molecular pharmaceutics

ACS AuthorChoice

Brief Article

pubs.acs.org/molecularpharmaceutics

Terms of Use CC-BY

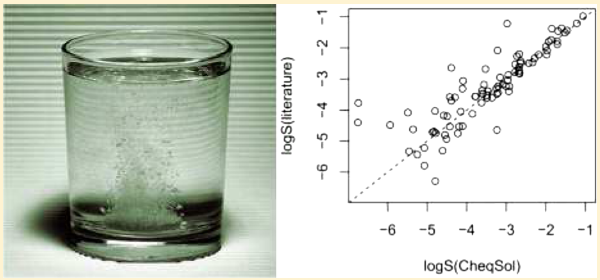
Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules?

David S. Palmer*,[†] and John B. O. Mitchell*,[‡]

[†]Department of Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland G1 1XL, U.K.

[‡]Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, University of St. Andrews, Purdie Building, North Haugh, St. Andrews, Scotland KY16 9ST, U.K.

S *Supporting Information*



ABSTRACT: We report the results of testing quantitative structure−property relationships (QSPR) that were trained upon the same druglike molecules but two different sets of solubility data: (i) data extracted from several different sources from the published literature, for which the experimental uncertainty is estimated to be 0.6−0.7 log S units (referred to mol/L); (ii) data measured by a single accurate experimental method (CheqSol), for which experimental uncertainty is typically <0.05 log S units. Contrary to what might be expected, the models derived from the CheqSol experimental data are not more accurate than those derived from the "noisy" literature data. The results suggest that, at the present time, it is the deficiency of QSPR methods (algorithms and/or descriptor sets), and not, as is commonly quoted, the uncertainty in the experimental measurements, which is the limiting factor in accurately predicting aqueous solubility for pharmaceutical molecules.

KEYWORDS: *solubility, bioavailability, QSPR, QSAR, druglike, ADME, Random Forest, dissolution, experimental error, CheqSol, Noyes−Whitney, Henderson−Hasselbalch, polymorph, crystal, machine learning, general solubility equation, ADMET, pharmaceutical, rule-of-five*

INTRODUCTION

Interest in the prediction of solubility by quantitative structure−property relationships (QSPRs) has risen dramatically in recent years.[1−5] Currently, the state-of-the-art tool allows the prediction of solubility with root-mean-square errors (RMSEs) of approximately 0.3−0.4 log units for simple organic molecules and 0.7−1.0 log units for drug molecules.[6] One frequently cited reason for the difficulty in predicting solubility for drug molecules is that published methods are derived from data taken from multiple sources from the literature, for which RMSEs in experimental data have been estimated to be 0.6−0.7 logS units.[6] The implicit assumption is that if existing models could be retrained and tested upon more accurate data, then the predictive error would decrease. Until recently, it has, however, not been possible to test this hypothesis, due to the absence of a definitive "gold-standard" data set containing accurate solubility data for fully characterized drug molecules.

There are many different definitions of aqueous solubility in common use in the published literature, but the majority of QSPR models have focused on the prediction of intrinsic aqueous solubility, which is the property we consider here. The intrinsic aqueous solubility of an ionizable molecule is defined as the concentration of the unionized molecule in saturated aqueous solution at thermodynamic equilibrium at a given temperature.[7,8] It is used to calculate dissolution rate and pH-dependent solubility in models such as the Noyes−Whitney equation[9] and the Henderson−Hasselbalch equation,[10,11] respectively. There has been great interest in predicting the intrinsic aqueous solubility of bioactive molecules in the biochemical sciences because it is a key determinant in the bioavailability of novel pharmaceuticals and the environmental fate of potential pollutants.[12−14]

Recently, the intrinsic aqueous solubilities of 132 drug molecules have been measured by the CheqSol method.[15,16] The data are highly reproducible and standard errors for ten repeat assays (i.e., the random errors between repititions of the same experiment) are typically around 0.05 log units (referred to mol/L). CheqSol measurements of solubility have been shown to be very reliable, with very good agreement reported for experiments carried out in different laboratories in different countries and continents. The results also agree very well with carefully performed shake-flask experiments carried out under the correct conditions (which are not always the conditions used for data reported in the published literature). The CheqSol data were originally published as part of a blind

Received: February 3, 2014
Revised: May 8, 2014
Accepted: June 11, 2014
Published: June 11, 2014

Table 1. Experimental Solubility Data Measured by the CheqSol Method (log S (exp)), and Experimental Solubility (log S (lit)), Melting Point (M.P.) and Octanol−Water Partition Coefficient Data (logP) Taken from the Published Literature; the Data Are Collated for 85 Druglike Molecules

molecule	logS (exp) (mol/L)	σ(S (exp)) (μ mol/L)	logS (lit) (mol/L)	M.P. (°C)	logP
acebutolol	−2.675	410	−2.20[27]	123[32]	1.71[32]
acetaminophen	−1.064	7000	−0.98[27,29]	170[32]	0.46[32]
acetazolamide	−2.435	80	−2.49[29]	260.5[32]	−0.26[32]
alprenolol	−2.634	40	−2.43[27]	109[32]	3.10[32]
amantadine	−1.854	1180	−1.38[33]	180[32]	2.44[32]
amitriptyline	−4.550	2.9	−4.80[27,29]	196.5[32]	4.92[32]
amoxicillin	−1.972	540	−2.09[27]	194[33]	0.87[32]
atropine	−2.004	460	−1.96[27,29]	118.5[32]	1.83[32]
azathioprine	−3.208	16	−3.44[34]	243.5[32]	0.10[32]
bendroflumethiazide	−4.298	6	−3.59[35]	223[32]	1.89[32]
benzocaine	−2.336	900	−2.47[27,29]	92[32]	1.86[32]
benzthiazide	−4.829	0.23	−4.69[33]	231.5[32]	1.73[32]
bupivacaine	−3.222	33	−2.08[32]	108[32]	3.41[32]
cephalothin	−2.938	3	−3.40[33]	160.5[32]	0.00[32]
chlorpromazine	−5.071	0.1	−5.22[27]	53[36]	5.41[32]
chlorpropamide	−3.249	10	−3.30[28]	128[32]	2.27[32]
chlorprothixene	−6.750	0.008	−4.40[37]	97.5[32]	5.18[32]
chlorzoxazone	−2.663	10	−2.83[29]	191.5[32]	1.66[33]
cimetidine	−1.692	480	−1.46[27,29,38]	142[32]	0.40[32]
ciprofloxacin	−3.597	27	−3.48[27,38]	263[39]	0.28[32]
clozapine	−3.238	2.14	−4.64[28]	184[32]	3.23[32]
desipramine	−3.627	8	−3.76[27]	25[40]	4.90[32]
diazoxide	−3.363	23	−3.60[28]	330.5[32]	1.20[32]
dibucaine	−4.390	5.8	−3.70[41]	64[42]	4.40[32]
diclofenac	−5.456	0.1	−5.33[27,29]	183[43]	4.51[32]
diethylstilbestrol	−4.429	11	−4.53[27,28]	170.5[32]	5.07[32]
diflunisal	−5.936	0.08	−4.48[29]	210.5[29]	4.44[32]
diltiazem	−3.159	45	−2.95[33]	212[32]	2.70[32]
5,5-diphenylhydantoin	−3.857	16	−4.12[27−29]	286[32]	2.47[32]
famotidine	−2.648	50	−2.49[27]	163.5[32]	−0.64[32]
flufenamic acid	−5.355	0.08	−4.62[29]	133.5[32]	5.25[32]
flurbiprofen	−4.152	0.4	−4.05[27,29]	111[32]	4.16[32]
folic acid	−5.247	0.45	−5.44[29]	250[32]	−2.80[44]
furosemide	−4.227	6	−4.75[27]	295[32]	2.03[32]
glipizide	−5.488	0.09	−4.08[33]	209[33]	1.91[32]
guanine	−4.432	9	−3.58[27,29]	360[32]	−0.94[45]
hexobarbital	−2.674	100	−2.73[32]	146.5[32]	1.98[32]
hydrochlorothiazide	−2.678	0.007	−2.61[27−29,38]	274[32]	−0.07[32]
hydroflumethiazide	−2.967	300	−3.04[29]	270.5[32]	0.36[32]
4-hydroxybenzoic acid	−1.464	670	−1.44[32]	214.5[32]	1.58[32]
ibuprofen	−3.595	16	−3.47[27−29]	76[32]	3.97[32]
imipramine	−4.105	3.5	−4.52[27]	174.5[32]	4.80[32]
indomethacin	−4.609	14	−5.31[27−29]	158[32]	4.16[46]
4-iodophenol	−1.714	600	−1.89[37]	93.5[32]	2.91[32]
ketoprofen	−3.209	16	−3.23[27−29,38]	94[32]	3.12[32]
lidocaine	−1.874	420	−1.74[27,29]	68.5[32]	2.44[32]
mefenamic acid	−6.738	0.001	−3.77[29]	231[32]	5.12[32]
metoclopramide	−3.565	16	−3.18[29]	147.25[32]	2.62[32]
metronidazole	−1.222	1500	−1.21[29]	160.5[32]	−0.02[32]
miconazole	−5.071	0.3	−5.79[27]	185[47]	6.02[33]
nalidixic acid	−3.611	5	−3.37[29]	229.5[32]	1.59[32]
naphthoic acid	−3.774	0.06	−3.56[32]	185.5[32]	3.28[32]
1-naphthol	−1.983	400	−2.22[32]	95[32]	2.85[32]
naproxen	−4.496	0.5	−4.20[27−29]	153[32]	3.18[32]
niflumic acid	−4.585	5	−4.17[41]	204[42]	4.43[32]
nitrofurantoin	−3.239	27	−3.48[29]	263[32]	−0.47[32]
oxytetracycline	−3.086	180	−3.17[32]	184.5[32]	−0.90[32]
phenazopyridine	−4.194	2	−4.53[27]	139[32]	2.80[44]
phenobarbital	−2.293	280	−2.34[27,29]	174[32]	1.47[32]

Table 1. continued

molecule	logS (exp) (mol/L)	σ(S (exp)) (μ mol/L)	logS (lit) (mol/L)	M.P. (°C)	logP
phenylbutazone	−4.391	0.04	−2.64[29]	105[32]	3.16[32]
phthalic acid	−1.606	860	−1.37[32]	230[32]	0.73[32]
pindolol	−3.788	4	−3.56[27,38]	171[32]	1.75[32]
piroxicam	−4.801	0.5	−4.03[28]	200[32]	3.06[32]
probenecid	−4.864	0.35	−4.71[27,28,38]	195[32]	3.21[32]
procaine	−1.719	820	−1.78[27]	61[32]	2.14[32]
propanolol	−3.495	16	−3.52[27,38]	96[32]	3.48[32]
pyrimethamine	−4.108	11	−3.31[33]	233.5[32]	2.69[32]
quinine	−2.786	36	−2.78[27,29]	195[32]	3.44[32]
salicylic acid	−1.931	290	−1.80[27,29]	158[32]	2.26[32]
sulfacetamide	−1.52	240	−1.51[29]	183[32]	−0.96[32]
sulfamerazine	−3.121	23	−2.98[32,33,48]	236[32]	0.14[32]
sulfamethazine	−2.732	23	−2.27[29]	198.5[32]	0.89[32]
sulfamethizole	−2.779	150	−2.53[37]	208[32]	0.54[32]
sulfathiazole	−2.688	180	−2.81[29]	189[32]	0.05[32]
sulindac	−4.509	4	−4.93[27−29]	183[32]	3.42[32]
tetracaine	−3.011	22	−3.23[37]	147[49]	3.51[32]
tetracycline	−2.924	190	−3.24[37]	172.5[32]	−1.30[32]
thiabendazole	−3.484	20	−3.60[37]	300[32]	2.47[32]
thymol	−2.186	38	−2.22[32]	51.5[32]	3.30[32]
tolbutamide	−3.463	18	−3.39[50]	128.5[32]	2.34[32]
tolmetin	−4.092	4	−3.06[33]	156[32]	2.79[32]
trichlomethiazide	−3.529	5	−2.68[32]	270[32]	0.62[32]
trimethoprim	−2.951	96	−2.87[28,29]	203[32]	0.91[32]
trimipramine	−4.796	3	−6.29[27]	45[32]	4.45[33]
warfarin	−4.783	0.03	−4.74[27]	161[32]	2.60[32]
minimum	−6.750		−6.29	25	−2.80
maximum	−1.064		−0.98	360	6.02
mean	−3.469		−3.31	175	2.24
σ	1.238		1.16	68	1.83

challenge to predict solubility.[15,16] Although this exercise provided a useful benchmark for the field, the results of the blind challenge were difficult to interpret in the scope of the questions addressed here since entrants were not asked to provide details of the methods they used or of any additional experimental data they employed in making the predictions.

Here, to test the hypothesis that experimental data quality is the limiting factor in predicting aqueous solubility, we develop QSPRs from the CheqSol solubility data set and compare these with models derived from literature solubility data for the same molecules. The models are derived by Random Forest regression, which has a number of attributes that make it useful for this purpose. First, a Random Forest model trained on the molecular descriptors used here has been shown to perform well in comparison to other published methods for the prediction of solubility.[17,18] Therefore, the models may be considered to be representative of QSPR models in the literature. Second, Random Forest is a well-defined algorithm for QSPR model building that does not overfit, which means that its use eliminates the influence that inconsistent modeling procedures would have on the experiment. To provide a comparison to the Random Forest model, we also implement a regression model against experimental melting point and logP, which may be considered to be a parametrized model based on the general solubility equation.[19]

The results are important as a benchmark of the prediction of solubility by QSPR methods but may also be considered to be illustrative of the effect that noise (in the dependent variable) has on QSPR models and what this means when attributing physical significance to QSPR variables. The work is also timely given the recent development of molecular simulation methods to predict the solubility of druglike molecules,[20,21] which, although currently more computationally expensive and less accurate than QSPR models, may in the future offer an alternative to QSPR models in some applications, depending on how both fields develop.[22−24]

■ METHODS

Data Sets. For each of the 85 drug molecules in the data set, two solubility values were obtained. First, the thermodynamic solubility of the nonionized form (intrinsic solubility) at 298 K was determined by the CheqSol method ("CheqSol Data set"). Second, an experimental intrinsic solubility value at 298 K was taken from the literature ("Literature Data set").

CheqSol Data Set. The intrinsic aqueous solubilities of 132 molecules were published by Llinas et al. as part of a recent blind challenge to predict solubility.[15,16] In the current work, we consider only a subset of this data because solubility data for only 85 of these molecules were found to be available from other sources in the published literature. The intrinsic aqueous solubility was measured for all 85 compounds by Llinas et al. using the CheqSol method.[16,25,26] For indomethacin, which was observed to hydrolyze in the original assay, we took the revised solubility data point reported by Comer et al.[26] In each CheqSol assay, intrinsic solubility was measured 10 or more times. In addition, the complete CheqSol assay was repeated multiple times starting from separate samples in different vials. The solubility or standard error of each molecule in the data set is reported as a statistical result across all separate CheqSol assays (Table 1). The standard errors for the measured intrinsic solubilities were typically less than 0.05 logS (referred to mol/L).

For five molecules in the data set, a change in polymorphic form was reported to occur during the solubility assay. In each case, the final polymorphic form was reported to be stable to repeated cycling between sub- and supersaturated states, and the solubility of this polymorph was used in the QSPR analyses. The ratios of the solubilities of the observed polymorphs are discussed in the results section.

Literature Solubility Data. Data for intrinsic solubility in water at 25 °C were obtained from the literature for each of the 85 molecules. For 26 of the 85 molecules, more than one value for solubility was found in the literature and an arithmetic average was taken. Of these molecules, three had a standard deviation between 0.5 and 1 log solubility units (diethylstilbestrol, indomethacin, and propanolol) and the remainder had standard deviations less than 0.5 log solubility units. Solubility data and references are provided in Table 1.

The literature solubility data set was compiled from 123 experimental solubility values taken from 12 sources from the literature. In general, the data are taken from well-known QSPR solubility data sets; the majority comes from papers on the prediction of solubility by Bergström et al. (35%),[27] Wassvik et al. (11%),[28] and Rytting (28%).[29] The remainder of the data are taken from other QSPR papers and well-known databases. Accurate measurement of intrinsic aqueous solubility requires that thermodynamic equilibrium is established and several factors are controlled, including purity of the solute and solvent, temperature, physical form of the precipitate, and solution pH and ionization state. Although care was taken to select intrinsic aqueous solubility from reliable sources in the literature (and those that are representative of good QSPR data sets), mistakes in experimental methodology and reporting of data have undoubtedly introduced some unidentified errors, which may mean that some of the literature data do not correspond perfectly to intrinsic aqueous solubility (they may be kinetic or total aqueous solubility). The problems with literature solubility are widely known and have been well discussed by many other authors.[30,31] Simple statistics of the distribution of both the literature and the new experimental data set are provided in Table 1. The literature and CheqSol solubility data are plotted against each other in Figure 1.

Melting Point Data and logP Data. For each molecule, experimental melting point and logP data were also obtained from the literature. All data and references are given in Table 1.

QSPR Models. Two models were built for both new experimental and literature solubility data: (i) a Random Forest model using all 2D and 3D descriptors; (ii) a multilinear regression equation using two variables: experimental melting point and experimental logP. The Random Forest model is similar to that used in a previous study on the prediction of solubility.[17] The multiple linear regression equation is a reparameterization of the general solubility equation and is included for comparison.

For each molecule, molecular structure files were taken from PubChem and were checked using SciFinder. A single low energy molecular conformer was selected by a low-mode conformation search using the MMFF94s force field with a Generalized Born Surface Area (GB/SA) model for water as solvent (as implemented in MacroModel).[51] Both 2D and 3D molecular descriptors were calculated from the lowest energy conformer using the Molecular Operating Environment (MOE) software.[52] The MOE software provides a method for predicting logS, which when compared to our new experimental solubility data for all 85 molecules gave a rather unsatisfactory, r^2 = 0.27, RMSE = 1.05 logS units, and bias = −0.36 logS units (where S is referred to units of mol/L). These predicted solubility values were excluded from further analysis and were not used in model building.

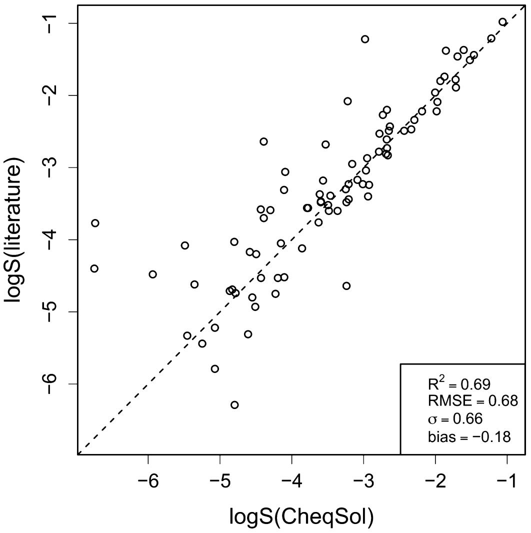


Figure 1. Correlation diagram for solubility data taken from the literature and as measured by the CheqSol method.

The 2D descriptors included calculated physical properties (logP, molar refractivity), charged surface properties (from Gasteiger−Marsili PEOE charge distributions on VDW surfaces), constitutional descriptors (counts of atoms and functional groups), connectivity and topological indices (including the chi, Kier−Hall, kappa, Wiener, and Balaban indices), and hydrogen bonding propensities (numbers of hydrogen bond donors and acceptors). The 3D descriptors included energy terms (total potential energy and contributions of angle bend, electrostatic, out of plane, and solvation terms to the molecular mechanics force-field energy), molecular shape descriptors (water accessible surface areas), volume descriptors, and surface area descriptors. These descriptors have previously been shown to be successful for the prediction of solubility.[53] All of the descriptors used in this work are listed in the Supporting Information. Each model was trained upon a subset of the data set (n = 60) and then validated by both cross-validation on the training data set and by prediction of the remaining 25 molecules. To reduce the influence of the choice of training and test split, this procedure was repeated 20 times (i.e., 20 different random partitions; a method sometimes referred to as Monte Carlo cross-validation), and statistically averaged results are reported. All comparisons were made on a like-for-like basis, i.e., the same randomly selected training and test sets and the same cross-validation folds were used for each pair of Random Forest or multilinear regression models trained upon literature solubility and new experimental data. Thus, the results for row 1 in Table 4 (or Table 7) and row 1 in Table 5 (or Table 8) are directly comparable, and likewise for rows 2, 3, 4, etc. Three statistics are reported to assess the accuracy of each regression model:

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y^i - y^i_{\mathrm{exp}})^2}{\sum_{i=1}^{n} (y^i_{\mathrm{exp}} - M(y^i_{\mathrm{exp}}))^2} \qquad (1)$$

$$\mathrm{RMSD}(y, y_{\mathrm{exp}}) = \sqrt{\frac{1}{N} \sum_{i} (y^i - y^i_{\mathrm{exp}})^2} \qquad (2)$$

$$\mathrm{bias} = M(y - y_{\mathrm{exp}}) = \frac{1}{N} \sum_{i \in S} (y^i - y^i_{\mathrm{exp}}) \qquad (3)$$

where index i runs through the set of N selected molecules, and y^i and y^i_{exp} are the calculated and the experimental values, respectively, for molecule i for the given property (i.e., $\log_{10}S$). A parentheses nomenclature is adopted to indicate whether the results refer to fit-to-the-training data (tr), 10-fold cross-validation (cv), out-of-bag validation (oob), or prediction of the test set (te).

Random Forest is robust to overfitting and can be used without optimization of the training parameters,[54] which allows the results for the two data sets to be compared without concern that the results are influenced by modeling errors. Each Random Forest was trained with the parameters of ntree = 500, nodesize = 5, and mtry = (total number of descriptors)/3, which have been used successfully in previous QSPR models to predict solubility.[17,18]

Comparison between Experimental and Literature Solubility Data. Figure 1 shows the correlation diagram between the log solubility data taken from the literature against the new experimental data (r^2 = 0.69, RMSE = 0.68 logS units, σ = 0.66 logS units, and *bias* = −0.18 logS units). Table 2 shows summary statistics for the distribution of log solubility data in these two data sets. The largest deviations between the two sources of solubility data are observed toward the bottom left of the graph (at low solubility values). Nine molecules have a difference in reported solubilities (ΔlogS = logS(exp) − logS(lit)) with a magnitude larger than 1 logS unit (ΔlogS given in parentheses): bupivacaine (−1.14), chlorprothixene (−2.35), clozapine (1.40), diflunisal (−1.46), glipizide (−1.41), mefenamic acid (−2.97), phenylbutazone (−1.75), tolmetin (−1.03), and trimipramine (1.49).

Table 2. Simple Statistics for Both the New Experimental Log Solubility Data (Referred to mol/L) and the Literature Log Solubility Data (Where Units of Solubility Are mol/L)

	σ	maximum	minimum	kurtosis	skew
new experimental data	1.24	−1.06	−6.75	−0.30	−0.31
literature data	1.16	−0.98	−6.29	−0.54	−0.17

Accurate measurement of intrinsic aqueous solubility requires the careful control of many factors, including purity of the solute and solvent, temperature, physical form of the precipitate, and solution pH and ionization state. In Figure 1, the predominance of molecules whose literature solubility values are higher than our measured intrinsic values suggests that some of these literature data points may correspond more closely to total aqueous solubilities than intrinsic solubilities (i.e., the ionized form of the molecule may contribute to the measured solubility). The literature solubility values that are lower than the measured intrinsic values may indicate that in the former a different crystalline form (e.g., a hydrate) was present at thermodynamic equilibrium. The inaccuracies in literature solubility caused by both experimental and reporting errors have been well discussed previously.[8,30] The CheqSol method is designed to measure specifically the intrinsic aqueous solubility. The crystalline form of the precipitates identified in the CheqSol experiments were characterized by thermogravimetric analysis, differential scanning calorimetry, and powder and single-crystal X-ray diffraction.

Polymorphism. The cycling between subsaturated, saturated and supersaturated states that is inherent to the CheqSol method was reported to cause a polymorphic change for five molecules during the solubility assay.[15,25] This polymorphic change was first evident as a change in solubility and was verified by repeating the assay in order to isolate and fingerprint the precipitate by powder X-ray diffraction (PXRD). The final polymorphic form was reported to be stable to repeated cycling between saturated and subsaturated states in each experiment. In Table 3, the ratios of the average solubilities of each observed polymorph to that of the least soluble polymorph for that molecule are reported. For each molecule, the solubility is observed to decrease as consecutive polymorphs are formed, which is in agreement with Ostwald's "Law of Stages" that states that metastable polymorphs are often observed to form prior to more thermodynamically stable ones.[55]

Table 3. Polymorph Solubility Ratios for Five Molecules That Were Observed to Undergo a Polymorphic Change during the CheqSol Assay[a]

molecule	form I	form II	form III	form IV
trazadone	3.6	1	n/a	n/a
trichlomethiazide	2.2	1	n/a	n/a
sulindac	6.7	1	n/a	n/a
phthalic acid	1.3	1	n/a	n/a
diflunisal	89	27	3.5	1

[a] The number represents the ratio relative to the least soluble polymorph for the given molecule. The polymorphs are named in the order that they appeared in the experiments, e.g., form I to form IV.

The largest ratio of the solubility of observed polymorphs is for diflunisal, for which form I is 89 times more soluble than form IV, a surprisingly large difference in solubility. Unfortunately, although the result for diflunisal was reproducible, it was not possible to solve the three-dimensional crystal structure from the PXRD patterns, which prevented further analysis of the structural differences between forms I and IV. The remaining four molecules in the data set have an average polymorph ratio of 3.45, which provides further evidence that diflunisal is an uncommon example. It has previously been reported, based on a statistical analysis of known crystal structures and existing solubility data, that the average difference between the solubility of polymorphs of druglike molecules is approximately 2-fold.[56] Here the observed average polymorph solubility ratio is 3.45; this may be an artifact caused by the use of a small data set (only five molecules with solubilities for more than one polymorph).

■ RESULTS

We have used two different methods, (i) Random Forest regression with a combination of 2D and 3D molecular descriptors and (ii) multilinear regression using only experimental melting point and logP values, to model the 85 molecule data set for both literature solubility data and new solubility data.

Table 4. Statistics for Random Forest Regression with Literature Solubility Data[a]

model	r^2(tr)	RMSE(tr)	bias(tr)	r^2(oob)	RMSE(oob)	bias(oob)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)	
1	0.91	0.36	−0.00	0.40	0.91	0.00	0.40	0.83	0.23	1.18	1.09	−3.42	−3.07	
2	0.89	0.36	0.01	0.32	0.90	0.02	0.59	0.82	0.16	1.11	1.31	−3.31	−3.33	
3	0.90	0.33	0.00	0.32	0.86	0.00	0.58	0.89	−0.09	1.05	1.41	−3.29	−3.38	a,c
4	0.92	0.34	0.01	0.47	0.88	0.01	0.37	0.78	−0.10	1.22	1.01	−3.23	−3.50	
5	0.92	0.31	0.00	0.47	0.78	0.01	0.45	0.96	0.24	1.08	1.33	−3.41	−3.08	e
6	0.92	0.32	0.00	0.45	0.81	0.01	0.43	0.98	0.03	1.10	1.32	−3.28	−3.39	a,e
7	0.92	0.31	0.00	0.47	0.81	0.00	0.46	0.88	0.18	1.12	1.22	−3.45	−2.98	a
8	0.92	0.32	0.01	0.49	0.82	0.02	0.43	0.87	0.06	1.16	1.17	−3.35	−3.21	a
9	0.93	0.31	−0.00	0.55	0.79	−0.01	0.08	1.01	−0.30	1.20	1.08	−3.25	−3.47	
10	0.92	0.31	−0.00	0.51	0.78	−0.01	0.43	0.95	−0.00	1.12	1.28	−3.31	−3.31	e
11	0.89	0.35	0.02	0.31	0.88	0.05	0.52	0.94	0.10	1.06	1.39	−3.31	−3.32	a
12	0.91	0.36	0.01	0.45	0.89	0.04	0.44	0.76	0.03	1.22	1.04	−3.30	−3.34	
13	0.90	0.35	0.00	0.37	0.88	0.01	0.48	0.91	−0.14	1.12	1.28	−3.29	−3.36	e
14	0.90	0.33	0.01	0.35	0.83	0.04	0.42	1.01	0.51	1.04	1.35	−3.48	−2.91	a,c,d
15	0.91	0.33	−0.00	0.38	0.86	−0.00	0.54	0.88	−0.02	1.10	1.32	−3.27	−3.41	e
16	0.92	0.32	0.01	0.49	0.81	0.02	0.38	0.92	−0.16	1.15	1.19	−3.24	−3.49	
17	0.91	0.35	0.01	0.40	0.88	0.03	0.43	0.89	0.53	1.14	1.20	−3.41	−3.09	
18	0.91	0.35	−0.01	0.44	0.89	−0.01	0.30	0.87	−0.03	1.20	1.07	−3.35	−3.22	
19	0.91	0.36	−0.00	0.41	0.90	−0.01	0.55	0.73	0.17	1.18	1.10	−3.22	−3.54	
20	0.91	0.35	−0.01	0.40	0.88	−0.01	0.51	0.83	0.13	1.15	1.21	−3.29	−3.38	c,e

[a] Extrapolation in logS for (a) acetaminophen, (b) miconazole, (c) metronidazole, (d) phthalic acid, and (e) trimipramine, in test set.

Table 5. Statistics for Random Forest Regression with New Experimental Solubility Data[a]

model	r^2(tr)	RMSE(tr)	bias(tr)	r^2(oob)	RMSE(oob)	bias(oob)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)	
1	0.92	0.37	0.01	0.48	0.94	0.03	0.22	0.85	0.46	1.30	0.99	−3.63	−3.09	
2	0.90	0.33	0.00	0.36	0.86	0.01	0.49	1.10	−0.12	1.08	1.56	−3.41	−3.62	c,d,f
3	0.91	0.35	0.01	0.43	0.88	0.03	0.59	0.88	−0.01	1.18	1.40	−3.45	−3.52	a,b,g
4	0.92	0.35	0.01	0.52	0.87	0.02	0.26	0.97	−0.27	1.26	1.15	−3.35	−3.76	c
5	0.92	0.36	0.01	0.48	0.88	0.03	0.27	1.02	0.50	1.23	1.21	−3.61	−3.13	
6	0.91	0.37	0.02	0.42	0.93	0.05	0.58	0.80	0.11	1.24	1.26	−3.45	−3.51	b
7	0.92	0.36	−0.00	0.50	0.88	0.01	0.48	0.83	0.12	1.25	1.17	−3.60	−3.16	b
8	0.93	0.32	0.02	0.54	0.82	0.04	0.39	1.02	−0.01	1.21	1.33	−3.50	−3.39	b,c
9	0.93	0.33	0.01	0.56	0.83	0.03	0.29	0.98	−0.35	1.25	1.19	−3.35	−3.74	
10	0.93	0.34	0.01	0.54	0.85	0.03	0.40	0.89	0.12	1.27	1.17	−3.51	−3.37	
11	0.91	0.37	0.01	0.39	0.95	0.01	0.61	0.78	0.36	1.22	1.28	−3.54	−3.31	b
12	0.92	0.38	0.02	0.49	0.96	0.06	0.52	0.63	0.06	1.35	0.93	−3.50	−3.40	
13	0.90	0.39	0.02	0.35	1.00	0.05	0.64	0.73	0.07	1.25	1.24	−3.48	−3.44	
14	0.89	0.38	0.02	0.31	0.94	0.07	0.50	0.95	0.48	1.15	1.37	−3.64	−3.06	b,g
15	0.90	0.38	0.01	0.39	0.94	0.03	0.59	0.83	−0.02	1.21	1.32	−3.43	−3.57	c
16	0.91	0.37	0.02	0.42	0.95	0.04	0.54	0.79	−0.11	1.27	1.19	−3.42	−3.58	
17	0.91	0.35	−0.00	0.43	0.86	0.00	0.58	0.92	0.24	1.16	1.44	−3.49	−3.43	c
18	0.93	0.36	0.02	0.54	0.90	0.05	0.15	0.87	0.14	1.34	0.96	−3.54	−3.30	
19	0.90	0.35	0.00	0.38	0.89	0.00	0.49	0.97	−0.17	1.13	1.38	−3.28	−3.91	c,d,e,f
20	0.92	0.36	0.01	0.47	0.91	0.02	0.53	0.81	0.17	1.26	1.20	−3.43	−3.55	

[a] Extrapolation in logS for (a) 4-hydroxybenzoic acid, (b) acetaminophen, (c) chlorprothixene, (d) diflunisal, (e) glipizide, (f) mefenamic acid, and (g) metronidazole, in test set.

Table 6. Average Statistics for Random Forest Regression with Both Literature and Experimental Solubility Data

data	r^2(tr)	RMSE(tr)	bias(tr)	r^2(oob)	RMSE(oob)	bias(oob)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)
experimental	0.91	0.36	0.01	0.45	0.90	0.03	0.45	0.88	0.09	1.23	1.24	−3.48	−3.44
literature	0.91	0.34	0.00	0.42	0.85	0.01	0.44	0.89	0.08	1.13	1.22	−3.32	−3.29

The results for each of the 20 Random Forest regression models derived from different random partitions of the data set are presented for literature solubility data in Table 4 and for new experimental data in Table 5. The statistically averaged results are presented in Table 6. For the literature solubility data set, the averaged values for out-of-bag cross-validation were r^2(oob) = 0.42, RMSE(oob) = 0.85 logS units, and bias(oob) = 0.01 logS units, and prediction of the molecules in the external test sets were r^2(te) = 0.44, RMSE(te) = 0.89 logS units, and bias(te) = 0.08 logS units. These results are comparable to recent global models for the prediction of solubility of drugs such as those of Bergström et al.[27] who

Table 7. Statistics for Multilinear Regression against Two Variables (Experimental Melting Point and Experimental logP) for Literature Solubility Data

model	$r^2(tr)$	RMSE(tr)	bias(tr)	r^2(cv)	RMSE(cv)	bias(cv)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)
1	0.22	0.99	0.00	0.11	1.06	−0.01	0.53	0.81	−0.00	1.13	1.20	−3.19	−3.62
2	0.41	0.88	0.07	0.34	0.93	−0.01	0.13	1.06	−0.18	1.16	1.16	−3.25	−3.46
3	0.40	0.87	0.10	0.37	0.89	−0.02	0.05	1.09	−0.24	1.13	1.14	−3.14	−3.73
4	0.53	0.74	−0.02	0.50	0.77	−0.00	−0.01	1.31	0.05	1.09	1.33	−3.38	−3.16
5	0.22	0.98	0.05	0.13	1.03	−0.01	0.49	0.84	−0.13	1.12	1.20	−3.17	−3.67
6	0.36	0.82	−0.03	0.31	0.86	−0.01	0.29	1.18	0.07	1.04	1.43	−3.38	−3.16
7	0.35	0.90	0.01	0.28	0.95	−0.01	0.30	1.04	−0.03	1.12	1.26	−3.34	−3.24
8	0.48	0.82	−0.01	0.47	0.83	−0.00	0.02	1.18	0.02	1.15	1.22	−3.29	−3.36
9	0.31	0.94	0.08	0.23	0.99	−0.01	0.34	0.94	−0.18	1.14	1.18	−3.20	−3.60
10	0.31	0.93	−0.01	0.23	0.99	0.00	0.37	0.96	0.01	1.14	1.24	−3.36	−3.21
11	0.31	1.01	−0.05	0.24	1.06	−0.01	0.38	0.76	0.12	1.23	0.99	−3.38	−3.15
12	0.34	0.93	0.08	0.25	0.99	−0.02	0.23	0.97	−0.19	1.16	1.13	−3.18	−3.63
13	0.30	1.03	0.06	0.17	1.13	0.01	0.39	0.67	−0.15	1.24	0.88	−3.19	−3.62
14	0.42	0.87	0.02	0.37	0.90	−0.02	0.15	1.09	−0.05	1.15	1.20	−3.27	−3.42
15	0.29	0.93	−0.06	0.17	1.01	−0.01	0.36	0.97	0.15	1.11	1.24	−3.44	−3.00
16	0.37	0.90	0.00	0.26	0.98	−0.00	0.26	1.03	−0.01	1.14	1.22	−3.31	−3.33
17	0.51	0.79	0.05	0.50	0.79	−0.01	−0.02	1.24	−0.12	1.13	1.25	−3.32	−3.30
18	0.23	1.03	−0.03	0.12	1.11	0.01	0.60	0.67	0.06	1.19	1.07	−3.41	−3.08
19	0.28	0.98	−0.05	0.21	1.03	−0.01	0.42	0.85	0.13	1.16	1.14	−3.42	−3.05
20	0.45	0.93	0.05	0.45	0.92	−0.01	−0.19	0.98	−0.11	1.25	0.91	−3.35	−3.22

Table 8. Statistics for Multilinear Regression against Two Variables (Experimental Melting Point and Experimental logP) for New Experimental Solubility Data

model	r^2(tr)	RMSE(tr)	bias(tr)	r^2(cv)	RMSE(cv)	bias(cv)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)
1	0.28	0.92	−0.01	0.21	0.97	−0.00	0.54	0.99	0.02	1.10	1.50	−3.33	−3.80
2	0.44	0.90	0.13	0.39	0.93	−0.00	0.31	1.05	−0.32	1.21	1.29	−3.35	−3.75
3	0.44	0.91	0.08	0.42	0.93	−0.02	0.24	1.03	−0.19	1.22	1.20	−3.30	−3.88
4	0.48	0.83	−0.02	0.42	0.88	−0.00	0.28	1.17	0.04	1.17	1.41	−3.54	−3.30
5	0.37	0.92	0.05	0.28	0.98	−0.02	0.41	1.00	−0.13	1.17	1.33	−3.30	−3.87
6	0.46	0.88	−0.05	0.41	0.92	−0.01	0.30	1.09	0.12	1.20	1.32	−3.56	−3.25
7	0.32	0.98	0.03	0.23	1.04	−0.01	0.58	0.86	−0.08	1.20	1.36	−3.49	−3.42
8	0.53	0.88	−0.04	0.54	0.88	−0.01	0.01	1.07	0.09	1.30	1.10	−3.47	−3.47
9	0.33	1.00	0.08	0.25	1.06	−0.01	0.56	0.79	−0.20	1.23	1.22	−3.34	−3.78
10	0.39	0.99	−0.05	0.30	1.06	0.01	0.45	0.83	0.13	1.27	1.15	−3.57	−3.24
11	0.41	1.00	−0.09	0.33	1.06	−0.02	0.37	0.79	0.22	1.31	1.02	−3.58	−3.19
12	0.31	0.95	0.14	0.23	1.00	−0.01	0.48	0.93	−0.34	1.15	1.32	−3.27	−3.96
13	0.39	0.98	0.08	0.28	1.07	0.01	0.38	0.84	−0.20	1.27	1.09	−3.31	−3.85
14	0.49	0.91	0.04	0.46	0.94	−0.01	0.12	1.02	−0.11	1.29	1.11	−3.40	−3.64
15	0.36	1.00	−0.09	0.24	1.09	−0.01	0.44	0.80	0.22	1.26	1.10	−3.64	−3.05
16	0.44	0.91	−0.01	0.36	0.98	−0.00	0.34	1.01	0.02	1.24	1.27	−3.48	−3.45
17	0.49	0.88	0.08	0.45	0.91	−0.02	0.23	1.09	−0.20	1.24	1.27	−3.46	−3.50
18	0.32	1.02	−0.02	0.18	1.13	0.01	0.62	0.72	0.05	1.25	1.19	−3.57	−3.22
19	0.37	1.03	−0.08	0.29	1.10	−0.01	0.50	0.69	0.20	1.31	0.99	−3.62	−3.11
20	0.50	0.88	0.14	0.47	0.90	−0.01	0.19	1.07	−0.34	1.25	1.22	−3.43	−3.57

Table 9. Average Statistics for Multilinear Regression against Two Variables (Experimental Melting Point and Experimental logP) for Both Literature and Experimental Solubility Data

data set	r^2(tr)	RMSE(tr)	bias(tr)	r^2(cv)	RMSE(cv)	bias(cv)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)
experimental	0.41	0.94	0.02	0.34	0.99	−0.01	0.37	0.94	−0.05	1.23	1.22	−3.45	−3.52
literature	0.35	0.91	0.02	0.29	0.96	−0.01	0.25	0.98	−0.04	1.15	1.17	−3.30	−3.35

reported RMSE(te) from 0.86 to 1.01 logS units and Chen et al.[34] who reported an RMSE(te) = 0.86 logS units, which suggests that the Random Forest model is performing as expected for the literature data set. When the model was retrained upon the new solubility data set, however, the expected improvement in results was not observed. For the new solubility data set, the averaged values for cross-validation were r^2(oob) = 0.45, RMSE(oob) = 0.90 logS units, and bias(oob) = 0.03 logS units, and prediction of the molecules in the external test sets were r^2(te) = 0.45, RMSE(te) = 0.88 logS units, and *bias*(te) = 0.09 logS units (Tables 5 and 6).

By comparison to the Random Forest models, the multilinear regression models against experimental values of melting point and logP were less accurate for the prediction of solubility. The results for all 20 models for literature and new experimental data are presented in Tables 7 and 8, respectively. The statistically averaged results are presented in Table 9. For the literature solubility data set, the averaged values for cross-validation were r^2(cv) = 0.29, RMSE(cv) = 0.96 logS units, and bias = −0.01 logS units, and for prediction of the molecules in the external test sets, r^2(te) = 0.25, RMSE(te) = 0.98 logS units, and bias = −0.04 logS units. When the multilinear regression models were retrained upon new experimental solubility data, the results for cross-validation were r^2(cv) = 0.34, RMSE(cv) = 0.99 logS units, and bias = −0.01 logS units, and for prediction of the molecules in the external test sets, r^2(te) = 0.37, RMSE(te) = 0.94 logS units, and bias = −0.05 logS units.

■ DISCUSSION

The observation that QSPR models for the prediction of solubility do not improve when trained and tested upon experimental data that are obtained under standardized conditions is surprising and is in disagreement with the assumptions made by other authors. Before this conclusion can be accepted, it is necessary to discuss potential sources of error in this experiment.

First, the work is only valid if the experimental data are more accurate than the literature data set. The average standard errors for the measured solubility data are typically <0.05 log solubility units, which is calculated as the standard deviation of multiple independent measurements made using the CheqSol method. The standard error is an indication of the reproducibility of the solubility measurements and does not preclude a systematic error. It is difficult to guarantee the accuracy of the data because there is no separate “gold-standard” set of solubility data with which to compare the results. However, the CheqSol results have proven to be consistent between different laboratories and in tests with carefully measured shake-flask results. Since QSPRs are empirical models, the presence of a systematic (rather than random) error would not necessarily imply that an accurate structure−property relationship could not be derived.

Second, the results may have been deleteriously affected when the random partitioning into training and test sets meant that the models had to extrapolate in logS. In Tables 5 and 8, the errors for the extrapolative predictions of the test set (marked with an a, b, or c) are observed to be of the same order as the nonextrapolative predictions; therefore, this possibility may be rejected.

Third, the regression model might not be optimal because the data set is small and diverse. Without measuring additional solubility data this is difficult to investigate, but a simple experiment was carried out to see whether the Random Forest out-of-bag cross-validation results would converge with data set size. Random Forest models were retrained on data sets of size 10, 20, 30, 40, 50, 60, 70, and 80 molecules, taken from the full data set (n = 85) with experimental solubility values. The selection of each data set was made at random, and as before, the results were averaged over 20 different random selections for each size of data set. A plot of average RMSE(oob) in logS units against the size of the data set is shown in Figure 2. The predictive accuracy is worst when the data set size is very small, as would be expected based upon the statistical averaging that is inherent to Random Forest and because of the general problems with working with small data sets. The average RMSE(oob) in logS units converges at around 60 molecules, which suggests that the size of the data set used in this work is acceptable. Furthermore, when leave-one-out cross-validation was carried out against the complete data set (n = 85) the results were similar for both the new experimental solubility data (q^2 = 0.51, RMSE = 0.86 logS units, and bias = 0.02 logS units) and the literature solubility data (q^2 = 0.49, RMSE = 0.82 logS units, and bias = 0.00 logS units), which suggest that the size of the data set is acceptable. It should be noted that the results illustrated in Figure 2 are dependent upon the diversity of the molecules that are selected (and the physical property of interest). Therefore, these results should not be interpreted as indicating that 60 molecules is an acceptable or minimum data set size for all QSPRs.

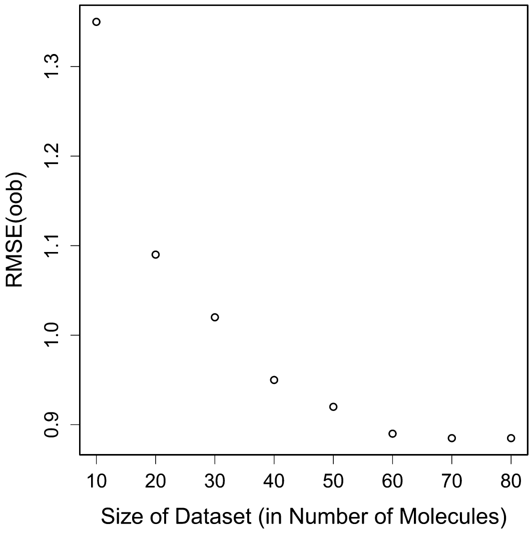


Figure 2. Average RMSE(oob) (in logS units) converges with data set size.

Fourth, QSPR models make a prediction of solubility from molecular structure (without knowledge of the crystal packing); therefore, it is reasonable to expect that the error might be due to the difference in solubility between polymorphs. However, the average difference in solubility between polymorphs has been measured to be 2-fold,[56] whereas the error in models to predict solubility is approximately 5- to 10-fold. Therefore, it would seem unlikely that this is the sole reason why the accuracy of QSPR models does not improve when trained upon accurate experimental data.

Our conclusion from this work is that experimental error in literature solubility data is not the limiting factor in predicting aqueous solubility. The predictive errors are similar for models constructed from both new solubility data and data extracted from the literature, even though the latter are known to contain experimental errors of the order of 0.6 to 0.7 log units. An interesting conclusion that can be drawn from this observation is that QSPR models are adept at modeling noise. This might suggest that caution should be exercised when attributing physical significance to QSPR variables.

The obvious question is, “how can QSPR models for the prediction of solubility be improved?” Various authors have demonstrated that the prediction of solubility of liquids and simple nondruglike organic molecules is possible with RMSEs of approximately 0.3 log solubility units.[4] Therefore, the answer may relate to the added complexity of modeling solubility for solid drug molecules. The solubility of a solid drug molecule depends upon the energy required for removing molecules from the crystal lattice as well as the energy gained by solvation.

Including these solid-state effects into models to predict solubility is unlikely to be completely successful based solely on single molecule properties because solid−solid interactions also depend upon the geometrical arrangement of molecules in the crystal lattice. The prediction of melting point for drug molecules exemplifies this problem, where the average RMSE for prediction is 40−50 °C,[18,57] even though the average experimental error is probably less than 5 °C. The most complete solution to this problem might involve the ab initio prediction of crystal structure, but despite recent advances, this remains a major challenge for drug molecules.[58] However, we note that it might not be necessary to know the correct polymorphic form, but only a plausible low energy polymorph, because average differences in solubility between polymorphs (2-fold) are considerably lower than average errors in QSPR models to predict solubility (5- to 10-fold).[59] This suggests that it might be possible to improve existing QSPR models by incorporating lattice energy terms calculated from a simulated or best-guess crystal form. Furthermore, separating crystal packing energy from solvation energy in QSPR models might permit a simpler linear regression model to be developed, which might alleviate the problems noted by Lipinski that solubility becomes more difficult to predict for complex molecules because it depends upon multiple intercorrelated factors.[60] In the long term, a more satisfactory solution to this problem might be to calculate solubility from first-principles using molecular simulation. Although there have been some significant recent advances in this field,[20,21,59,61] further development work is required before these methods become widely used in pharmaceutical research and development.[62,63]

Another consideration is that errors in the calculation of logP may contribute to errors in models to predict solubility. However, this conclusion is not supported by the results. First, it is observed that the error in the multilinear regression model is large despite being derived from experimental values of logP and melting point (this observation also implies that melting point and logP are not ideal descriptors to quantify the free energy changes associated with breaking cohesive interactions in the crystal and hydrating the free solute molecules; such contributions to the free energy of solution can be expressed more rigorously as sublimation and hydration free energies by a thermodynamic cycle via the gas-phase[21,59]). Second, when the Random Forest model was retrained using experimental logP values, rather than calculated logP values, the results did not improve.

■ CONCLUSIONS

We conclude from this work that existing QSPR methods for modeling solubility data do not improve when trained upon experimental data that is obtained under standardized conditions. Furthermore, QSPR models are adept at modeling noise, which suggests that caution should be exercised when attributing physical significance to QSPR variables. The results suggest that further work is required to develop novel QSPR methodologies that are more accurate and more reliable.

■ ASSOCIATED CONTENT

Supporting Information

List of all molecular descriptors used in this work and some further analysis. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*(D.S.P.) E-mail: david.palmer@strath.ac.uk.
*(J.B.O.M.) E-mail: jbom@st-andrews.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

D.S.P. is grateful for funding from the European Commission through a Marie Curie Intra-European Fellowship within the seventh European Community Framework Programme (FP7-PEOPLE-2010-IEF). D.S.P. thanks the University of Strathclyde for support through its Strategic Appointment and Investment Scheme. Computations were performed at the EPSRC funded ARCHIE-WeSt High Performance Computer (www.archie-west.ac.uk, EPSRC grant no. EP K000586l). J.B.O.M. thanks the Scottish Universities Life Sciences Alliance (SULSA) for financial support and EaStCHEM for access to the EaStCHEM Research Computing Facility.

■ REFERENCES

(1) van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(2) Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discovery* **2003**, *2*, 665−668.

(3) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.

(4) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems and solutions. *Curr. Med. Chem.* **2006**, *13*, 223−241.

(5) Salahinejad, M.; Le, T. C.; Winkler, D. A. Aqueous solubility prediction: Do crystal lattice interactions help? *Mol. Pharmaceutics* **2013**, *10*, 2757−2766.

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

(7) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; Oxford University Press, New York, 1999.

(8) Avdeef, A. *Absorption and Drug Development: Solubility, Permeability, and Charge State*; Wiley-Interscience, Hoboken, N. J., 2003.

(9) Noyes, A. A.; Whitney, W. R. The rate of solution of solid substances in their own solutions. *J. Am. Chem. Soc.* **1897**, *19*, 930−934.

(10) Hendersen, L. J. Concerning the relationship between the strength of acids and their capacity to preserve neutrality. *Am. J. Physiol.* **1908**, *21*, 173−179.

(11) Hasselbalch, K. A. Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebundenen Kohlensaure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl. *Biochemische Zeitschrift* **1917**, *78*, 112−144.

(12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.

(13) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(14) Walters, W. P.; Namchuk, M. Designing screens: How to make your hits a hit. *Nat. Rev. Drug Discovery* **2003**, *2*, 259−266.

(15) Llinas, 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.

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

(17) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.

(18) 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.

(19) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.

(20) Schnieders, M. J.; Baltrusaitis, J.; Shi, Y.; Chattree, G.; Zheng, L.; Yang, W.; Ren, P. The structure, thermodynamics, and solubility of organic crystals from simulation with a polarizable force field. *J. Chem. Theory Comput.* **2012**, 1721−1736.

(21) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V. First-principles calculation of the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Theory Comput.* **2012**, *8*, 3322−3337.

(22) Palmer, D. S.; Sergiievskyi, V. P.; Jensen, F.; Fedorov, M. V. Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *J. Chem. Phys.* **2010**, *133*, 044104.

(23) Palmer, D. S.; Chuev, G. N.; Ratkova, E. L.; Fedorov, M. V. In silico screening of bioactive and biomimetic solutes by integral equation theory. *Curr. Pharm. Des.* **2011**, *17*, 1695−1708.

(24) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Toward a Universal Model To Calculate the Solvation Thermodynamics of Druglike Molecules: The Importance of New Experimental Databases. *Mol. Pharmaceutics* **2011**, *8*, 1423−1429.

(25) Llinas, A.; Goodman, J. M. Polymorph control: past, present and future. *Drug Discovery Today* **2008**, *13*, 198−210.

(26) Comer, J.; Judge, S.; Matthews, D.; Towes, L.; Falcone, B.; Goodman, J.; Dearden, J. The intrinsic aqueous solubility of indomethacin. *ADMET DMPK* **2014**, *2*, 18−32.

(27) Bergstrom, 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) Wassvik, C. M.; Holmen, A. G.; Bergstrom, 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.

(29) 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.

(30) Pontolillo, J.; Eganhouse, R. P. *U.S. Geological Survey Water-Resources Investigations Report 01-4201*; U.S. Geological Survey: Reston, VA, 2001.

(31) Llinas, 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.

(32) US EPA. *EPISUITE* for Microsoft Windows, v 4.10; United States Environmental Protection Agency: Washington, DC, 2011.

(33) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.

(34) Chen, X. Q.; Cho, S. J.; Li, Y.; Venkatesh, S. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J. Pharm. Sci.* **2002**, *91*, 1838−1852.

(35) Ran, Y.; He, Y.; Yang, G.; Johnson, J. L.; Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere* **2002**, *48*, 487−509.

(36) Fabricius, I. L. *Aminoalkylation of Secondary Amines*; Novo Terapeutisk Laboratorium A/S, BE 620485 19630121, p 14, 1963.

(37) Yalkowsky, S. H.; He, Y. *The Handbook of Aqueous Solubility Data*; CRC Press LLC: Boca Raton, FL, 2003.

(38) Bergstrom, C.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* **2002**, *19*, 182−188.

(39) Sanchez, J. P.; Domagala, J. M.; Hagen, S. E.; Heifetz, C. L.; Hutt, M. P.; Nichols, J. B.; Trehan, A. K. Quinolone antibacterial agents. Synthesis and structure−activity relationships of 8-substituted quinoline-3-carboxylic acids and 1,8-naphthyridine-3-carboxylic acids. *J. Med. Chem.* **1988**, *31*, 983−991.

(40) Ran, Y.; Yalkowsky, S. H. Prediction of drug solubility by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(41) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.

(42) Milne, G. W. A. *Drugs: Synonyms and Properties*; Ashgate Publ. Co.: Brookfield, VT, 2000.

(43) Cambridge Structural Database; Cambridge Crystallographic Data Centre, 2007. See http:///www.ccdc.cam.ac.uk/ (accessed July 21, 2013).

(44) Moffat, A. C.; Osselton, M. D.; Widdop, B.; Watts, J. *Clarkes Analysis of Drugs and Poisons*; Pharmaceutical Press: London, 2007.

(45) Machatha, S. G.; Yalkowsky, S. H. Comparison of the octanol/water partition coefficients calculated by ClogP, ACDlogP and KowWin to experimentally determined values. *Int. J. Pharm.* **2005**, *294*, 185−192.

(46) Zhao, Y.; Jona, J.; Chow, D. T.; Rong, H.; Semin, D.; Xia, X.; Zanon, R.; Spancake, C.; Maliski, E. High-throughput logP measurement using parallel liquid chromatography/ultraviolet/mass spectrometry and sample-pooling. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 1548−1555.

(47) Yanez, E. C.; Sanchez, A. C.; Becerra, J. M. S.; Muchowski, J. M.; Almanza, R. C. Synthesis of Miconazole and Analogs Through a Carbenoid Intermediate. *Rev. Soc. Quim. Mex.* **2004**, *48*, 49−52.

(48) Delgado, D. R.; Martinez, F. Solubility and solution thermodynamics of sulfamerazine and sulfamethazine in some ethanol and water mixtures. *Fluid Phase Equilib.* **2013**, *360*, 88−96.

(49) Doser, H. Uber die Schmelzpunkte des Pantokains, Bromurals und Theophyllins. *Arch. Pharm.* **1943**, *281*, 251−256.

(50) 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.

(51) *Macromodel* v.9.1; Schrodinger: Portland, OR, 2006. See http://www.schrodinger.com/ (accessed July 21, 2013).

(52) *MOE*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2002. http://www.chemcomp.com (accessed July 21, 2013).

(53) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170−176.

(54) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(55) Threlfall, T. Structural and Thermodynamic Explanations of Ostwald's Rule. *Org. Process Res. Dev.* **2003**, *7*, 1017−1027.

(56) Pudipeddi, M.; Serajuddin, A. T. M. Trends in solubility of polymorphs. *J. Pharm. Sci.* **2005**, *94*, 929−939.

(57) O'Boyle, N. M.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Simultaneous feature selection and parameter optimization using an artificial ant colony: case study of melting point prediction. *Chem. Cent. J.* **2008**, *2*, 21.

(58) Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K. Towards crystal structure prediction of complex organic compounds: A report on the fifth blind test. *Acta Crystallogr., Sect. B: Struct. Sci.* **2011**, *67*, 535−551.

(59) Palmer, D. S.; Llinas, 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. Pharmaceutics* **2008**, *5*, 266−279.

(60) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(61) Luder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Persson, R.; Pedersen, M. In silico prediction of drug solubility: 4. Will simple potentials suffice? *J. Comput. Chem.* **2009**, *30*, 1859−1871.

(62) Frolov, A. I.; Ratkova, E. L.; Palmer, D. S.; Fedorov, M. V. Hydration thermodynamics using the reference interaction site model: Speed or accuracy? *J. Phys. Chem. B* **2011**, *115*, 6011−6022.

(63) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a universal method to calculate hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys.: Condens. Matter* **2010**, *22*, 492101.