JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

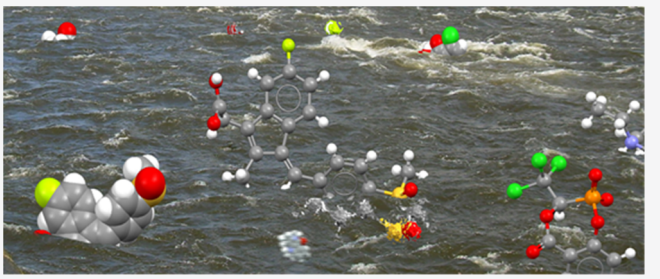
Findings of the Second Challenge to Predict Aqueous Solubility

Antonio Llinas,* Ioana Oprisiu, and Alex Avdeef

Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00701 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Ten years ago, we issued an open prediction challenge to the cheminformatics community: would participants be able to predict the equilibrium intrinsic solubilities of 32 druglike molecules using only a high-precision (CheqSol instrument, performed in one laboratory) set of 100 compounds as a training set? The "solubility challenge" was a widely recognized success and spurred many discussions about the prediction methods and quality of data. We revisited the competition a second time recently and challenged the community to a different challenge, not a blind test this time but using a larger test set of molecules, gathered and curated from published sources (mostly "gold standard" saturation shake-flask measurements), where the average interlaboratory reproducibility for the molecules was estimated to be ∼0.17 log unit. Also, a second test set was included, comprising "contentious" molecules, the reported (mostly shake-flask) solubility of which had higher average uncertainty, ∼0.62 log unit. In the second competition, the participants were invited to use their own training sets, provided that the training sets did not contain any of the test set molecules. We were motivated to revisit the competition to (1) examine to what extent computational methods had improved in 10 years, (2) verify that data quality may not be the main limiting factor in the accuracy of the prediction method, and (3) attempt to seek a relationship between the makeup of the training set data and the prediction outcome.



INTRODUCTION

Recently,[1] we initiated a repeat of the "solubility challenge" originally carried out 10 years ago,[2,3] to predict equilibrium intrinsic aqueous solubility (at 25 °C, in logarithmic molarity units, as log S_0) from the structure. In the original competition (SC-1), precise S_0 values of drug substances were measured by the same group, using the same CheqSol potentiometric method.[4,5] Participants in SC-1 were invited to predict the S_0 values of a 32-drug test set, using a provided training set of 100 S_0 values. Before SC-1, it was a widely held view that the lack of enough good experimental data had held back both the understanding of the equilibrium processes and the derivation of effective prediction models. Using precision data from a single source was intended to minimize the influence of interlaboratory variability in measurements on the prediction outcome, as there had been a persistent concern that the interlaboratory reproducibility in published solubility values for druglike molecules appeared to be 0.6 log unit or higher.[6−9]

It is important to emphasize at this point that neither the first solubility challenge (SC-1) nor this second one (SC-2) aimed to identify a "winner." It was our goal to advance our general understanding of predictive models and to examine the actual state of the art in prediction methods.

The findings of the SC-1 competition indicated that computational methods did not predict log S_0 any better than a root-mean-square-error (RMSE) of ∼0.6 log unit. In only two entries (of 99), the predictions were better than 60% "correct", defined as being less than ±0.5 log unit from the measured value, or a ±10% error for S. Five of the top-rank predictions used partial least squares (PLS) and support vector regression (SVR) methods, based on recursively optimized atom-type descriptors, originally developed for octanol−water partition coefficient (log P) prediction (Table 6.6 in ref 15).

SC-1 spurred critical discussions about the quality of solubility measurements[10−13] and about how computational techniques could be improved.[14−18] On reviewing the SC-1 competition, Palmer and Mitchell[9] compared the prediction quality of the models using the same compounds, training and testing both models on different data and concluding that experimental data quality might not be the limiting factor in predicting the solubility of druglike molecules. Evidently, opportunities still remained to improve the computational methods and the choice of descriptors used in such methods.

It was demonstrated recently that when legacy "gold-standard" saturation shake-flask (SSF) solubility data of druglike substances, which are more numerous and more diverse than those available from CheqSol measurements, are curated critically (e.g., correcting for ionization, temperature, and other effects), the interlaboratory reproducibility is

Received: June 19, 2020
Published: August 14, 2020

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Intrinsic Solubility—Ext. Test Set 1 (Interlab. SD ~0.17)[a]

compound	log S_0	SD	n	mp (°C)	log P	GSE − log S
acetazolamide	−2.38	0.18	11	259	−0.86	−0.98
acetylsalicylic acid	−1.67	0.15	16	142	1.31	−1.98
alclofenac	−4.40	0.16	4	92	2.53	−2.70
ambroxol	−3.87	0.17	3	234	3.19	−4.78
aripiprazole	−6.64	0.21	3	139	4.86	−5.50
atovaquone	−6.07	0.18	3	224	5.51	−7.00
atrazine	−3.69	0.15	6	173	1.78	−2.76
baclofen	−1.78	0.15	4	208	1.86	−3.19
barbital,buta-	−2.22	0.16	10	167	0.79	−1.71
benzthiazide	−4.84	0.22	6	232	2.43	−4.00
bromazepam	−3.39	0.13	3	193[b]	2.63	−3.81
candesartan cilexetil	−6.79	0.15	6	167	6.32	−7.24
carbamazepine	−3.22	0.16	15	192	3.39	−4.56
carbazole	−5.19	0.19	3	246	3.32	−5.03
carbendazim	−4.56	0.19	4	304	1.74	−4.03
cefmenoxime	−3.27	0.14	7	187	−0.87	−0.25
cefprozil	−1.68	0.20	4	222	0.71	−2.18
celecoxib	−5.89	0.18	6	158	3.51	−4.34
cephradine	−1.18	0.13	8	140	0.35	−1.00
chlorpropamide	−3.17	0.14	7	128	1.74	−2.27
cholic acid, deoxy-	−4.62	0.15	7	176	4.48	−5.49
cilostazol	−4.93	0.13	3	160	3.46	−4.31
cimetidine	−1.52	0.22	8	142	0.60	−1.27
ciprofloxacin	−3.57	0.18	20	267	1.58	−3.50
cisapride	−6.78	0.17	6	110	3.36	−3.71
corticosterone	−3.29	0.17	7	182	2.67	−3.74
cortisone acetate	−4.22	0.13	4	222	2.56	−4.03
cyclosporine A	−5.03	0.16	6	151	3.27	−4.03
daidzein	−5.23	0.13	5	330	2.87	−5.42
desipramine	−3.83	0.18	3	100	3.53	−3.78
dexamethasone	−3.56	0.18	16	263	1.90	−3.78
diazoxide	−3.43	0.22	4	329	1.87	−4.41
diclofenac	−5.34	0.18	33	168	4.36	−5.29
diflorasone diacetate	−4.82	0.16	3	223	2.99	−4.47
difloxacin	−3.83	0.21	3	211	2.72	−4.08
diltiazem	−3.02	0.13	3	210	3.37	−4.72
diphenylamine	−3.53	0.14	3	54	3.43	−3.22
DOPA, L-	−1.76	0.17	6	270	0.05	−2.00
enalapril	−1.36	0.21	3	144	1.60	−2.29
estradiol, 17α-	−5.00	0.18	5	215	3.61	−5.01
estrone	−5.38	0.19	8	255	3.82	−5.62
ethoxzolamide	−3.76	0.17	3	189	1.34	−2.48
etoposide	−3.60	0.20	4	244	1.34	−3.03
eucalyptol	−1.66	0.21	3	37	2.74	−2.36
fenbufen	−5.18	0.21	10	186	3.40	−4.51
flumequine	−3.90	0.19	3	253	2.35	−4.13
flurbiprofen	−4.34	0.20	23	111	3.68	−4.04
folic acid	−5.96	0.16	6	250	−0.04	−1.71
ganciclovir	−1.78	0.13	3	250	−1.97	0.22
glipizide	−5.61	0.21	9	209	2.08	−3.42
griseofulvin	−4.52	0.19	15	220	2.69	−4.14
haloperidol	−5.71	0.17	10	151	4.43	−5.19
ibrutinib	−4.85	0.19	7	155	4.22	−5.02
indinavir	−4.53	0.16	5	168	2.87	−3.80
indomethacin	−5.48	0.22	21	159	3.93	−4.77
indoprofen	−4.65	0.21	5	214	3.04	−4.43
ketoconazole	−5.47	0.14	11	146	4.21	−4.92
maprotiline	−4.62	0.22	5	92	4.21	−4.38
metolazone	−3.88	0.21	8	256	2.71	−4.52
nabumetone	−4.40	0.21	3	80	3.37	−3.42
naproxen	−4.23	0.16	17	153	3.04	−3.82

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. continued

compound	log S_0	SD	n	mp (°C)	log P	GSE − log S
nelfinavir	−6.21	0.20	3	350	4.75	−7.50
nevirapine	−3.41	0.14	6	248	2.65	−4.38
nifedipine	−4.71	0.15	11	173	2.18	−3.16
nimesulide	−4.74	0.14	5	144	2.76	−3.45
norfloxacin	−2.88	0.16	19	221	1.27	−2.73
nortriptyline	−3.93	0.16	5	214	3.83	−5.22
noscapine	−4.48	0.14	3	176	2.88	−3.89
ofloxacin	−2.03	0.13	14	254	1.54	−3.33
oxazepam	−4.03	0.17	5	206	2.45	−3.76
oxyphenbutazone	−3.94	0.19	3	96	3.49	−3.70
papaverine	−4.33	0.19	12	147	3.86	−4.58
perphenazine	−4.48	0.17	6	97	3.94	−4.16
phenacetin	−2.30	0.14	10	135	2.04	−2.64
phenazopyridine	−4.02	0.16	7	139	2.66	−3.30
pindolol	−3.75	0.15	9	170	1.91	−2.86
pravastatin	−4.86	0.15	10	326	2.44	−4.95
prednisolone, methyl-	−3.33	0.18	5	233	1.80	−3.38
primidone	−2.53	0.14	4	282	0.54	−2.61
probenecid	−4.83	0.20	4	197	2.20	−3.42
promazine	−4.45	0.13	4	33	4.24	−3.82
promethazine	−4.38	0.19	11	60	4.24	−4.09
repaglinide	−4.77	0.17	4	131	5.22	−5.78
resveratrol, trans-	−3.75	0.18	7	254	2.97	−4.76
ritonavir	−5.17	0.16	5	121	5.91	−6.37
rofecoxib	−4.61	0.16	5	207	2.56	−3.88
spironolactone	−4.21	0.16	6	135	4.85	−5.45
strychnine	−3.38	0.19	6	275	2.09	−4.09
sulfasalazine	−6.41	0.14	9	220	1.80	−3.25
sulfathiazole	−2.62	0.22	9	202	1.53	−2.80
sulfisomidine	−2.16	0.14	3	243	1.48	−3.16
sulfisoxazole	−3.13	0.14	3	191	1.67	−2.83
sulindac	−4.96	0.21	7	184	4.37	−5.46
tetracaine	−3.11	0.11	3	149	2.62	−3.36
tetracycline	−3.22	0.15	8	165	−0.37	−0.53
thiacetazone	−3.50	0.16	10	225	0.81	−2.31
triamcinolone	−3.52	0.21	5	270	0.62	−2.57
triamterene	−4.11	0.14	9	313	0.83	−3.21
warfarin	−4.78	0.20	11	161	3.61	−4.47
xanthine	−3.60	0.21	3	300	−1.06	−1.19
Min	−6.8	0.11		33	−2.0	RMSE = 1.1
Max	−1.2	0.22		350	6.3	GSE
Mean	−4.0	0.17		191	2.6	

[a]Log S_0 = interlab. mean equilibrium solubility (molarity units), 25 °C. SD = std. dev. of mean. n = no. of lit. refs. Log P = octanol−water partition coefficient, calc. by RDKit. [b]Melting point calc. by open-source program:www.qsardb.org/repository/predictor/10967/104?model=rf.

evidently ∼0.17 log unit,[10] much lower than the ∼0.6 log unit noted earlier.[6−9] In the second solubility challenge, SC-2,[1] we were motivated to revisit the competition to (1) examine the extent to which computational methods have improved over the last ten years, (2) verify that data quality (or rather test set quality) may not be the main limiting factor in the accuracy of the prediction, by considering two test sets (one characterized by good interlaboratory reproducibly, and the other with poorly reproducible measurements) with both sets comprising "gold standard" methods, and (3) examine to what extent public solubility databases (many older than ten years) have kept up with the expanding chemical space of today's drugs.

To address the first point, participants were asked in the submission forms to identify the computational method (e.g., MLR, PLS, PCR, ANN, kNN, SVM, RFR, or specify other) and the descriptors used.

To address the second point, two new test sets of druglike molecules were gathered. The first set consisted of 100 drugs whose log S_0 values (mostly SSF type) were collected and curated from multiple published sources, where the average interlaboratory standard deviation, SD, was estimated to be ∼0.17 log unit. These are drugs whose solubility is "well determined." It is noteworthy that although the CheqSol measurements in SC-1 indicated an internal precision of 0.05 log unit,[2,3] the comparisons of such determinations between different laboratories suggested an average SD = 0.15 log unit, and the comparison between CheqSol and high-quality SSF measurements suggested RMSE = 0.34 log unit.[5] It is also important to highlight that the reasons for the poor laboratory reproducibility (high SD) are multiple and diverse and not always due to experimental inconsistencies or systematic or random errors of the experimental protocol. There are other

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Intrinsic Solubility—Ext. Test Set 2 (Interlab. SD ~0.62)[a]

compound	log S_0	SD	n	mp (°C)	log P	GSE-log S
amantadine	−2.19	0.50	3	180	1.91	−2.96
amiodarone	−10.40	0.50	5	156	6.94	−7.75
amodiaquine	−5.49	0.65	3	208	5.18	−6.51
bisoprolol	−2.09	0.59	3	100	2.37	−2.62
bromocriptine	−5.50	0.51	5	217	3.19	−4.61
buprenorphine	−6.07	0.83	3	210	4.41	−5.76
chlorprothixene	−5.99	0.51	6	98	5.19	−5.42
clofazimine	−9.05	0.93	5	211	7.49	−8.85
curcumin	−5.36	0.68	3	177	3.37	−4.39
danazol	−6.10	0.52	10	229	4.22	−5.76
didanosine	−1.24	0.54	3	162	−0.21	−0.66
diflunisal	−4.99	0.56	11	214	3.04	−4.43
diphenhydramine	−3.21	0.55	4	169	3.35	−4.29
etoxadrol	−1.96	0.55	3	124[b]	2.81	−3.30
ezetimibe	−4.94	0.51	4	165	4.89	−5.79
fentiazac	−5.84	0.65	4	161	4.76	−5.62
iopanoic acid	−5.49	0.66	3	155	3.74	−4.54
itraconazole	−8.98	0.61	3	165	5.58	−6.48
miconazole	−5.82	0.50	6	161	6.45	−7.31
mifepristone	−5.22	0.75	4	194	5.41	−6.60
omeprazole	−3.70	0.50	3	156	2.90	−3.71
pioglitazone	−6.20	0.66	4	184	3.16	−4.25
procaine	−2.30	0.60	3	61	1.77	−1.63
quinine	−3.06	0.57	7	177	3.17	−4.19
raloxifene	−6.82	0.56	6	145	6.08	−6.78
rifabutin	−4.09	0.66	3	176[b]	4.62	−5.63
saquinavir	−5.92	0.58	3	350	3.09	−5.84
sulfadimethoxine	−3.74	0.70	3	204	0.88	−2.17
tamoxifen	−7.52	0.72	7	98	6.00	−6.23
telmisartan	−6.73	0.84	5	262	7.26	−9.13
terfenadine	−7.74	0.71	11	150	6.45	−7.20
thiabendazole	−3.97	0.50	4	305	2.69	−4.99
Min	−10.4	0.50		61	−0.2	RMSE = 1.2
Max	−1.2	0.93		350	7.5	GSE
Mean	−5.2	0.62		181	4.1	

[a]Log S_0 = interlab. mean equilibrium solubility (molarity units), 25 °C. SD = std. dev. of mean. n = no. of lit. refs. Log P = octanol−water partition coefficient, calc. by RDKit. [b]Melting point calc. by open-source program:www.qsardb.org/repository/predictor/10967/104?model=rf.

kind of errors contributing equally to the high variability of data which in general are more difficult (or even impossible) to identify (i.e. typographical errors and wrong compound/form/solid state, reporting a "different" solubility, intrinsic vs aqueous or just citing over and over a wrong value in the literature making it to become the "true" value over time).

The second test set consisted of 32 "difficult" drugs, characterized by poor interlaboratory reproducibility: SD ~0.62 log unit (mostly SSF type). Nearly, a third of these "inconsistently determined" drugs possesses intrinsic solubility less than 1 μM, which is probably the main reason for the poor overall reproducibility. Furthermore, several of these are located in a sparsely populated chemical space,[70] with very few nearby known similar molecules (e.g., amiodarone, clofazimine, and itraconazole). Therefore, accurate prediction of their solubility was expected to be challenging.

To address the third point, the participants were invited to select their own training set to construct the prediction model, with the expectation that "fresh" and large diverse collections might be presented. The entry form asked the participants to give references to the sources of the training set. The form also asked for the number of solubility values used in the model training (n_{tr}) and in model validation (n_{va}). The participants were explicitly asked not to include in their training sets any of the SC-2 test set molecules. The inclusion of the same molecule in both the training and test sets increases the performance of the prediction method. For example, the random forest regression (RFR) method would simply present the experimental value from the training set as the prediction value. This would not be a fair test of the prediction capability of the method.

RESULTS

Analysis of the Results. Molecular descriptors used in QSPR models quantify properties of single molecules and, in general, lack accurate descriptors describing the long-range order phenomenon. Because the solubility of a crystalline molecule depends upon the free energy required to remove molecules from the crystal lattice as well as the free energy for solvation (Gibbs free energy change from crystal to supercooled liquid to solution and from crystal to vapor to solution)[44] QSPR models might not be adequate to model the solid state. Because these energies are difficult to measure and even to predict, octanol−water partition coefficients (log

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

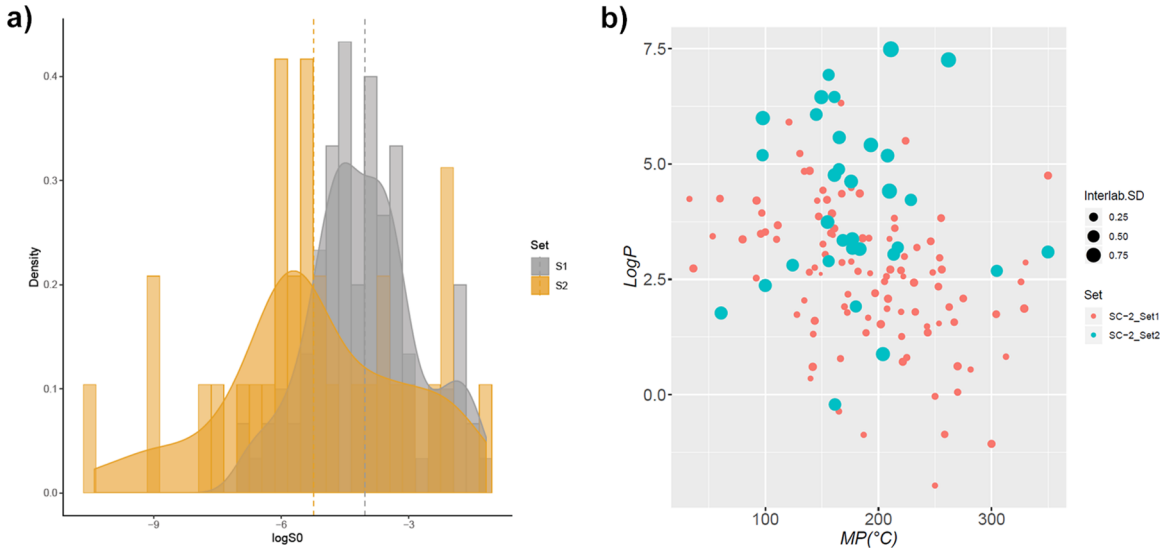


Figure 1. (a) Log S_0 histogram plot for each set. (b) Plot Log P vs Melting Point.

Table 3. Predictions of Intrinsic Solubility[a]

	SET 1				SET 2											
CODE	R^2	RMSE	bias	%±0.5 log	R^2	RMSE	bias	%±0.5 log	Method	Training Set Refs.	n(tr)	n(va)	n(descr.)	(i)[b]	(ii)[c]	(iii)[d]	Comments
ASLL_A	-0.24	1.41	-1.15	16	-0.23	2.38	-1.91	16	XGBoost Regr.	2,9,22,23,33,35-37	4741	1186	152	x	x		Selected CDK 2D descr. & in-house 2D descr. reflecting atom counts & molecular properties.
ASLL_B	0.17	1.16	-0.64	36	0.08	2.06	-1.48	19	Gradient Boosted Trees Regr.	2,9,22,23,33,35-37	4741	1186	152	x	x		Selected CDK 2D descr. & in-house 2D descr. reflecting atom counts & molecular properties.
ASLL_C	0.34	1.03	-0.68	46	0.02	2.12	-1.51	22	Extremely Rand. Trees Regr.	2,9,22,23,33,35-37	4741	1186	152	x	x		Selected CDK 2D descr. & in-house 2D descr. reflecting atom counts & molecular properties.
ASTU	0.35	1.02	-0.10	37	0.66	1.25	0.10	25	MLR	not specified			2			x	Method: logSw = -1.120E - 0.599ClogP, where E = excess molar refraction (Abraham descr.)
FLWMU	0.29	1.07	-0.57	36	-0.23	2.37	-1.74	16	Consensus: kNN, GBM, RF, SVM	60,61	5778	1445		x			online chem. database generated by e-Bitter program[63]
HPSU_A	-2.50	2.37	-0.03	19	-0.02	2.17	-0.67	25	ANN	19-22	222			x			distribution of molecular electrostatic potentials in 3D
HPSU_B	-1.78	2.11	0.64	23	-0.10	2.25	1.68	16	Extened Solvent Contact Model	22,57,58	371			x			solvation function
JCSU_A	0.48	0.92	-0.32	39	0.58	1.39	-0.64	38	RFR	2-4,24-28,31,32,38-41	333	NA	269	x			All MOE 2D descriptors except LogS, h_logS and Largest Ring Size
JCSU_B	0.37	1.01	-0.41	39	0.44	1.60	-0.76	28	RFR	2-4,24-28,31,32,38-41	333	NA	269	x			Descriptors calc. using Python package RDKit. Highly correlated (>0.95) descr. not used.
JHTNY	0.38	1.00	-0.50	45	0.19	1.93	-1.27	25	Message Passing NN	21,37,42,61	10720	1340	152	x			Message Passing Neural Network Code: https://github.com/swansonk14/chemprop/
JHUNC_A	0.57	0.83	-0.26	61	0.30	1.79	-0.89	28	ANN	2,3,19-34	312	0		x			RDKit MorganFingerprint, HashedAtomPairFingerprint, & other Fingerprints concatenated
JHUNC_B	-0.78	1.69	-1.41	17	0.07	2.07	-1.29	25	ANN	2,3,19-34	312	0		x			Molecular Graph, input to Graph Convolution Neural Network
JHUNC_C	-0.82	1.71	-1.43	14	0.34	1.74	-1.12	19	ANN	not specified				x			Molecular Graph, input to Graph Convolution Neural Network
JMSA_A	0.40	0.98	-0.40	46	0.52	1.49	-0.82	34	Consensus of ML methods	2,3,27,32,33	117	36	173	x			CDK descr. with non-zero variance used; some SMILES replaced with aromatised versions
JMSA_B	0.44	0.95	-0.36	46	0.50	1.52	-0.81	31	extraTrees Regr.	2,3,27,32,33	117	36	173	x			CDK descr. with non-zero variance used; some SMILES replaced with aromatised versions
JMSA_C	0.39	0.99	-0.41	45	0.51	1.49	-0.81	31	RFR	2,3,27,32,33	117	36	173	x			CDK descr. with non-zero variance used; some SMILES replaced with aromatised versions
KSMIT	0.45	0.94	-0.48	47	0.36	1.71	-0.95	31	Message passing NN	2,3,24-28,31-34,37,42	10237	1138		x			Atom & bond descr.[62] RDKit 200 descr.
MCSMD	0.46	0.93	-0.41	50	0.66	1.24	-0.51	38	ANN	42	2666	4x cross val.		x			LogP(calc), atom counts, ring count, bond count, TPSA, HBD, HBA
MLKC_A	0.60	0.80	-0.32	51	0.61	1.34	-0.64	38	lightGBM	2-5,23-28,30-32,34, 37,39,43-53	881	164		x			Dragon 6.0 all descriptors; RDKit FingerPrints
MLKC_B	0.60	0.80	-0.33	52	0.60	1.36	-0.63	44	lightGBM	2-5,23-28,30-32,34, 37,39,43-53	881	164		x			Dragon 6.0 all descriptors; RDKit FingerPrints
MLKC_C	0.60	0.80	-0.36	44	0.61	1.33	-0.65	44	lightGBM	2-5,23-28,30-32,34, 37,39,43-53	881	164		x			Dragon 6.0 all descriptors; RDKit FingerPrints
NMUPI	-2.63	2.41	-0.79	17	-0.96	3.00	-1.65	16	ANN	59	4394	941		x			Descr.- https://github.com/mordred-descriptor/mordred; doi: 10.1186/s13321-018-0258-y
PMSA_A	0.60	0.80	0.06	43	0.69	1.18	-0.16	44	RBF (radial basis function)	in-house pharma data	2220	554	168	x			StarDrop Standard Set
PMSA_B	0.54	0.86	0.12	39	0.72	1.13	-0.29	38	RBF (radial basis function)	in-house pharma data	704	174	167	x			StarDrop Standard Set
PMSA_C	0.62	0.78	0.09	54	0.65	1.27	-0.22	41	RBF (radial basis function)	in-house pharma data	7841	1959	164	x			StarDrop Standard Set
RFSP_A	0.38	1.00	-0.41	46	0.21	1.91	-1.40	25	ANN	19,22,30	2641	955	168		x		
RFSP_B	0.33	1.04	-0.51	44	0.10	2.03	-1.52	25	ANN	19,22,30	2641	955	168		x		
SGURV	0.28	1.07	-0.09	56	0.71	1.16	-0.40	34	ANN	24-26,28,31,39,54	102	12	60			x	https://github.com/mordred-descriptor/mordred
TDIPG	0.23	1.11	-0.38	35	0.66	1.24	-0.38	38	ANN	5,26	248	60		x			MOE 2D Descriptors (all)
UMUT_A	0.42	0.97	-0.27	48	0.74	1.10	-0.10	31	MLR	5	81	42		x			Software: Dragon descriptors: ALOGP2, SMO4_EA (bo)
UMUT_B	0.45	0.94	-0.16	42	0.62	1.32	-0.31	34	MLR	2,3,5,24-28,31-34,55	346	90		x			RDKit descriptors: MolLogP, TPSA
UMUT_C	0.38	0.99	-0.27	42	0.75	1.06	-0.37	44	MLR	2,3,5,24-28,31-34,55	346	90		x			xlogs, SpMax1_Bhp (PaDEL-Descriptor), SHBd (PaDEL-Descriptor)
XWUC_A	0.27	1.08	-0.51	42	0.60	1.36	-0.91	28	Graph Convolution NN	42	8000	1961		x			Molecule graph as descritors
XWUC_B	0.30	1.06	-0.52	39	0.60	1.35	-0.53	25	Graph Convolution NN	42	8000	1961		x			Molecule graph as descritors
YTACU	-0.05	1.30	0.10	28	0.55	1.44	0.53	22	XGBoost	56	212		11	x			Lipinski desc., log P, rel. neg. partial charge, principal moment of inertia, and others
YUMPU_A	0.64	0.76	-0.05	59	0.75	1.08	-0.25	47	light GBM	5	124	5x cross val.		x	x		Mordred descr. & predicted logS & solubility, total 13 kinds of descr.
YUMPU_B	-0.23	1.41	0.18	51	0.46	1.58	0.55	31	ANN	5	93	31		x	x		Dragon descr. & predicted logS & solubility, total 21 kinds of descr.
mean	0.09	1.14	-0.36	40	0.39	1.62	-0.67	30									
min	-2.63	0.76	-1.43	14	-0.96	1.06	-1.91	16									
max	0.64	2.41	0.64	61	0.75	3.00	1.68	47									

[a]Summary of information regarding methods used and training set data references, and comments were provided by the participants in the competition forms. n(tr) and n(va) refer to the number of molecules used in the training and internal validation sets, respectively. [b](i) "x" refers to participant agreeing that all test set molecules also found in the training set have been removed. The grey highlight indicates that there may possibly be an overlap between user-provided training and competition test sets. [c](ii) "x" refers to data coming from a commercial source, where (presumably) training set molecules could not be filtered to avoid overlaps with test set molecules. [d](iii) "x" refers to the case that neither (i) or (ii) conditions were met.

P) and melting points (mp) have been traditionally used as surrogates for crystal packing and sublimation energy contributions to the solubility. It is therefore important to double check that compounds belonging to the "brick dust" category (compounds with high melting points) and to the "grease ball" one (compounds very lipophilic, with high log P) are present in both test sets, and their distribution is not too dissimilar. These class of compounds will be naturally more challenging to measure, often requiring deviation from the standard methodology and having to adapt the experimental methods to each specific case, making it more likely to have higher interlab variability and likely more difficult to predict.

Tables 1 and 2 list the averaged 25 °C intrinsic aqueous solubility values, as log S_0, of the two test sets of molecules,

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

along with their interlaboratory standard deviations (SD) and the number (n) of independent literature sources used in the averaging. The details regarding the test sets are described elsewhere.[1]

Figure 1a shows a log S_0 histogram plot for the two test sets. Distributions of the two sets are overlapping well for the mid-to-high soluble fraction of compounds (with a higher weight of a more soluble fraction of compounds for the set 1); however, set 2 shows a higher fraction of poorly soluble molecules, which supports the hypothesis of a more challenging set to measure S_0 accurately, therefore with a higher interlaboratory variability. Test Set 1 log S_0 values ranged from −1.2 to −6.8, with a mean value of −4.0. The interlaboratory SD ranged from 0.11 to 0.22 log, with a mean value of 0.17 log. Test Set 2 values ranged from −1.2 to −10.4, with a mean value of −5.2. The corresponding SD ranged from 0.50 to 0.93 log, with a mean of 0.62 log.

Figure 1b shows that "grease balls" (log P > 4 and mp < 160 °C) and "brick dust" (log P < 3 and mp > 200 °C) compounds are included in both test sets showing a reasonable good broad distribution in both, with a slightly higher fraction of more lipophilic compounds in Set 2 and a significant higher fraction of higher melting point compounds in Set 1 (however, note that despite the high melting points, these are not true "brick dust" compounds because only six compounds (6%) in this set have intrinsic solubilities <1 μM). The size of the circles is scaled by the average error and represents the SD for each compound. It is noteworthy to see that there is not a clear correlation between extreme case compounds and the SD.

This second solubility challenge formally started 1 May 2019, and submissions were accepted until 8 September 2019. Twenty different groups submitted their entries; in several cases, multiple entries were submitted from the same group. The total number of complete and accepted entries was 37. Table 3 lists the statistical analysis of the 37 predictions provided by 20 participating groups, arranged by entry codes. The literature references to the training sets used are provided.[19−63] The R^2, RMSE, bias (the usual equations for the three terms are defined in ref 36), and "%±0.5 log" (percentage of predicted values within 0.5 log unit of the test set values) statistics measure the degree to which the predictions were effective. The prediction methods used are also listed. Even though we did not limit the challenge to any particular model, all competitors submitted predictions based on QSPR approaches. The five most frequently used methods were neural networks (30%: artificial neural networks and 8%: radial basis functions), MLR (11%: multiple linear regression), methods based on decision trees (11%: light gradient boosting machines, 8%: random forest regression, EXTremely RAndomized TREES, and LightGBM, XGBoost), two consensus models, and one extended solvent contact model.

As a benchmark against which to compare proposed methods, solubility predictions using the general solubility equation (GSE), log S_0 = 0.5 − log P −0.01(mp—25),[64−66] were also provided to participants and listed in Tables 1 and 2. The experimental melting points (mp, °C) and the RDKit-calculated[67] octanol−water partition coefficients, log P, are also listed in Tables 1 and 2. The GSE RMSE values for the two test sets are nearly the same, 1.1 and 1.2 log unit, respectively, and do not appear to be affected by the differences in the precision of the two test sets. As shown previously in Figure 1a,b, the drugs in the "contentious" Test Set 2 are more lipophilic (mean log P 4.1) and less soluble (see above) than those in the "agreeable" Test Set 1 (mean log P 2.6). The traditional GSE is a simple method which requires no training.[64−66]

Training Sets, the Impact of Non-Druglike Molecules, and the Search for "Missing" Molecules. The training set references in Table 3 indicate that new databases have been collected since the original SC-1. In the SC-2 competition, CheqSol solubilities were used most often to build training sets. There were 16 entrant building prediction models based on the original SC-1 training set[2] and 13 based on the SC-1 test set.[3] Other studies featuring the potentiometric method included refs 4(5 entrants) 5,(9 entrants) 31,(11 entrants) 34, (8 entrants), and 47−51 (3 entrants each). The original potentiometric method, pSOL,[68] had a smaller following: refs 24 and 25 (11 entrants each) 39, (6 entrants) 46, (4 entrants), and 54 (1 entrant). There appear to be 233 published CheqSol values for 145 molecules.[5] The count is estimated to be 75 for the pSOL method.[69] The miniaturized SSF measurements as a function of pH from the Bergström's group (converted to intrinsic values) were extensively used: refs 26, 28, 32, 55 (12, 13, 11, 13, and 2 entrants, resp.). With the exception of the collection in ref 5, all of the above references point to single-source measurements of drugs.

Databases of solubility values collected from multiple sources[9,19−23,33,35−38,40,42,52,53,56,58−60] were also the makings of training sets in SC-2. Some of these databases were cherry-picked from earlier databases. This can be still useful, provided these include references to the original publications where the measurements were performed, in case clarifications of values are required. Even well-known databases have some minor mistakes in them (e.g., not identifying "water solubility" in a sole measurement as that of the salt and not the free acid/base), so listing references to the original publications is helpful. Rytting et al.[29,30] measured the water solubilities of 122 drugs using a consistent SSF method (without reporting pH). Because many of these drugs are ionizable, it is necessary to calculate the intrinsic solubility values from the knowledge of the corresponding pK_a values, with the assumption that the Henderson−Hasselbalch equation[69] is valid. Abraham and Le[52] have a very useful discussion of the calculation procedure. A concern is that this calculation is not always carried out when water solubilities (S_w) are mixed into an intrinsic solubility (S_0) database. Sorkun et al.[42] compiled the impressive AqSolDB database consisting of 9982 solubility values (freely downloadable). Unfortunately, the specific references to the original measurements are not available in the downloaded set. Because many of the molecules are ionizable, it is quite a feat to convert values to the intrinsic scale. Also, AqSolDB, as well as the compilations of Yalkowsky,[19−21] Howard and Meylan,[22] and Huuskonen,[23] consists of industrial organic molecules, not all solids at room temperature, and many agrochemicals (herbicides, pesticides, insecticides, rodenticides, and acaricides), which may not be the sort of molecules occupying the chemical space of drugs.

Unfiltered Use of Ref 5 as the Training Set. Nine of the 37 entries included the CheqSol-SSF data from Ref 5 in their training sets, without indicating which of the two compared sets (CheqSol, SSF) was used. Three entries (UMUT_A, YUMPU_A, and YUMPU_B) solely used the 124 values for their method training. Twenty-seven molecules in the CheqSol-SSF publication[5] are also found in Test Set 1 (27%), and 13 molecules overlap with those in Test Set 2 (41%). All nine entries checked the box in the entry form,

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

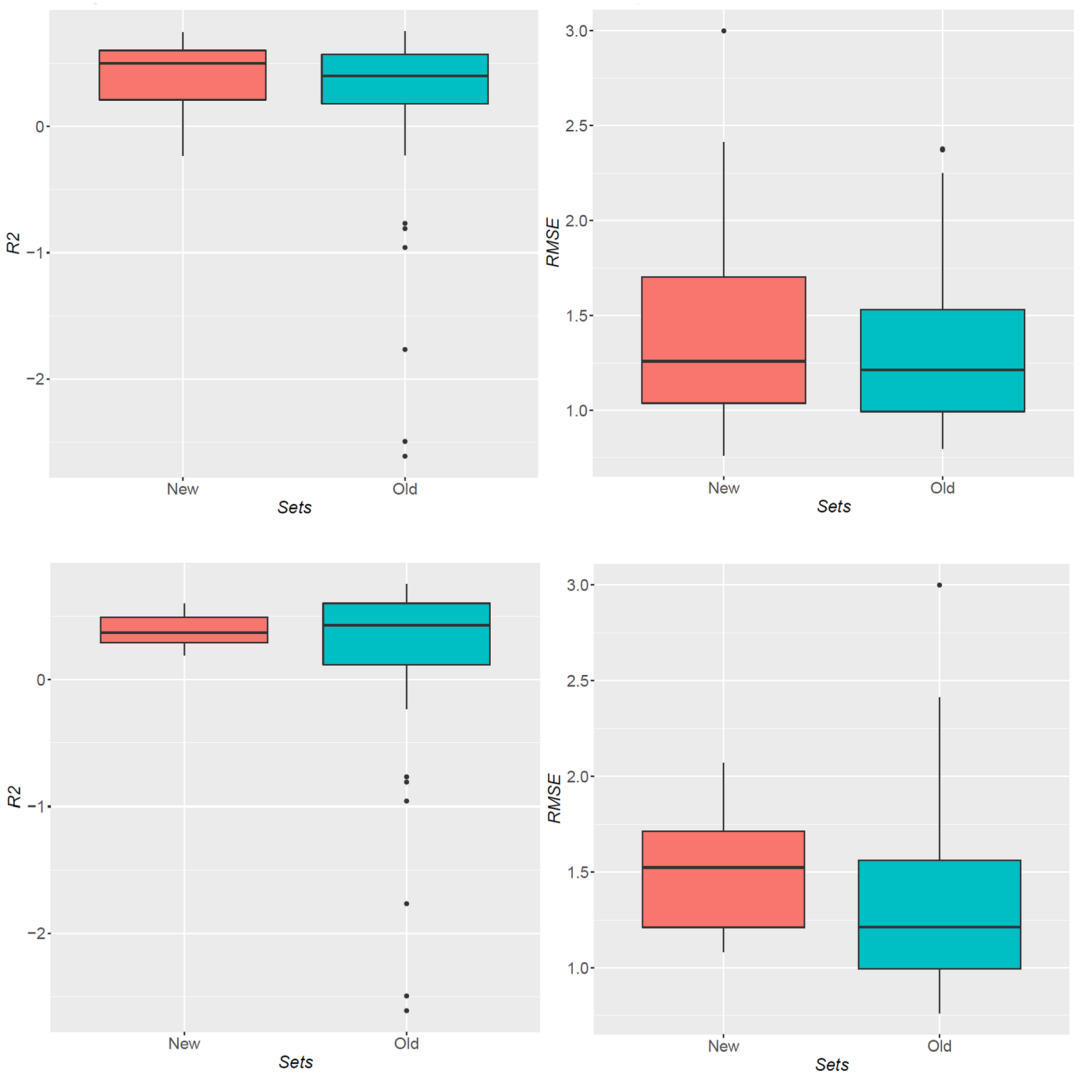


Figure 2. Comparison predictive accuracy (R^2 and RMSE) of "Old" and "New" methods.

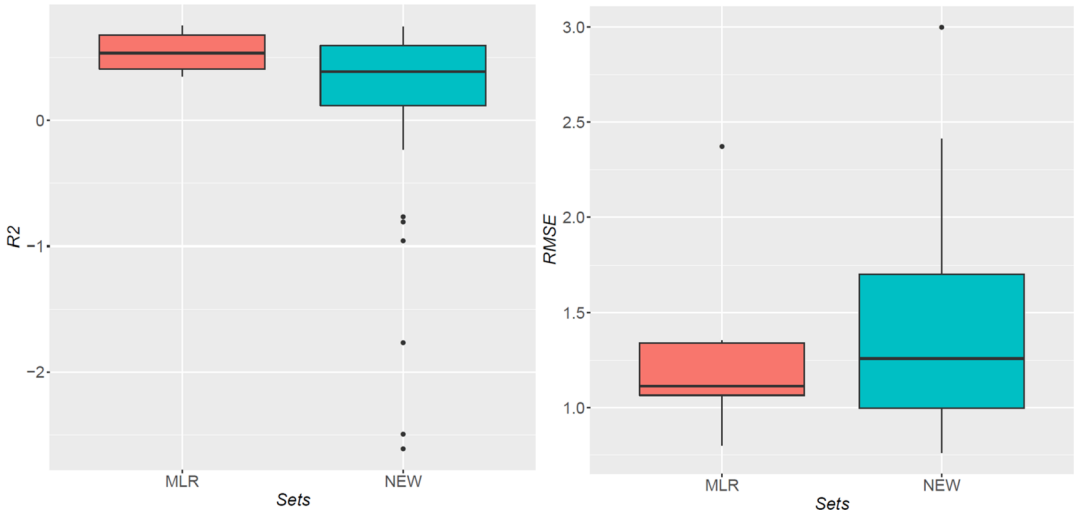


Figure 3. Comparison predictive accuracy (R^2 and RMSE) of "New" methods against MLR.

acknowledging that all molecules in the training set that are also found in the test set have been removed prior to the training procedure. However, the n(tr) counts in three entry forms suggested that all 124 of the CheqSol-SSF molecules were used, which is an inconsistency. Although the use of such overlapping molecules improves the prediction, it cannot be a recommended statistical procedure; therefore, we have colored

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

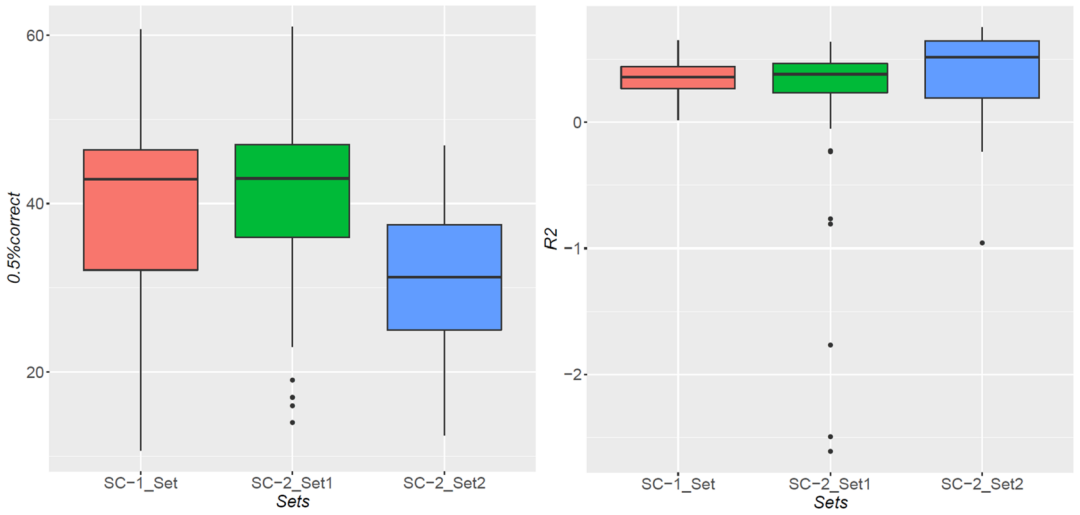


Figure 4. Comparison predictive accuracy (±0.5 log S_0 and R^2) using test sets from SC-1 and both test sets from this challenge: SC-2_Set1 (low SD) and SC-2_Set2 (high SD).

in grey (Table 3) those entries which appear to possess overlapping test and training sets.

DISCUSSION

It is important to highlight that even though this challenge was open to any kind of approach to solubility prediction, all participants chose to use some sort of QSPR or ML (Machine Learning) approach. Therefore, our conclusions can only be extended to these methods.

From the comparison of the outcome of SC-1 with SC-2, it is clear that there is no significant difference regarding prediction performances, especially when "tight" data (Test Set 1) are compared. The first solubility challenge could not identify any definitive methods performing better than others, but all methods and combinations performed "equally" well.

Figure 2 shows a comparison between the prediction performance (R^2 and RMSE) obtained with the methods used by contestants 10 years ago in the first SC (old methods—MLR, RFR) with the newest ones used in this challenge (Graph Convolution NN, Message passing NN, light GBM, XGBoost, and EXTremely RAndomized TREES). The ranges in the predicted versus measured R^2 and RMSE for log S_0 are not significantly different from the original SC-1 results. This comparison clearly shows that the use of the more sophisticated state-of-the-art methods does not improve the outcome of predictions. In fact, we reach the same conclusions as we did 10 years ago: we observe no prediction improvement of the new methods over the classic MLR ones (see Figure 3).

Comparing the prediction accuracy, defined as % correct predictions for ±0.5 log S_0, also shows that the new methods used in this challenge are not performing better than the ones used in the previous one, at least when comparing the low-variance sets (Figure 4).

The prediction accuracy (defined as % correct predictions for ±0.5 log S_0) also shows the same performance when comparing the low variance sets (see Figure 4). Both "high quality" sets (the one from the previous SC-1 and Test Set 1 from this challenge) can be predicted equally well: the same mean number of compounds accurately predicted (±0.5 log S_0) and the same mean R^2. However, it is very clear that the predictions become worse when a high variability set is used. When the accuracy of the measured test set decreases, all the prediction performance parameters significantly worsen (see table below). Figure 5 clearly shows a significant increase in the RMSE when both sets are compared. The RMSE mean for the "high quality" set is ca. 1.10 (comparable to the GSE value), while the RMSE mean for the "highly variable" set is ca. 1.58. The prediction of the "highly variable" set also performs significantly worse when compared to the outcome of the first SC (% correct predicted ~30 against 43%), which was done on a "high-quality" very tight solubility data with an internal precision of 0.05 log unit (Figure 4). In addition to the quality of measurement issues, the diminished prediction performance, in part, may be due to the absence of training set molecules that occupy the edges of the chemical space of such practically insoluble drugs as amiodarone, clofazimine, and itraconazole.[70]

Figure 6 shows the log S_0 correlations for the top performers (based on RMSE) for five of the most representative methods (RBF, lightGBM, ANN, RFR, and MLR). Figures 7 and 8 show a graphical overview of the prediction accuracy (RMSE and % correct predictions) of all 37 valid entries. It is clear that all entries are able to predict more accurately the intrinsic solubilities using the tight set (blue bars) than the loose set (red bars). Typically, the RMSE differences between both sets are between 0.3 and 1.0. Almost all entries showing a similar prediction accuracy between both sets (ΔRMSEs < 0.3) have a RMSE > 1, indicating that the model predicts "equally badly" for both sets. However, several entries with a RMSE < 1 show differences below 0.2 indicating that the predictive accuracy is similar for both sets. It would be interesting to know more details about the methods and training sets used by the authors in these few cases. In general, good prediction models have RMSEs between 0.7 and 1.1. In this challenge, 73% of predictions for the tight set of compounds is below 1.1, while only 5% has RMSEs below 1.1 for the loose set, indicating that the quality of the test set (or possibly the dissimilar chemical space of training set molecules) has a big impact on the accuracy of the predictions. However, for a model to be a "useful" predictive model, it should provide better predictions than the naive estimate of the mean of all predictions, reducing the randomness more efficiently than the SD, and therefore, the SD is the value that a good model should beat.

It is clear (Figure 5) that modeling the tight set of data generates significant smaller RMSE values than those obtained

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

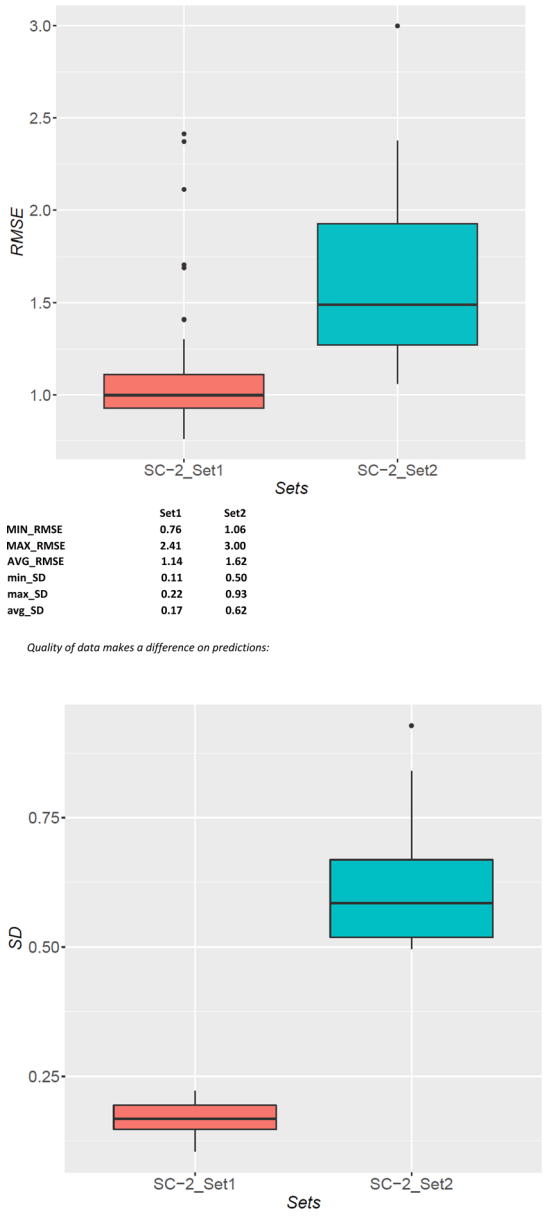


Figure 5. Comparison test sets from this challenge: SC-2_Set1 (low SD) and SC-2_Set2 (High SD).

by modeling the loose set (1.14 vs 1.62 on average). On this measure, the tight data are better predicted. Similarly, the average percentage of correct predictions (40% vs 30% within the same 0.5-unit tolerance) again suggests that the tight set is better predicted, but not in terms of R^2. The loose set has a better average predicted R^2 (0.30 tight vs 0.40 loose). This is likely explained by the wider range of extreme solubilities in the loose set (see Figure 1a), with 9.16 log S units span between the min and max solubility values compared to only 5.61 for the tight set. The deterioration seen in the prediction quality of the models (decreased RMSE from 1.14 to 1.62) for the loose set is as expected in line with the increase in the average interlaboratory SD (Figure 5). In this sense, benchmarking the RMSE against the SD would give 1.14 − 0.17 = 0.97 log S_0 units for the tight set while for the loose set gives 1.62 − 0.62 = 1.00 log S_0 units. Similarly, we could do the same with the average SD for the 100 (or 32) quoted gold standard values; then, the tight set gives 1.14 − 1.27 = −0.13 log S_0 units, and the loose set gives 1.62 − 2.14 = −0.52 log S_0 units. The negative values show that the "average" models for both sets are "useful" (better than predict-average-for-all). From the total of 37 submitted models, 29 give a negative RMSE-SD value (useful) for the tight set and 32 for the loose set. Mitchell in a very recent paper describes the use of the ratio RMSE/SD.[71] The tight set gives 1.14/1.27 = 0.90, and the loose set gives the rather better (smaller) ratio 1.62/2.14 = 0.76. Hence, in absolute terms, the tight set is predicted better, but taking into account the greater variance of the loose set, there is no significant difference of the prediction quality between the models based on the different data sets.

This is the same conclusion Palmer and Mitchell reached in the 2014 paper for the same set of compounds. In that paper, the same compounds were used to do the comparison of the prediction quality of the models (even if models were trained and tested on different data), and the conclusions were that there was no significant difference in the prediction quality for the low and high quality sets (even when models were trained and tested on more accurate solubility data).

This challenge has clearly shown that there is room for improvement of the accuracy of the predictive models and that it is therefore critical that careful data curation and validation goes into the generation of the data sets, with adequate coverage of the chemical space of druglike molecules, but it has also shown that with the present design, it is beyond the power of the currently used machine learning modeling methods to answer this question unambiguously.

CONCLUSIONS

1 No improvement in the prediction of solubility is recognizable in 10 years since the first competition. New methods appear to perform equally well (or badly) as older ones, even similarly to MLR (same result as from the SC-1).

2 Test set quality matters. In absolute terms, the tight set is predicted better. The solubility of inconsistently determined molecules (Test Set 2), especially of practically insoluble molecules from the sparsely populated chemical space, is more difficult to predict, compared to the prediction of the consistently determined solubility of molecules selected from well-represented parts of the chemical space of druglike molecules (Test Set 1 and molecules from SC-1). However, when model prediction accuracies are considered in the context of the average errors of each data set (SD), performances of the models built on Set 1 (low SD) and Set 2 (high SD) are about the same. It is therefore beyond the power of currently used machine learning modeling methods in the present design of this challenge to demonstrate this unambiguously.

3 There is room for predictive accuracy improvement based on improving data quality of training sets used. In future competitions, it would be desirable to draw on a single critically curated training set of intrinsic aqueous solubility values of at least several thousand published druglike molecules, which demonstrably cover the chemical space of drugs. This would make it easier to recognize significant improvements in the prediction methodology and the selection of descriptors used therein.

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

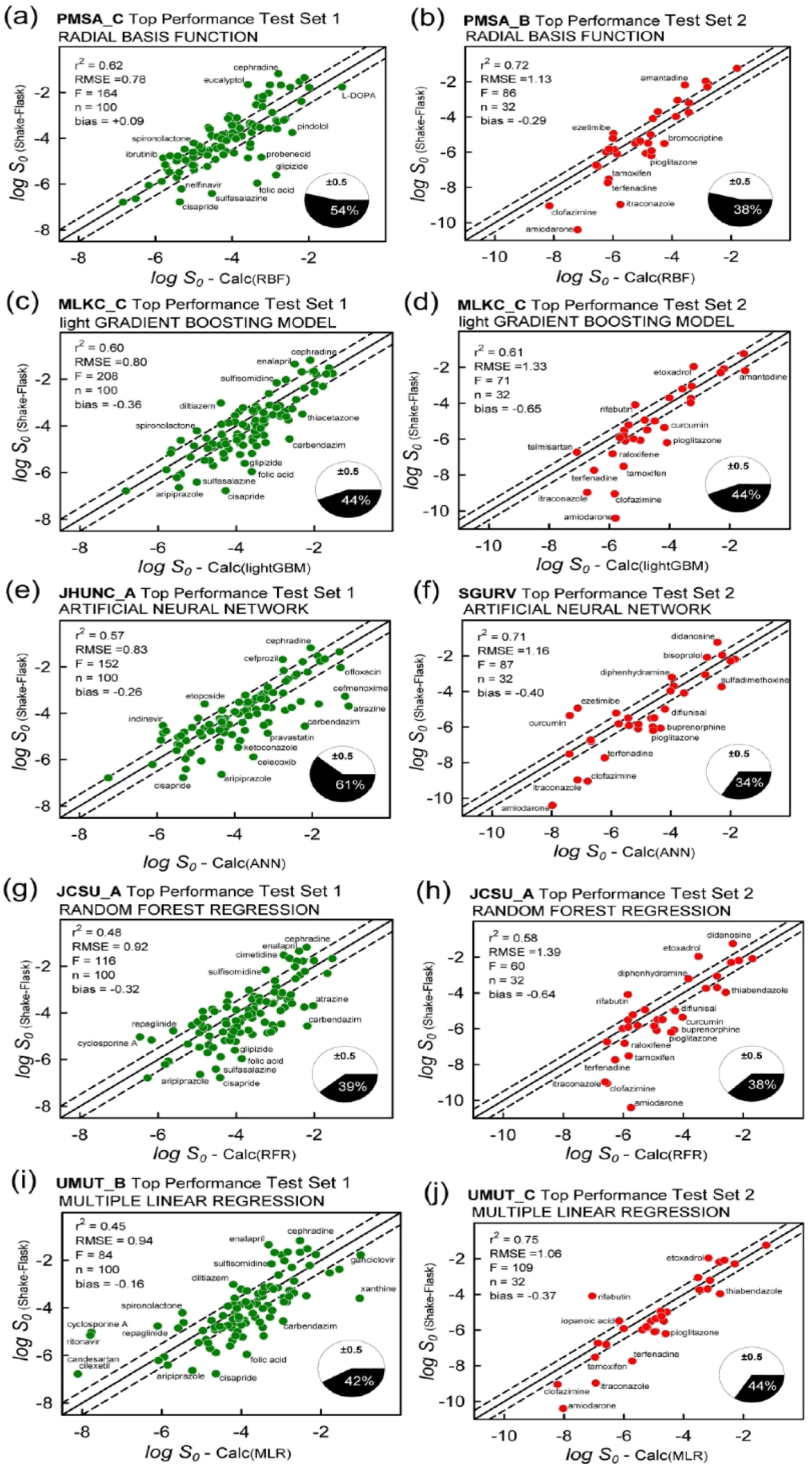


Figure 6. Log S_0 correlations for the top performers (based on RMSE) for five of the most representative methods (RBF, lightGBM, ANN, RFR, and MLR).

Finally, there is plenty of room for argument about how much scope for improvement there is in descriptors, regression and learning algorithms, and accuracy of experimental data. Open, objective, and transparent challenges to predict important physicochemical properties are needed as a way to evaluate the state of the art and progress of our computational capabilities. However, as in the first solubility challenge, it became clear that this is a time-consuming activity which requires careful selection of the data used and careful attention to how results are analyzed and scored. In carrying out this second solubility challenge, we confirmed (again) that bigger and better databases (with citations to the original literature) are needed, especially containing druglike molecules. Many of the datasets used traditionally for training the models are compiled from very old publications, not comprising very druglike molecules. Better measurement methods, sound experimental designed protocols following "good practices" and proper reporting of the measured data (with details about

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

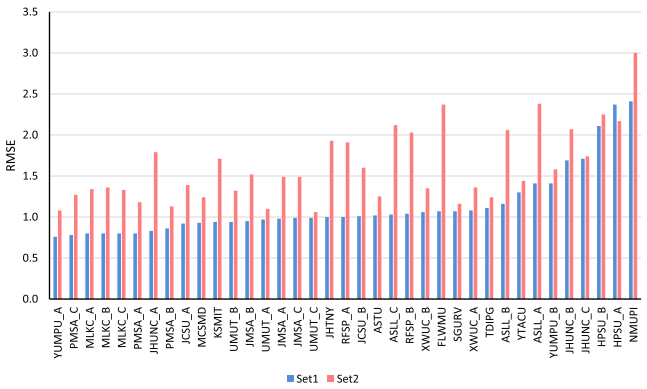


Figure 7. RMSE for each prediction showing both sets: Set 1 (blue) and Set 2 (red).

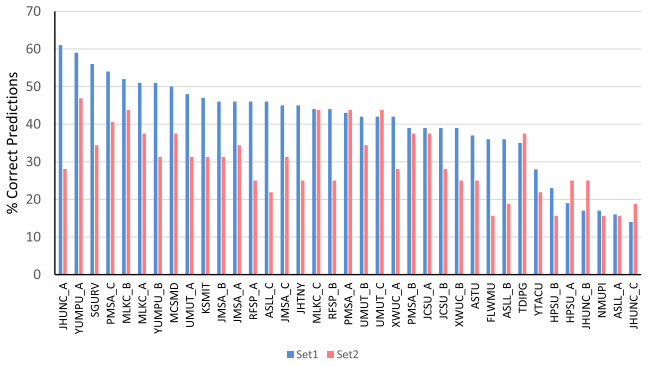


Figure 8. % Correct predictions (±0.5 log S_0) for each prediction showing both sets: Set 1 (blue) and Set 2 (red).

the experimental conditions), will increase the quality of the data sets and reduce variances.[11] Finally, better 3D descriptors, both in solution and describing the solid state, are needed.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00701.

SC-2 data and results (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Antonio Llinas − *DMPK, Research and Early Development, Respiratory & Immunology (R&I), BioPharmaceuticals R&D, AstraZeneca, Gothenburg SE 431 50, Sweden;* orcid.org/0000-0003-4620-9363; Email: Antonio.llinas@astrazeneca.com

Authors

Ioana Oprisiu − *Data Science & Artificial Intelligence, Imaging & Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg SE 431 50, Sweden*

Alex Avdeef − *in-ADME Research, New York, New York 10128, United States;* orcid.org/0000-0002-3139-5442

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00701

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Prof. John Mitchell (University of St. Andrews) for very helpful discussions and the Journal of Chemical Information and Modeling for its help and great assistance in setting up this second solubility challenge. We would like to thank, especially, Professor Kenneth M. Merz, Editor-in-Chief of this journal for his constant advice, support, and great patience.

■ DEDICATION

Dedicated to the memory of Anton J. Hopfinger and Oleg A. Raevsky.

■ REFERENCES

(1) Llinas, A.; Avdeef, A. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ∼ 0.17 log) and Loose (SD ∼ 0.62 log) Test Sets. *J. Chem. Inf. Model.* **2019**, *59*, 3036.

(2) Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289−1303.

(3) Hopfinger, A. J.; Esposito, E. X.; Llinás, A.; Glen, R. C.; Goodman, J. M. Findings of the challenge to predict aqueous solubility. *J. Chem. Inf. Model.* **2009**, *49*, 1−5.

(4) Stuart, M.; Box, K. Chasing equilibrium: Measuring the intrinsic solubility of weak acids and bases. *Anal. Chem.* **2005**, *77*, 983−990.

(5) Avdeef, A. Multi-lab intrinsic solubility measurement reproducibility in CheqSol and shake-flask methods. *ADMET & DMPK* **2019**, *7*, 210−219.

(6) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Model.* **1998**, *38*, 720−725.

(7) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **2002**, *54*, 355−366.

(8) Taskinen, J.; Norinder, U. In silico predictions of solubility. In *Comprehensive Medicinal Chemistry II*; Testa, B., van de Waterbeemd, H., Eds.; Elsevier: Oxford, U.K., 2007; pp. 627−648.

(9) Palmer, D. S.; Mitchell, J. B. O. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? *Mol. Pharm.* **2014**, *11*, 2962−2972.

(10) Avdeef, A. Suggested improvements for measurement of equilibrium solubility-pH of ionizable drugs. *ADMET & DMPK* **2015**, *3*, 84−109.

(11) Avdeef, A.; Fuguet, E.; Llinàs, A.; Ràfols, C.; Bosch, E.; Völgyi, G.; Verbić, T.; Boldyreva, E.; Takács-Novák, K. Equilibrium solubility measurement of ionizable drugs - consensus recommendations for improving data quality. *ADMET & DMPK* **2016**, *4*, 117−178.

(12) Birch, H.; Redman, A. D.; Letinski, D. J.; Lyon, D. Y.; Mayer, P. Determining the water solubility of difficult-to-test substances: A tutorial review. *Anal. Chim. Acta* **2019**, *1086*, 16.

(13) Ono, A.; Matsumura, N.; Kimoto, T.; Akiyama, Y.; Funaki, S.; Tamura, N.; Hayashi, S.; Kojima, Y.; Fushimi, M.; Sudaki, H.; Aihara, R.; Haruna, Y.; Jiko, M.; Iwasaki, M.; Fujita, T.; Sugano, K. Harmonizing solubility measurement to lower inter-laboratory variance - progress of consortium of biopharmaceutical tools (CoBiTo) in Japan. *ADMET & DMPK* **2019**, *7*, 183−195.

(14) Walters, W. P. What are our models really telling us? A practical tutorial on avoiding common mistakes when building predictive models. In *Chemoinformatics for Drug Discovery*; Bajorath, J., Ed.; John Wiley & Sons: Hoboken, NJ, 2014: pp. 1−31.

(15) Sun, H. *A Practical Guide to Rational Drug Design*; Elsevier: Amsterdam, 2015; pp. 193−223.

(16) Pirashvili, M.; Steinberg, L.; Belchi Guillamon, F.; Niranjan, M.; Frey, J. G.; Brodzki, J. Improved understanding of aqueous solubility modeling through topological data analysis. *J. Cheminf.* **2018**, *10*, 54.

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(17) Bergström, C. A. S.; Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int. J. Pharm.* **2018**, *540*, 185−193.
(18) Matos, G. D. R.; Mobley, D. L. Challenges in the use of atomistic simulations to predict solubilities of drug-like molecules. *F1000Research* **2019**, *7*, 686.
(19) Yalkowsky, S. H.; Dannenfelser, R. M. *Aquasol Database of Aqueous Solubility*, Version 5; College of Pharmacy, Univ. of Ariz: Tucson, AZ, 1992.
(20) Yalkowsky, S. H.; He, Y.; Jain, P. *Handbook of Aqueous Solubility Data*; CRC Press—Taylor & Francis Group: Boca Raton, FL, 2003.
(21) Yalkowsky, S. H.; He, Y.; Jain, P. *Handbook of Aqueous Solubility Data*, 2nd.; CRC Press—Taylor & Francis Group: Boca Raton, FL, 2010.
(22) Howard, P.; Meylan, W. *PHYSPROP DATABASE*; Syracuse Research Corp., Environmental Science Center: N. Syracuse, NY, 1999.
(23) Huuskonen, J. Estimation of aqueous Solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(24) Avdeef, A.; Berger, C. M.; Brownell, C. pH-metric solubility. 2. Correlation between the acid-base titration and the saturation shake-flask solubility-pH methods. *Pharm. Res.* **2000**, *17*, 85−89.
(25) Avdeef, A.; Berger, C. M. pH-metric solubility. 3. Dissolution titration template method for solubility determination. *Eur. J. Pharm. Sci.* **2001**, *14*, 281−291.
(26) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* **2002**, *19*, 182−188.
(27) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1477−1488.
(28) Bergström, C. A. S.; Luthman, K.; Artursson, P. Accuracy of calculated pH-dependent aqueous drug solubility. *Eur. J. Pharm. Sci.* **2004**, *22*, 387−398.
(29) Rytting, E.; Lentz, K. A.; Chen, X.-Q.; Qian, F.; Venkatesh, S. A Quantitative Structure-Property Relationship for Predicting Drug Solubility in PEG 400/Water Cosolvent Systems. *Pharm. Res.* **2004**, *21*, 237−244.
(30) Rytting, E.; Lentz, K. A.; Chen, X.-Q.; Qian, F.; Venkatesh, S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS J.* **2005**, *7*, E78−E105.
(31) Sköld, C.; Winiwarter, S.; Wernevik, J.; Bergström, F.; Engström, L.; Allen, R.; Box, K.; Comer, J.; Mole, J.; Hallberg, A.; Lennernäs, H.; Lundstedt, T.; Ungell, A.-L.; Karlén, A. Presentation of a structurally diverse and commercially available drug data set for correlation and benchmarking studies. *J. Med. Chem.* **2006**, *49*, 6660−6671.
(32) Wassvik, C. M.; Holmén, A. G.; Bergström, C. A. S.; Zamora, I.; Artursson, P. Contribution of solid-state properties to the aqueous solubility of drugs. *Eur. J. Pharm. Sci.* **2006**, *29*, 294−305.
(33) Boobier, S.; Osbourn, A.; Mitchell, J. B. O. Can human experts predict solubility better than computers? *J. Cheminf.* **2017**, *9*, 63.
(34) Baek, K.; Jeon, S. B.; Kim, B. K.; Kang, N. S. Method validation for equilibrium solubility and determination of temperature effect on the ionization constant and intrinsic solubility of drugs. *J. Pharm. Sci. Emerg. Drugs* **2018**, *6*, 1−6.
(35) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(36) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and log P. *J. Chem. Inf. Model.* **2008**, *48*, 220−232.
(37) Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(38) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.
(39) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption classification of oral drugs based on molecular surface properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(40) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(41) Shoghi, E.; Fuguet, E.; Bosch, E.; Ràfols, C. Solubility-pH profiles of some acidic, basic and amphoteric drugs. *Eur. J. Pharm. Sci.* **2013**, *48*, 291−300.
(42) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci. Data* **2019**, *6*, 143.
(43) Narasimham, Y. S. L.; Barhate, V. D. Kinetic and intrinsic solubility determination of some b-blockers and antidiabetics by potentiometry. *J. Pharm. Res.* **2011**, *4*, 532−536.
(44) Palmer, D. S.; Llinàs, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharm.* **2008**, *5*, 266−279.
(45) Shareef, A.; Angove, M. J.; Wells, J. D.; Johnson, B. B. Aqueous Solubilities of Estrone, 17β-Estradiol, 17α-Ethynylestradiol, and Bisphenol A. *J. Chem. Eng. Data* **2006**, *51*, 879−881.
(46) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355−1359.
(47) Box, K.; Comer, J. Using Measured pKa, LogP and Solubility to Investigate Supersaturation and Predict BCS Class. *Curr. Drug Metab.* **2008**, *9*, 869−878.
(48) Etherson, K.; Halbert, G.; Elliott, M. Determination of excipient based solubility increases using the CheqSol method. *Int. J. Pharm.* **2014**, *465*, 202−209.
(49) Fornells, E.; Fuguet, E.; Mañé, M.; Ruiz, R.; Box, K.; Bosch, E.; Ràfols, C. Effect of vinylpyrrolidone polymers on the solubility and supersaturation of drugs; a study using the CheqSol method. *Eur. J. Pharm. Sci.* **2018**, *117*, 227−235.
(50) Llinàs, A.; Burley, J. C.; Box, K. J.; Glen, R. C.; Goodman, J. M. Diclofenac Solubility: Independent Determination of the Intrinsic Solubility of Three Crystal Forms. *J. Med. Chem.* **2007**, *50*, 979−983.
(51) Schönherr, D.; Wollatz, U.; Haznar-Garbacz, D.; Hanke, U.; Box, K. J.; Taylor, R.; Ruiz, R.; Beato, S.; Becker, D.; Weitschies, W. Characterisation of selected active agents regarding pKa values, solubility concentrations and pH profiles by SiriusT3. *Eur. J. Pharm. Biopharm.* **2015**, *92*, 155−170.
(52) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(53) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.
(54) Faller, B.; Wohnsland, F. Physicochemical parameters as tools in drug discovery and lead optimization. In *Pharmacokinetic Optimization in Drug Research*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta, Zürich and Wiley-VCH: Weinheim, 2001; pp. 257−274.
(55) Bergström, C. A. S.; Avdeef, A. Perspectives in solubility measurement and interpretation. *ADMET & DMPK* **2019**, *7*, 88−105.
(56) De Stefano, C.; Lando, G.; Malegori, C.; Oliveri, P.; Sammartano, S. Prediction of water solubility and Setschenow coefficients by tree-based regression strategies. *J. Mol. Liq.* **2019**, *282*, 401−406.
(57) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: the accuracy of density functionals across a broad

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120476.
(58) Mobley, D. L.; Guthrie, J. P. FreeSolv: a database of experimental and calculated hydration free energies, with input files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711−720.
(59) Experimental Water Solubility from EPI Suite Database. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface (accessed Sep 2, 2020).
(60) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533−554.
(61) Online Chemical Database with Modeling Environment. https://ochem.eu (accessed Sep 2, 2020).
(62) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing learned molecular representations for property prediction. *J Chem Inf Model.* **2019**, *59*, 3370−3388.
(63) Zheng, S.; Jiang, M.; Zhao, C.; Zhu, R.; Hu, Z.; Xu, Y.; Lin, F. e-Bitter: bitterant prediction by the consensus voting from the machine-learning methods. *Front. Chem.* **2018**, *6*, 82.
(64) Yalkowsky, S. H.; Valvani, S. C. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **1980**, *69*, 912−922.
(65) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility: Methods of Estimation for Organic Compounds*; Marcel Dekker, Inc.: New York, 1992; p 142.
(66) Jain, P.; Yalkowsky, S. H. Prediction of aqueous solubility from SCRATCH. *Int. J. Pharm.* **2010**, *385*, 1−5.
(67) Landrum, G.; Lewis, R.; Palmer, A.; Stiefl, N.; Vulpetti, A. Making sure theres a "give" associated with the "take": Producing and using open-source software in big pharma. *J. Cheminf.* **2011**, *3*, 3.
(68) Avdeef, A. pH-metric solubility. 1. Solubility-pH profiles from Bjerrum plots. Gibbs buffer and pK_a in the solid state. *Pharm. Pharmacol. Commun.* **1998**, *4*, 165−178.
(69) Avdeef, A. *Absorption and Drug Development*, 2nd ed.; Wiley-Interscience: Hoboken NJ, 2012.
(70) Avdeef, A. Prediction of aqueous intrinsic solubility of druglike molecules using Random Forest regression trained with Wiki-pS0 database. *ADMET & DMPK* **2020**, *8*, 29−77.
(71) Mitchell, J. Three machine learning models for the 2019 Solubility Challenge. *ADMET & DMPK* **2020**, DOI: 10.5599/admet.835. In Press.

https://dx.doi.org/10.1021/acs.jcim.0c00701
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX