 April 2019
Accepted: 12 July 2019
Published online: 08 August 2019

Murat Cihan Sork

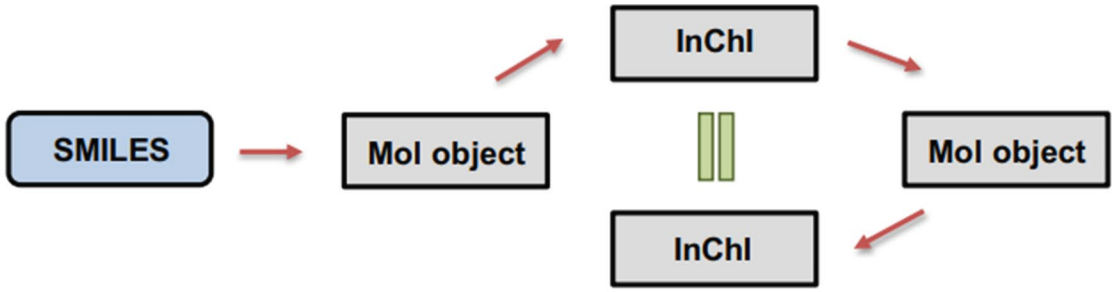


Fig. 2 Validation steps of compound representations. Blue box represents the SMILES values from the dataset and gray boxes represent the generated values using RDKit. Red arrows represent the conversion steps and green equal sign represents the validation of consistency.

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

AqSolDB, a curate… set of aqueous solu… descriptors for a di… compounds

Received: 16 April 2019
Accepted: 12 July 2019
Published online: 08 August 2019

Murat Cihan Sorkun[1,2], Abhishek Khetan[1,2] & Sü…

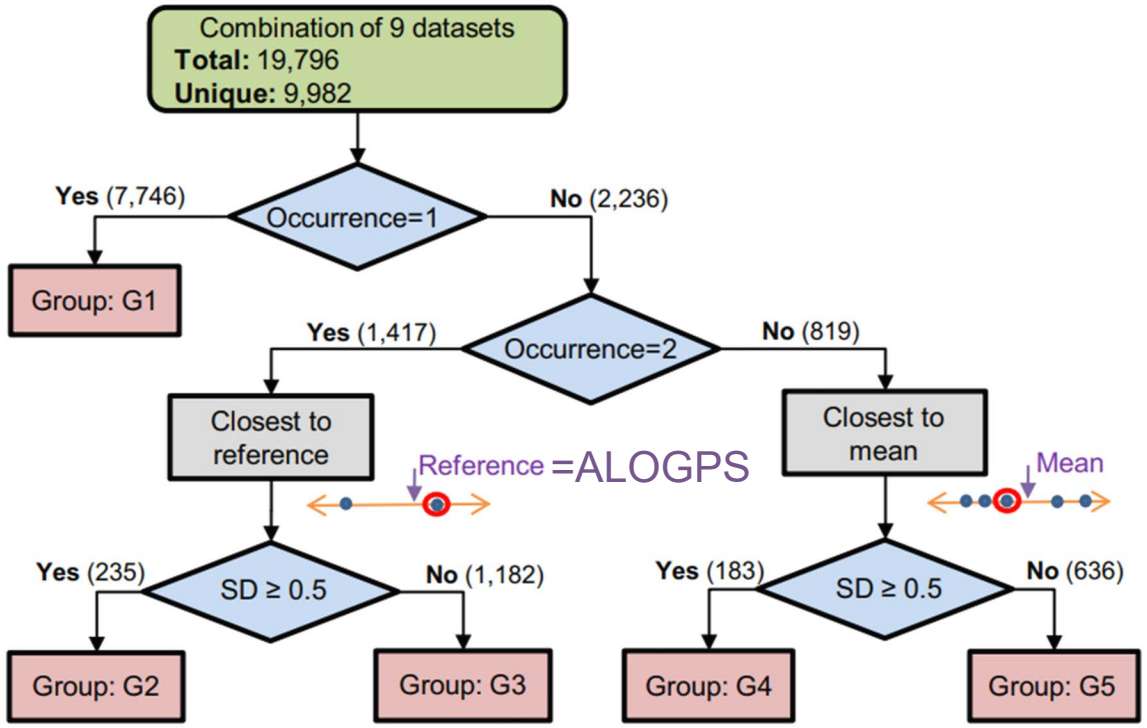


Fig. 4 Flowchart of the curation algorithm. Green box represents the initial state. Blue diamond shapes represent a decision according to the number of occurrences of a compound and the SD of multiple occurrences. Pink boxes represent the reliability group. Gray boxes represent the selection method for multiple occurrences. The numbers over the arrows represent the number of unique compound in the corresponding classification path.

Dataset preparation & creation

- Calculate median water solubility for each QSAR-ready SMILES
- Randomly split data into training/test sets (75%/25%)
- 1,444 1D & 2D PaDEL descriptors from QSAR-ready SMILES.
- Eliminate constant, near constant, and highly correlated descriptors
- Center and scale descriptors
- Random forest
  - An ensemble of decision trees made with a random subset of descriptors (bagging) to overcome overfitting
  - Algorithmically selected/expert selected descriptor set

Dataset preparation & creation

Standardized SMILES	median_WS	nAcid	apol	naAromAtom	nAromBor	nAtom	nHeavyAt	nH	nB	nC	nN
[O-][N+](=O)C(C=O)(C(O)=O)C1C=CC=CC=1	-1.84	1	25.6	6	6	22	15	7	0	9	1
[O-][N+](=O)C1(Br)COCOC1	-1.65	0	18.4	0	0	16	10	6	0	4	1
[O-][N+](=O)C1=C(Cl)C(=C(O)C(=C1Cl)[N+]([O-])=O)[N+]([O-])=O	-0.60	0	24.5	6	6	19	18	1	0	6	3
[O-][N+](=O)C1=CC(=C(C=C1)N=NC1=C(O)C=CC2=CC=CC=C21)[N+]([O-])=O	-7.73	0	43.2	16	17	35	25	10	0	16	4
[O-][N+](=O)C1=CC(=C(C=C1)ON=CC1=CC(Br)=C(O)C(Br)=C1)[N+]([O-])=O	-6.66	0	41.8	12	12	31	24	7	0	13	3
[O-][N+](=O)C1=CC(=CC(=C1Cl)[N+]([O-])=O)C(O)=O	-3.52	1	23.5	6	6	19	16	3	0	7	2
[O-][N+](=O)C1=CC(=CC=C1)C(O)=O	-1.68	1	20	6	6	17	12	5	0	7	1
[O-][N+](=O)C1=CC(=CC=C1F)[N+]([O-])=O	-2.67	0	18.5	6	6	16	13	3	0	6	2
[O-][N+](=O)C1=CC(=CC=C1N=NCC=CC1=CC=C(O1)[N+]([O-])=O)[N+]([O-])=C	-5.06	0	40	11	11	34	25	9	0	13	5
[O-][N+](=O)C1=CC(Cl)=C(Cl)C(Cl)=C1Cl	-4.55	0	22.7	6	6	14	13	1	0	6	1
[O-][N+](=O)C1=CC(Cl)=C(Cl)C=C1Cl	-3.89	0	21.1	6	6	14	12	2	0	6	1
[O-][N+](=O)C1=CC=C(C=C1C(=O)OCC(O)=O)OC1=CC=C(C=C1Cl)C(F)(F)F	-5.84	1	44.7	12	12	37	28	9	0	16	1
[O-][N+](=O)C1=CC=C(C=CC=NN2CC(=O)NC2=S)O1	-3.96	0	33.4	5	5	27	19	8	0	10	4
[O-][N+](=O)C1=CC=C(CN=O)O1	-2.19	0	16.9	5	5	15	11	4	0	5	2
[O-][N+](=O)C1=CC=C(O)C2N=CC=CC1=2	-1.84	0	25.8	0	0	22	14	8	0	9	2
[O-][N+](=O)C1=CC=C(O1)C1NC2=CC=CC=C2N=1	-3.89	0	29.7	14	16	24	17	7	0	11	3
[O-][N+](=O)C1=CC=C(O1)C1NC2C=C(F)C=CC=2N=1	-4.00	0	29.6	14	16	24	18	6	0	11	3
[O-][N+](=O)C1=CC=C2CCC3C=CC=C1C=32	-5.34	0	29.8	10	11	24	15	9	0	12	1
[O-][N+](=O)C1=CC=CC=C1C(O)=O	-1.35	1	20	6	6	17	12	5	0	7	1
[O-][N+](=O)C1=CC=CC2C=CC=C(C(O)=O)C=21	-2.75	1	28.3	10	11	23	16	7	0	11	1
[O-][N+](=O)C1=CC2=CC=CN=C2C2=NC=CC=C12	-3.92	0	30.7	14	16	24	17	7	0	12	3

Random forest results

Results for OPERA & AqSolDB's curation and/or modeling strategy

Dataset	No. Descriptors	R^2 5CV	RMSE 5CV	Training Size	R^2 Training	RMSE Training	Test Size	R^2 Test	RMSE Test
AqSolDB	17	0.77	1.13	7469	0.95	0.56	2490	0.78	1.10
OPERA	11	0.87	0.81	3158	0.87	0.82	1066	0.86	0.86

Results for this work's curation & modeling strategy

Dataset	No. Descriptors	R^2 5CV	RMSE 5CV	Training Size	R^2 Training	RMSE Training	Test Size	R^2 Test	RMSE Test
Entire Dataset	16	0.82	0.96	8037	0.97	0.41	2680	0.82	0.97
AqSolDB	16	0.86	0.88	5424	0.98	0.38	1809	0.86	0.90
OPERA	15	0.86	0.89	3812	0.98	0.38	1271	0.87	0.86

Final thoughts

- This curation and modeling workflow established here will be transferred to other physicochemical endpoints of interest to EPA
- A manuscript describing this work and the work of our earlier presentation on modeling water solubility is currently in late development
- These data and models will be made available within a future EPA cheminformatics module
  - See https://hcd.rtpnc.epa.gov/#/ for currently available modules such as the standardizer


Thank you for
Listening!