JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

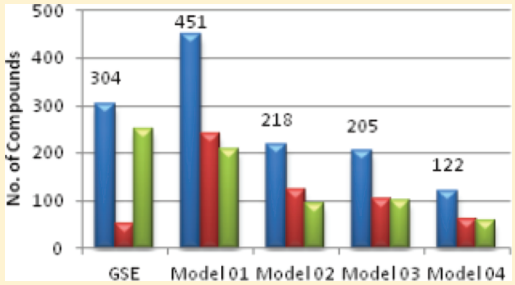
Revisiting the General Solubility Equation: *In Silico* Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area

Jogoth Ali,[†] Patrick Camilleri,[‡] Marc B. Brown,[†,§] Andrew J. Hutt,[†] and Stewart B. Kirton*[,†]

[†]School of Pharmacy, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, United Kingdom

[‡]Bio-Chemical Solutions, 5 Morgan Close, Stevenage, Hertfordshire, SG1 4TG, United Kingdom

ABSTRACT: The General Solubility Equation (GSE) is a QSPR model based on the melting point and log *P* of a chemical substance. It is used to predict the aqueous solubility of nonionizable chemical compounds. However, its reliance on experimentally derived descriptors, particularly melting point, limits its applicability to virtual compounds. The studies presented show that the GSE is able to predict, to within 1 log unit, the experimental aqueous solubility (log *S*) for 81% of the compounds in a data set of 1265 diverse chemical structures (−8.48 < log *S* < 1.58). However, the predictive ability of the GSE is reduced to 75% when applied to a subset of the data (1160 compounds −6.00 < log *S* < 0.00), which discounts those compounds occupying the sparsely populated regions of data space. This highlights how sparsely populated extremities of data sets can significantly skew results for linear regression-based models. Replacing the melting point descriptor of the GSE with a descriptor which accounts for topographical polar surface area (TPSA) produces a model of comparable quality to the GSE (the solubility of 81% of compounds in the full data set predicted accurately). As such, we propose an alternative simple model for predicting aqueous solubility which replaces the melting point descriptor of the GSE with TPSA and hence can be applied to virtual compounds. In addition, incorporating TPSA into the GSE in addition to log *P* and melting point gives a three descriptor model that improves accurate prediction of aqueous solubility over the GSE by 5.1% for the full and 6.6% for the reduced data set, respectively.



INTRODUCTION

Consideration of the aqueous solubility of compounds is important in numerous fields including the medicinal, physical, and environmental sciences. Aqueous solubility can be simply described as a measure of the dissolution of an un-ionized substance in water, and from the perspective of the pharmaceutical industry, the accurate measurement/prediction of the aqueous solubility of chemical compounds is very useful at several stages of the drug discovery process.[1] In conjunction with intestinal permeability, aqueous solubility may also be used to categorize the intestinal absorption of potential drugs using the Biopharmaceutics Classification System. Aqueous solubility must also be considered by scientists designing and synthesizing agrochemicals such as pesticides, where the agricultural benefits of the chemicals being used must be considered, for example, alongside the environmental impact of such chemicals being washed into waterways. Experimental determination of aqueous solubility, denoted as *S*, can be measured in moles per liter and conventionally converted to log values. Experimental determination is the preferred method of ascertaining aqueous solubility, as it offers the most accurate values. However the amount of time and significant cost is prohibitive when considering large libraries of compounds,[2] and hence accurate, reliable, and robust alternatives, such as in silico models for predicting aqueous solubility, are required.[3]

In recent years, there have been numerous models proposed for predicting aqueous solubility, the majority of which seek to exploit the relationship between chemical structure and aqueous solubility via quantitative structure−property relationships (QSPRs).[4,5] The importance in the development of a valid QSPR model is the avoidance of numerous types of errors,[6] especially in terms of data accuracy and selection. As such, the most important aspect of QSPR modeling is the selection, critical appraisal, and exploitation of appropriate data sets which contain only the highest quality experimental data.

The development of a variety of in silico models of aqueous solubility, including both QSPR-based and non-QSPR models, has been widely reported.[7−10] Arguably, the most well-known of these models is the general solubility equation (GSE) proposed by Yalkowsky and Jain.[11]

The GSE (eq 1) is a simple QSPR model which uses only two descriptors: a modified melting point (°C) (m.p. °C− 25) and the octanol−water partition coefficient of the un-ionized molecule (log *P*). The predictive power of the equation is impressive (coefficient of determination (r^2) = 0.96 and root-mean-square error (RMSE) = 0.53 in a data set of 1026 organic compounds[12]), and the simplicity of the model means it has

Received: August 17, 2011
Published: December 24, 2011

been widely adopted for use by the pharmaceutical industry.

$$\log S = 0.5 - 0.01(\text{m.p. } ^\circ\text{C} - 25) - \log P \quad (1)$$

A problem arises in relation to the GSE for compounds where experimental melting points have not been determined (e.g., virtual compounds). In addition, predicting values for large libraries of compounds, where the uniform determination of experimental melting points to a rigorously high standard cannot be guaranteed, may result in calculation errors. Therefore, models of aqueous solubility based on alternate descriptors which can be calculated unambiguously, e.g., from chemical structure, need to be identified.

The reliance of the GSE on experimental descriptors has been revisited in recent work by Yalkowsky and Jain, who have developed an alternative model for the prediction of aqueous solubility, SCRATCH. SCRATCH uses only two descriptors:[13] the molar aqueous activity coefficient (γ_w) and a complex algorithm which predicts melting point, rather than relying on a measured value. Predictive results from SCRATCH ($r^2 = 0.956$, RMSE = 0.859) for a data set of 883 compounds are comparable to those for the GSE.

However, despite their simplicity, neither the GSE nor SCRATCH explicitly accounts for the effect of polar and polarizable atoms on aqueous solubility. The aqueous solubility of a drug has been shown to affect its molecular absorption and transport. The molecular polar surface area of a drug has also been shown to affect molecular transport, and models have been created for the prediction of intestinal absorption and blood brain barrier penetration based on this observation.[14,15] Given these similarities, it is not unreasonable to hypothesize that polar surface area may affect aqueous solubility. The following studies investigate modifications to the GSE using a simple, time effective descriptor for calculating polar surface area (TPSA) and its effects on the predictive ability of the model.

METHODS

Data Set. A data set of pharmaceutical, agrochemical, and general organic compounds was curated using Molecular Operating Environment (MOE)[16] in conjunction with the Web-based ChemSpider[17] and Reaxys[18] software. The data set consists of 1256 structurally unique compounds collected from a range of reliable literature sources. Comparisons of canonical SMILES[19] strings ensured that all compound structures were unique.

For the un-ionized form of each compound, an aqueous solubility value (log *S*), an octanol−water partition coefficient (log *P*), and an experimental melting point (m.p. °C) value were identified. The log *P* values gathered from the literature were validated using the Web site Virtual Computational Chemistry Laboratory.[20] This Web site returns an experimental log *P* value and a number of calculated log *P* values from a range of different models to give an "average" log *P*. Any compound which had predicted log *P* values which were significantly different (±1 log unit) from the experimentally determined log *P* cited in the literature were removed from the data set to reduce possible erroneous experimentally determined values. As such, all log *P* values identified for the data set were experimental values which were subsequently shown to be accurately predicted using the suite of in silico log *P* algorithms available via the Virtual Computational Chemistry Laboratory. This reduces the likelihood of an erroneous prediction for aqueous solubility based upon a calculated log *P* value for virtual compounds with similar chemical structures to those represented in the data set. TPSA was calculated for each compound in MOE using the fast method based on a sum of fragment-based contributions.[21]

Defining Training and Test Sets—Diverse Subset Algorithm and Fingerprint Analysis. The full data set consists of 1256 unique compounds with a range of log *S* values from −8.48 to +1.58 (mean −3.20, standard deviation 1.70). The data set was separated into a training set of 1004 (79.9% of total data set) and a test set of 252 (20.1% of total data set) compounds, using MACCS Structural Keys[16] and the diverse subset tool within MOE.

The Reduced Data Set. Sparsely populated regions of a data set can skew regression-based analyses and artificially inflate or depress the predictive ability of models generated by these techniques. As such, a reduced data set concentrating on the region of the data populated by the majority of compounds was constructed. This reduced data set was used as a test of model robustness for all equations generated, and for comparison with model performance on the full data set. In order to generate the reduced data set, compounds in sparsely populated regions of the data space (defined as compounds outside of the range $-6.00 < \log S < 0.00$) were discounted, resulting in a reduced data set of 1160 compounds (92.4% of the full data set).

Applying GSE to Full and Reduced Data sets. The GSE (eq 1) was applied to the full and reduced data sets. Linear regression analysis was used to correlate the measured log *S* values against the GSE-predicted value and produce corresponding r^2 and RMSE values for both data sets.

Model Creation and Validation. A full factorial design of QSPR models was constructed using combinations of the molecular descriptors log *P*, melting point, and TPSA in conjunction with the training set of 1004 compounds and MOE—specifically its QuaSAR-Model function with the partial least-squares (PLS) regression method.[16] No limit was placed on the degree of the fit for these investigations.[22] The maximum number of conditions before the model finds the perfect fit was set at 10^5 for each iteration, and overall quality of significant models was assessed by considering a number of respective r^2 and RMSE values, i.e., leave-one-out cross-validated training set analysis,[23] predictive ability of models with regard to test set compounds, and overall performance of the equations when considering both full and reduced data sets in their entirety.

Definition of Poor Predictions. Poor predictions of a compound were defined as instances where a model failed to predict the aqueous solubility of a compound to within one log unit of the reported experimentally determined value. This is consistent with previous models in the field.[7]

RESULTS AND DISCUSSION

Defining Training and Test Sets—Fingerprint Analysis and Diverse Subset Algorithm. The importance of data set selection is highlighted in a number of studies[6] with the overall indication that "whilst selection of the training set of chemicals for a QSAR should be based on diversity, the selection of the chemicals for validation should be based on representivity."[24] Initially, the creation of the training and test sets was carried out using a diverse subset tool within MOE to randomly select 252 test set compounds using log *S* as the discriminating variable. However, analysis of the data sets generated showed that, despite the fact that the mean log *S*

dx.doi.org/10.1021/ci200387c | J. Chem. Inf. Model. 2012, 52, 420−428

Table 1. Summary of Regression Statistics after Selection of Test (252 Compounds) and Training (1004 Compounds) Sets Using log *S* and MACCS Structural Keys As the Discriminating Variable[a]

	log *S*			MACCS Structural Keys	
data set	full	training	test	training	test
mean	−3.2048	−3.1963	−3.2386	−3.2892	−2.8686
standard deviation	1.7095	1.4738	2.4353	1.7007	1.7062
sample variance	2.9223	2.172	0 5.9307	2.8925	2.9110
range	10.0600	9.2300	10.0600	10.0400	9.5900
minimum	−8.4800	−8.4600	−8.4800	−8.4800	−8.4800
maximum	1.5800	0.7700	1.5800	1.5800	1.1100
count	1256	1004	252	1004	252

[a]The statistical data for the full data set (1256 compounds) are provided for comparison purposes.

values in both test and training sets were similar and comparable to that of the data set as a whole (Table 1), the test set had a much greater standard deviation with respect to log *S* values when compared with both the training and full data sets. This implied that the test set was not representative of either the training or full data sets, which could bias the results, making the model appear better or worse than it actually was. This inflated standard deviation for the test set can be attributed to the fact that the diverse subset tool selected a relatively large number of compounds in the sparsely populated regions of the data set (i.e., those with log *S* values <−6.00 or >0.00) for the test set, highlighting that log *S* was an inappropriate descriptor for use with the diverse subset tool.

As such, an alternative discriminating variable was required to ensure a representative distribution of compounds in both the test and training sets. MACCS Structural Keys proved to be a good descriptor for this purpose (Figure 1). On the basis of the

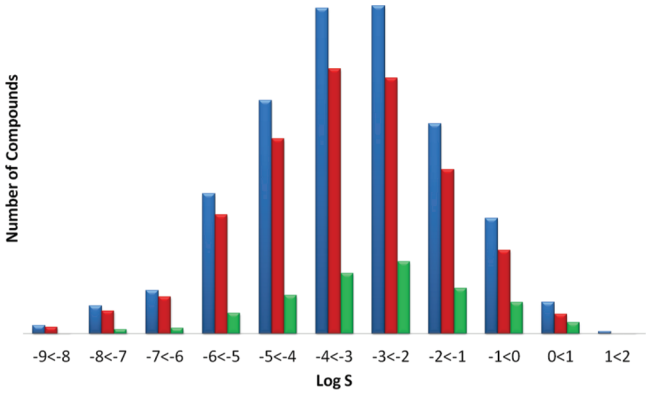


Figure 1. Distribution of compounds with respect to log *S* in the Full (blue), training (red), and test (green) sets selected by the diverse subset tool in conjunction with MACCS Structural Keys, illustrating the similarity of the distributions.

MACCS fingerprints for each compound, the diverse subset tool was used to identify a 1004 compound training set and a 252 compound test set. When compared to the previous distribution using log *S*, a more representative distribution of the molecules was seen for both test and training sets (Figure 1). It should be noted that the diverse subset tool ensures that the compounds in both the training and test sets are representative of the structural diversity in the data set as a whole, as well as ensuring that an appropriate distribution of log *S* values is considered in both training and test sets.

Applying the GSE to the Full and Reduced Data Sets. The GSE (eq 1) was applied to the full and reduced data sets, and linear regression was used to compare the measured log *S* values with the values predicted by the GSE (Figure 2).

The results from applying the GSE (eq 1) to the full and reduced data sets (Figure 2) suggest that the sparsely populated regions of the data set actually result in inflation of the predictive ability of the GSE by 6.4%. The predictive ability of the GSE for both data sets is considerably lower than that quoted in the literature,[12] but this is attributable to the fact that the data sets examined are different in each case.

Compounds Poorly Predicted by the GSE. Overall, the GSE predicted the aqueous solubility (logs) of 952 (75.8%) compounds in the full data set to within ±1 log unit of their experimentally determined values, the prediction of the remaining 304 (24.2%) compounds being outside this range.

An important observation when examining the poorly predicted compounds was the number of failures for compounds containing polar/polarizable atoms, e.g., nitrogen, oxygen, sulfur, and phosphorus atoms. A total of 1167 compounds of the full 1256 data set contain polar/polarizable atoms. A total of 76.9% of these polar/polarizable compounds are predicted within ±1 log unit of their experimentally determined values. Of the 304 compounds that were poorly predicted by the GSE, 269 (88.5%) contained at least one polar or polarizable atom. As such, it was of interest to determine whether or not improvements over the GSE predictions could be made by simply incorporating a term which accounted for the number of polar/polarizable atoms in a molecule. To this end, investigations into the effects of incorporating a term for topographical polar surface area (TPSA) on the predictive ability of the GSE were carried out.

Model Construction and Validation. Four significant models (Table 2) were designed using different combinations of the log *P*, melting point, and TPSA descriptors in conjunction with PLS regression. Each model was constructed using the training set of 1004 compounds and was validated using internal cross validation and by applying the models to the test set of 252 compounds.

Model 01. Model 01 explores the use of the log *P* component only as a descriptor for predicting aqueous solubility. Log *P* has been shown in previous studies to have a good correlation with the experimentally determined log *S*.[11,25]

$$\log S = -0.7897 \log P - 1.3674 \quad (2)$$

The equation (eq 2) produced from model 01 yielded values of r^2 = 0.632 and RMSE = 1.032 for the training set, and similar cross-validated statistics (r^2 = 0.63, RMSE = 1.034) are indicative that the model produced is robust. The predictive

dx.doi.org/10.1021/ci200387c | J. Chem. Inf. Model. 2012, 52, 420−428

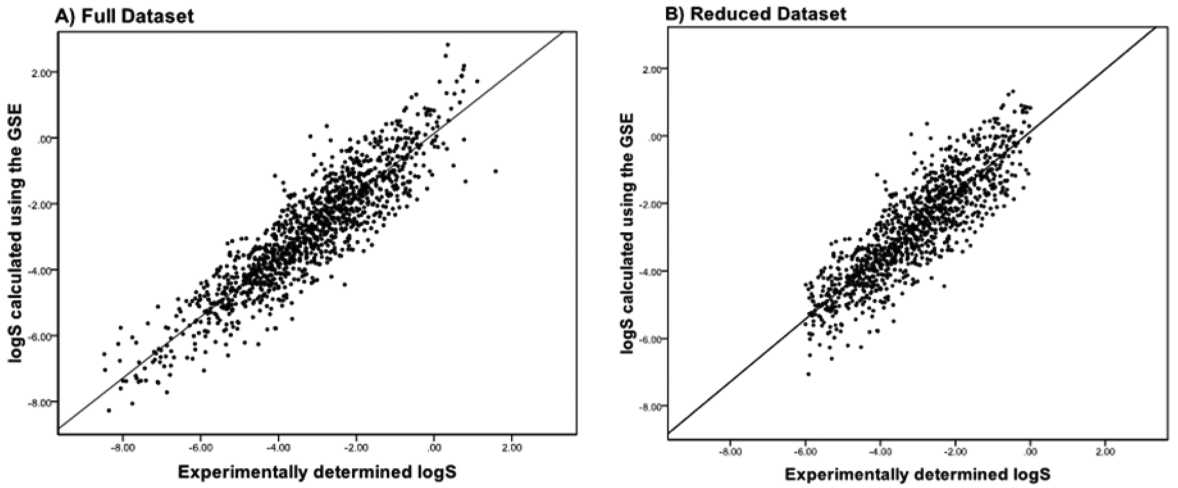


Figure 2. Correlation of predicted versus observed aqueous solubilities after applying the GSE to (A) the full data set (1256 compounds, r^2 = 0.816, RMSE = 0.719) and (B) the reduced data set (1160 compound, r^2 = 0.752, RMSE = 0.691).

Table 2. Summary of Regression Statistics for Models 01−04 Showing Results of Model Construction and Leave-One-out Cross-Validation (CV) Analysis for the Training Set and the Predictive Quality of Each Model When Applied to the Test Set

model	description	training set r^2	RMSE	CV r^2	CV RMSE	test set r^2	RMSE
01	log *P*	0.6315	1.0319	0.6300	1.0340	0.6982	0.9809
02	log *P*, m.p. − 25	0.8144	0.7324	0.8132	0.7348	0.8252	0.7123
03	log *P*, TPSA	0.8060	0.7488	0.8047	0.7513	0.8319	0.6986
04	log *P*, m.p. − 25,TPSA	0.8663	0.6215	0.8651	0.6242	0.8734	0.6105

ability of model 01 is relatively poor (r^2 = 0.698 for test set compounds). A comparison of statistical data from the training sets with the test sets shows an increase in r^2 and a decrease in RMSE, but this is believed to be a consequence associated with a single descriptor model.

Model 02. Model 02 uses the same two components as the GSE: log *P* and (m.p. °C − 25) and generated values of r^2 = 0.814 and RMSE = 0.732 for the training set, with similar values from cross-validation studies supporting the premise that the model is robust (r^2 = 0.813, RMSE = 0.735). The ability of model 02 to accurately predict the aqueous solubility of the compounds in the test set is good (r^2 = 0.825, RMSE = 0.712) and comparable with the results for the literature GSE with respect to the full data set (r^2 =0.816).

$$\log S = -0.8877 \log P - 0.0083(\text{m.p.} - 25) - 0.3007 \quad (3)$$

However, it is obvious that that the coefficients for the descriptors in the resultant equation (eq 3) are different from those of the original GSE (eq 1). This is not unexpected, given the data set used to derive model 02 is likely to be different from the data set used to generate the original GSE. However, without creating this "*in situ*" variation of the GSE model complete with its modified coefficients, it would have been impossible to compare predictive results from a GSE-type model to the results from any modified models proposed by these investigations. As such, model 02 establishes a comparable benchmark for other models.

Model 03. Model 03 (eq 4) was produced using log *P* and TPSA as the two components (eq 4) generating r^2 = 0.806 and RMSE = 0.749 for the initial model. The regression statistics obtained from the cross-validation studies support the fact that the model is robust (r^2 = 0.805, RMSE = 0.751), and the test set prediction (r^2 = 0.832, RMSE = 0.699) demonstrates that model 03 is as predictive as model 02 and the literature variation of the GSE.

$$\log S = -1.0377 \log P - 0.0210\text{TPSA} + 0.4488 \quad (4)$$

The results were essentially identical for model 03 when compared with model 02, suggesting, for our data set at least, that the replacement of an experimentally determined melting point as a descriptor with calculated TPSA would result in a model of equal predictive ability. This provides a robust and predictive alternative to the GSE which can be used for novel and virtual compounds where the melting point is unknown.

Model 04. Model 04 comprises all three of the descriptors, log *P*, TPSA, and melting point as components (eq 5) and generates r^2 = 0.866 and RMSE = 0.622 in the first instance. The model is robust, as evidenced by the fact that the cross-validated regression statistics have similar values when compared to the initial results (r^2 = 0.865, RMSE = 0.624), and of all the models generated, model 04 gives the best results when used to predict the aqueous solubilities of the test set molecules (r^2 = 0.873, RMSE = 0.611). This lends credence to the hypothesis that polar/polarizable atoms in a molecule must be accounted for when determining the aqueous solubility, and including a descriptor in the equation to account for this adds something beyond the implicit treatment of polarity/polarizability by log *P*.

$$\log S = -1.0144 \log P - 0.0056(\text{m.p.} - 25) - 0.0134\text{TPSA} + 0.5134 \quad (5)$$

Relative Importance of Descriptors. The relative importance of descriptors in every model was calculated by normalizing the respective coefficients in each equation (Table 3). The data indicate that for all models, log *P* is the most important descriptor, although all descriptors in each of the models are significant.

dx.doi.org/10.1021/ci200387c | *J. Chem. Inf. Model.* 2012, 52, 420−428

Table 3. Relative Importance of Descriptors As Described by Normalization of Equation Coefficients for Models 01−04

model	description	normalized coefficients		
		log *P*	m.p. − 25	TPSA
01	log *P*	1.0000		
02	log *P*, m.p. − 25	1.0000	0.4913	
03	log *P*, TPSA	1.0000		0.4660
04	log *P*, m.p. − 25, TPSA	1.0000	0.2896	0.3050

This analysis indicates that both melting point and TPSA have similar significance across all models. Again, this supports the hypothesis that an accurate estimate of aqueous solubility for a compound with no known melting point can be arrived at if a modified version of the GSE which substitutes the melting point descriptor with TPSA (e.g., model 03) is employed.

However, a caveat must be applied to this assertion. Unlike the GSE, such a model would not be universally applicable. Compounds which have no polar atoms will have a TPSA value of zero. This would lead to the prediction for such compounds being based entirely on log *P*, which model 01 shows to be an inadequate predictor for aqueous solubility (r^2 = 0.632, RMSE = 1.032). It is possible that the aqueous solubility of compounds containing no polar atoms can be accurately predicted using log *P* alone. However, this is unlikely given that 32% of the total number of nonpolar compounds in the data set were predicted poorly by model 01, including examples such as benzene and numerous halobenzenes. The risk of an inaccurate prediction when using a model such as model 03 must therefore be mitigated by consideration of the molecular structure before application of the model.

Model 04, which incorporates all three molecular descriptors, indicates that the relative importance of TPSA and the melting point parameters are similar and that combining all three descriptors into a single model results in the highest predictive ability of all models (r^2 = 0.873, RMSE = 0.611 for test set). High correlation between descriptors is a critical error in model development and needs to be avoided (Table 4). Significantly,

Table 4. Correlation Matrix Describing the Linear Correlation Coefficient (*r*) between Descriptors Used in the QSPR Models and log *S*

	log *S*	log *P*	m.p. − 25	TPSA
log *S*	1.00	−0.80	−0.25	0.03
log *P*	−0.80	1.00	−0.20	−0.49
m.p. − 25	−0.25	−0.20	1.00	0.55
TPSA	0.03	−0.49	0.55	1.00

the TPSA descriptor shows minimal correlation with melting point, with an *r* value of 0.55 and r^2 value of 0.304.

This suggests that the apparently interchangeable nature of TPSA and the melting point (models 02 and 03), implying at least partial overlap of encapsulation of molecular properties responsible for aqueous solubility by these descriptors, is not absolute, and where possible, predictions should be based on a knowledge of both melting point and the total polar surface area of a compound (in addition to log *P*) in order to ensure a greater accuracy of prediction.

Comparison of Models - Full and Reduced Data Sets. Model 01, which uses log *P* only as a descriptor for predicting aqueous solubility, showed some correlation with experimentally determined values but was considerably poorer in predicting solubility than the GSE (r^2 = 0.644 for model 01 (eq 2) c.f. 0.816 for GSE). The results from both the GSE and model 01 illustrate the inflationary impact of including the sparsely populated regions of the data set on the predictive ability of the model. Removal of the 10% of the data at either end of the solubility spectrum resulted in a significant reduction in predictive ability (model 01 r^2 = 0.524, GSE r^2 = 0.752). Importantly, this analysis serves to reiterate that reliance on log *P* alone to predict aqueous solubility is inadequate.

Model 02 uses the same two descriptors as the GSE but coefficients derived specifically for the data set being investigated, rather than those published in the literature. This provides a benchmark for comparing the performance of a GSE-type model with the alternatives proposed in these studies. The model produced is essentially identical when compared to the "literature" GSE (model 02 r^2 = 0.818 vs 0.816 for GSE for the full data set and model 02 r^2 = 0.754 vs 0.752 for GSE for the reduced data set, see Figure 3). These results imply that direct comparison of the models generated in this study with the GSE documented in the literature is valid. In addition, it again highlights the limitations of a GSE-type model being skewed by peripheral data, as on analysis of the results for the reduced data set there is a decrease in excess of more than 6% in quality, as evidenced by lower r^2 values for both model 02 and the GSE in comparison to the full data set.

Application of model 03 to the full data set yielded an r^2 = 0.813 with a corresponding value of r^2 = 0.742 for the reduced data set. These results are comparable to the results obtained for the GSE and model 02 and support the hypothesis that TPSA can be used to replace melting point as a descriptor for accurately predicting aqueous solubility. This is reinforced by the increase in the number of compounds containing polar/polarizable atoms predicted within ±1 log unit of their experimentally determined values (83.2% compared to 76.9% by GSE). Therefore, this allows the application of the model to virtual compounds, and those with no known melting points, with the caveat that such compounds contained a number of polar/polarizable atoms, as previously discussed.

Model 04, comprising all three descriptors, was shown to have the greatest predictive ability of all models (r^2 = 0.869 for the full data set and r^2 = 0.818 for the reduced data set) with a 5.1% improvement over the GSE for the full data set and a 6.6% improvement over the GSE for the reduced data set (Figure 4). In addition, 90.8% of compounds containing polar/polarizable atoms were predicted within ±1 log unit of their experimentally determined values by this model. These improvements support the initial hypothesis that inclusion of an additional descriptor which accounts for polarizable/polar atoms, such as TPSA, can lead to an enhancement in the predictive ability of the GSE, and overall the model shows that using both melting point and TPSA alongside log *P* in the same model gives better predictions than considering either TPSA or melting point exclusively in conjunction with log *P*.

Poorly Predicted Compounds—Breakdown of Results. The distribution of poorly predicted compounds, with respect to the full data set, from the GSE and models 01 to 04 is illustrated in Figure 5.

Model 01 using log *P* only as a descriptor unsurprisingly shows the greatest number of poorly predicted compounds, which is further evidence that log *P* alone is not a sufficient indicator of the likely aqueous solubility of a compound.

dx.doi.org/10.1021/ci200387c | J. Chem. Inf. Model. 2012, 52, 420−428

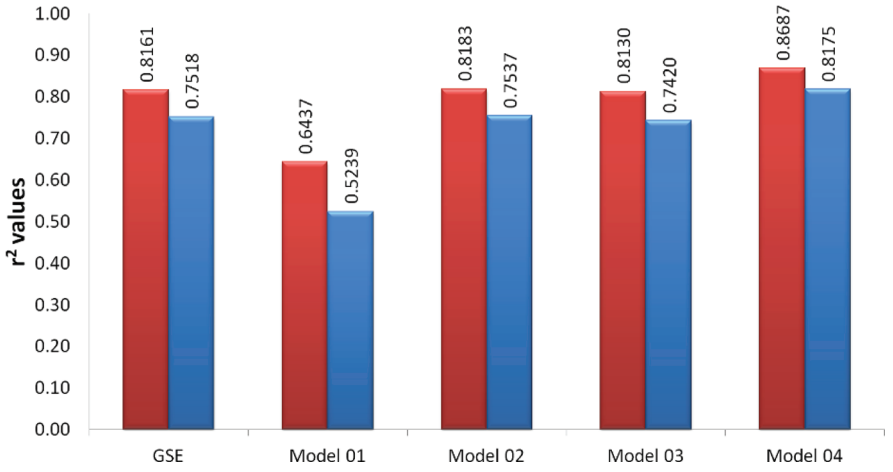


Figure 3. Summary of the regression statistics for the GSE and models 01−04 when applied to the full (blue) and reduced (red) data sets.

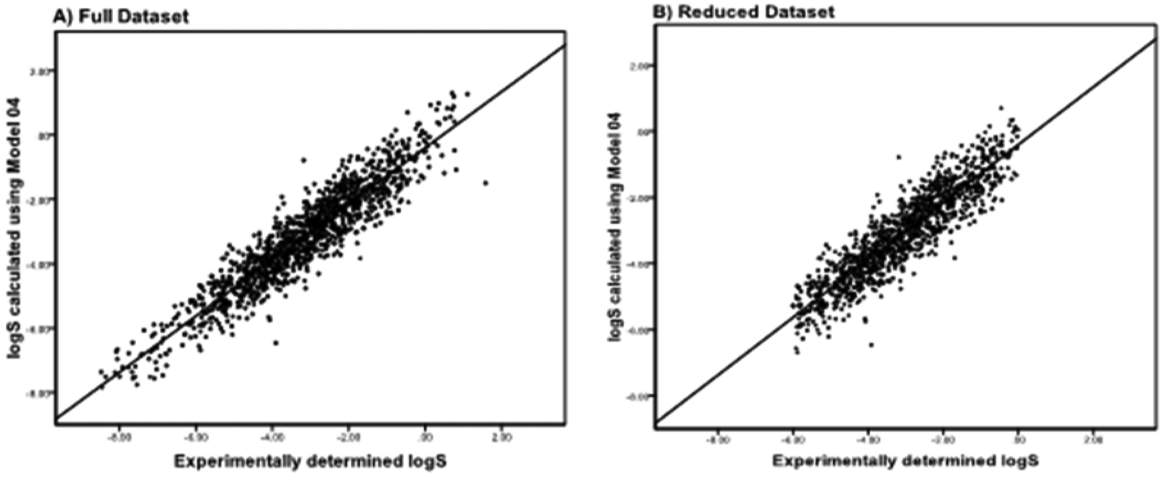


Figure 4. Correlation of predicted versus observed aqueous solubilities after applying model 04 to (A) the full data set (1256 compounds, r^2 = 0.869 and (B) the reduced data set (1160 compound, r^2 = 0.818).

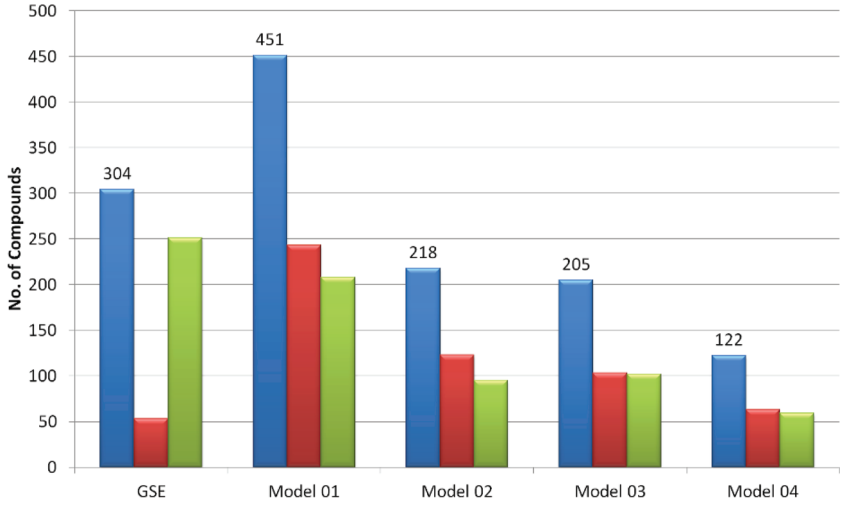


Figure 5. Distribution of poorly predicted compounds in the full data set for the GSE and models 01 to 04. The total number of poorly predicted compounds (i.e., log $S \pm 1$ log unit from the experimentally determined value) for each model (blue) together with an analysis of both under- (green) and overpredictions (red) of aqueous solubility are illustrated.

Analysis on the number of poorly predicted compounds showed 422 compounds containing polar and/or polarizable atoms.

Model 02, which utilizes the same two descriptors as the GSE (log *P* and melting point), shows some notable differences at this level of analysis. The first is that the total number

dx.doi.org/10.1021/ci200387c | *J. Chem. Inf. Model.* 2012, 52, 420−428

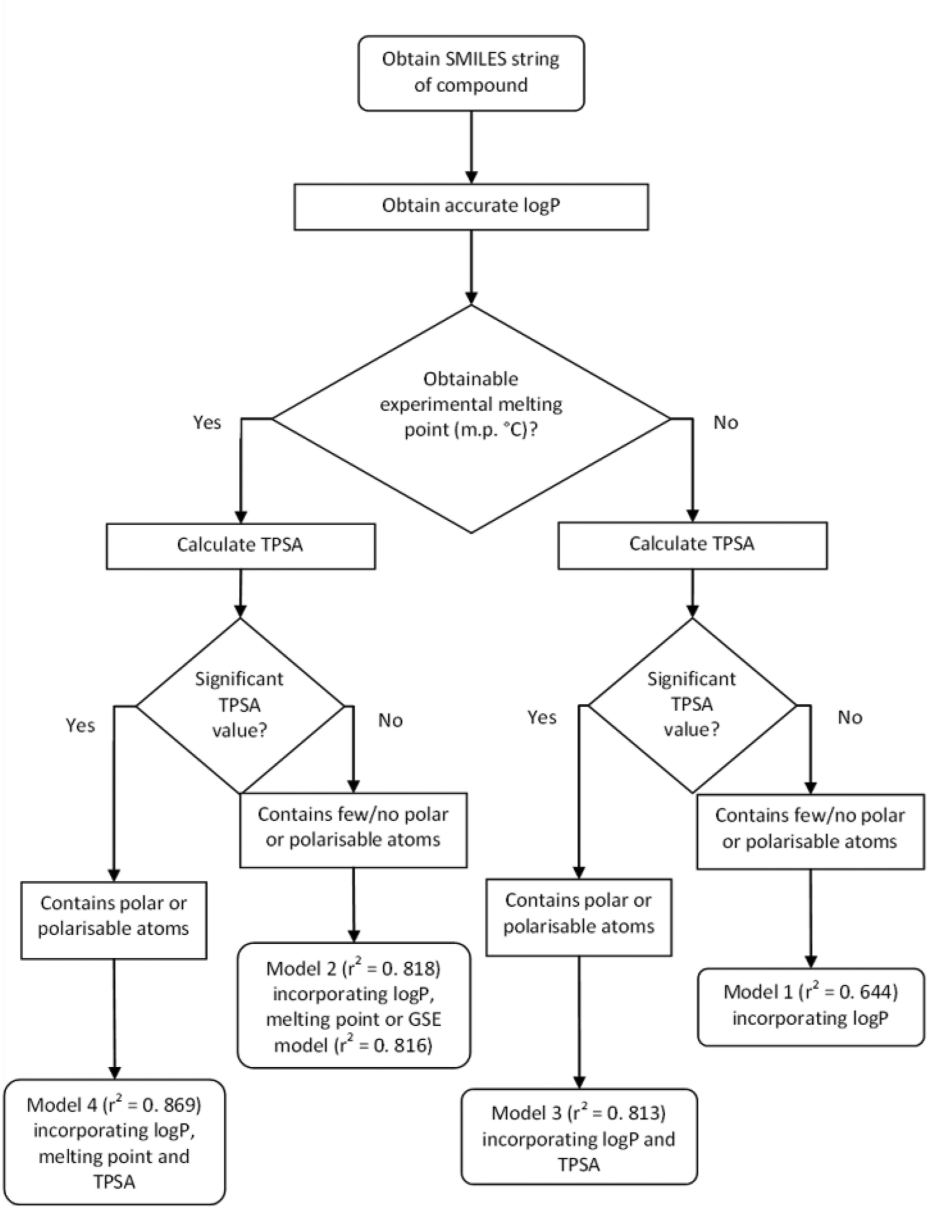


Figure 6. Workflow for prediction of aqueous solubility based on log *P*, melting point, and TPSA descriptors.

of poorly predicted compounds in model 02 is reduced (by 28.3%) in comparison to the GSE even though the r^2 results are similar for both model 02 and the GSE. This is evidence to advocate the generation of coefficients related specifically to the data set being investigated when comparing published models to ones which are being developed. Failure to do so could result in an overinterpretation of any favorable comparisons between models developed, and those in the public domain, which could otherwise be misleading. Comparisons of coefficient values for model 02 compared to GSE are shown to be "less negative;" therefore a bias to underprediction is reduced in model 02 compared to the GSE. The relative number of molecules underpredicted, compared to those which are overpredicted, are similar for model 02, which suggests a robust and well balanced model.

Model 03, which also uses log *P* but substitutes TPSA in place of melting point, also reduced the number of poorly predicted compounds in comparison to the GSE (32.6% reduction) and performs in a similar manner overall to model 02. The number of under- and overpredicted compounds is approximately equal, and again this suggests a robust and well-balanced model. Analysis on the number of poorly predicted compounds containing polar and/or polarizable atoms shows both models 02 and 03 have a reduced number of poorly predicted compounds containing polar and/or polarizable atoms (208 compounds and 196 compounds, respectively) compared with GSE (269 compounds). One reason for the slightly lower number of poorly predicted compounds for model 03 in comparison to model 02 could be that the incorporation of the TPSA descriptor not only improves prediction of compounds containing polar and/or polarizable atoms but also removes erroneous experimental melting point values for some compounds and hence improves their predictions. However, there is not a significant reduction in the total number of poorly predicted compounds between models 02 and 03 (13 compounds), and if errors in melting points are being ameliorated, another source of error must be being exacerbated to an equal magnitude. One hypothesis is

dx.doi.org/10.1021/ci200387c | J. Chem. Inf. Model. 2012, 52, 420−428

that the improvements for compounds with erroneous melting points are balanced out by a concomitant reduction in the ability of model 03 to accurately predict the aqueous solubility of compounds containing no or few polar or polarizable atoms. In effect, molecules which have few or no polar/polarizable atoms have their predicted solubility determined solely by log *P*, and as model 01 has shown, log *P* is not an adequate indicator of aqueous solubility. There is little enrichment of poorly predicted compounds with no polar/polarizable atoms when using model 03 as opposed to model 02. This is evident from the fact that 12 of the 13 compounds for which predictions of aqueous solubility were improved between model 02 and model 03 contained at least one polar/polarizable atom.

The argument centered on the counterbalance between losses and gains dependent upon whether TPSA or melting point is used (models 02 and 03) is further supported by the improvements in predictive ability over the other models when considering model 04. Inclusion of both TPSA and melting point descriptors which have been shown to display similar levels of significance would both reduce the impact of an incorrect prediction if TPSA is, or is close to, zero, and mitigate against an experimental melting point being incorrect. Overall, model 04 improves accurate prediction for 182 compounds and generates poor predictions for only 9.7% of the compounds in the full data set, a 6.6% improvement in predictive ability over its closest competitor (model 03). This indicates that the initial hypothesis that TPSA is an important determinant when predicting aqueous solubility for this data set, and that where the data is available, TPSA should be used alongside log *P* and accurate melting point values for GSE-style models in order to improve predictive quality.

4. CONCLUSION

The GSE's reliance on experimentally determined descriptors limits its applicability, and its predictive ability can be artificially inflated by data sets which are sparsely populated at the limits of their data range, as demonstrated by studies on the full and reduced data sets described above. Additionally, analysis of compounds that were poorly predicted by the GSE showed a large number of compounds which contained polar/polarizable atoms. As such, investigations into the impact of the TPSA descriptor to account for such atoms were incorporated into studies based on improvements of the GSE.

For this data set, a number of models were created using different descriptors with varying predictive ability. The method and effectiveness of predicting solubility is dependent upon the information available (Figure 6). Substituting the melting point descriptor in a GSE-style equation (model 02) with a TPSA descriptor generates a model of similar predictive quality (model 03). Hence, this model could be used to predict aqueous solubility in place of GSE-style models where experimental melting points are unknown, e.g., for virtual compounds, with the caveat that such a model would not be applicable to compounds containing few/no polar or polarizable atoms. Ultimately, a model which incorporates log *P*, melting point, and TPSA (model 04) shows a 5.1% improvement over the GSE in predicting aqueous solubility for the full data set and a 6.6% improvement over the GSE for the reduced data set.

In conclusion, the addition of TPSA, a simple time-effective descriptor which is easily and inexpensively calculated and removes the possibility of inaccurate prediction due to error associated with experimentally derived descriptors, is extremely beneficial in the prediction of aqueous solubility. As shown, the TPSA descriptor can be used in place of melting point to generate a model of similar predictive quality. Consequently, we would advocate the use of a modified version of the GSE which includes TPSA in addition to the familiar log *P* and melting point parameters in order to improve predictions of aqueous solubility for a diverse range of compounds. However, a number of compounds are still poorly predicted by the modified version of the GSE. Therefore, further investigations and continual analysis of poorly predicted compounds shall be undertaken to identify additional structural descriptors for the improvement of aqueous solubility prediction. It is also important to establish the performance of TPSA-based models with respect to other published models of aqueous solubility.

AUTHOR INFORMATION

Corresponding Author

*E-mail: s.b.kirton3@herts.ac.uk.

Present Address

[§]MedPharm Ltd., Unit 3, Chancellor Court, 50 Occam Road, Guildford, Surrey,GU2 7YN, United Kingdom

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge the University of Hertfordshire for a research studentship funding of this work.

REFERENCES

(1) 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.* **2001**, *46* (1−3), 3−26.

(2) Waterbeemd, H. v. d.; Testa, B. *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2009; pp xxv, 624.

(3) Selick, H. E.; Beresford, A. P.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today* **2002**, *7* (2), 109−116.

(4) Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative structure-property relationships in pharmaceutical research - Part 1. *Pharm. Sci. Technol. Today* **2000**, *3* (1), 28−35.

(5) Grover, M.; Singh, B.; Bakshi, M.; Singh, S. Quantitative structure-property relationships in pharmaceutical research - Part 2. *Pharm. Sci. Technol. Today* **2000**, *3* (2), 50−57.

(6) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20* (3−4), 241−266.

(7) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1* (1), 31−52.

(8) Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discovery Today* **2005**, *10* (4), 289−295.

(9) Faller, B.; Ertl, P. Computational approaches to determine drug solubility. *Adv. Drug Delivery Rev.* **2007**, *59* (7), 533−545.

(10) Wang, J.; Hou, T. Recent Advances on Aqueous Solubility Prediction. *Comb. Chem. High Throughput Screening* **2011**, *14* (5), 328−338.

(11) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90* (2), 234−252.

(12) Ran, Y. Q.; He, Y.; Yang, G.; Johnson, J. L. H.; Yalkowsky, S. H. Estimation of aqueous solubility of organic compounds by using the general solubility equation. *Chemosphere.* **2002**, *48* (5), 487−509.

(13) Jain, P.; Yalkowsky, S. H. Prediction of aqueous solubility from SCRATCH. *Int. J. Pharm.* **2010**, *385* (1−2), 1−5.

(14) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14* (5), 568−571.
(15) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *J. Pharm. Sci.* **1999**, *88* (8), 815−821.
(16) *Molecular Operating Environment* (MOE), 2010.10; Chemical Computing Group Inc: Montreal, Canada, 2010.
(17) ChemSpider. www.chemspider.com (accessed May 2011).
(18) Reaxys. www.reaxys.com (accessed May 2011).
(19) SMILES. www.daylight.com (accessed May 2011).
(20) Virtual Computational Chemistry Laboratory. www.vcclab.org (accessed May 2011).
(21) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43* (20), 3714−3717.
(22) Haenlein, M.; Kaplan, A. M. A Beginner's Guide to Partial Least Squares Analysis. *Understanding Statistics* **2004**, *3* (4), 283−297.
(23) Cross-validation. www.qsarworld.com/qsar-ml-cross-validation.php (accessed May 2011).
(24) Schultz, T. W.; Netzeva, T. I.; Cronin, M. T. D. Selection of data sets for qsars: Analyses of tetrahymena toxicity from aromatic compounds. *SAR QSAR Environ. Res.* **2003**, *14* (1), 59−81.
(25) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33* (1), 347−350.