JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

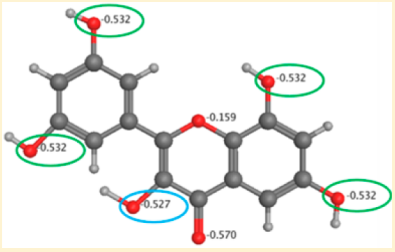
In Silico Prediction of Aqueous Solubility Using Simple QSPR Models: The Importance of Phenol and Phenol-like Moieties

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, and

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

ABSTRACT: Recently the authors published a robust QSPR model of aqueous solubility which exploited the computationally derived molecular descriptor topographical polar surface area (TPSA) alongside experimentally determined melting point and logP. This model (the "TPSA model") is able to accurately predict to within ± one log unit the aqueous solubility of 87% of the compounds in a chemically diverse data set of 1265 molecules. This is comparable to results achieved for established models of aqueous solubility e.g. ESOL (79%) and the General Solubility Equation (81%). Hierarchical clustering of this data set according to chemical similarity shows that a significant number of molecules with phenolic and/or phenol-like moieties are poorly predicted by these equations. Modification of the TPSA model to additionally incorporate a descriptor pertaining to a simple count of phenol and phenol-like moieties improves the predictive ability within ± one log unit to 89% for the full data set (1265 compounds −8.48 < logS < 1.58) and 82% for a reduced data set (1160 compounds 6.00 < logS < 0.00) which excludes compounds at the sparsely populated extremities of the data range. This improvement can be rationalized as the additional descriptor in the model acting as a correction factor which acknowledges the effect of phenolic substituents on the electronic characteristics of aromatic molecules i.e. the generally positive contribution to aqueous solubility made by phenolic moieties.



INTRODUCTION

Knowledge of the intrinsic aqueous solubility of compounds is important for numerous disciplines including the medicinal, physical, and environmental sciences. It can be simply described as a measure of the dissolution of an un-ionized substance in water. The accurate measurement/prediction of aqueous solubility of a new chemical entity (NCE) is very useful in the pharmaceutical industry, given that it may impact on the likelihood of a new drug being developed as a medicine.[1] In addition, aqueous solubility must also be considered by the agrochemical industry, particularly with respect to the design and synthesis of pesticides where the agricultural benefits must be considered alongside the environmental impact of these chemicals being washed into waterways. Experimental determination of aqueous solubility (logS) is the preferred method for establishing values as it provides the most reliable measurements. However, especially when considering large libraries of compounds, the time and significant costs associated with experimental determination of aqueous solubility become prohibitive.[2] This provides considerable motivation to identity methods for accurate prediction of aqueous solubility via reliable and robust *in silico* models.[3]

Numerous computational models have been proposed for predicting aqueous solubility,[4−7] with many utilizing Quantitative Structure−Property Relationships (QSPR). Two of the most widely-known QSPR-based models for prediction of aqueous solubility (logS) are the General Solubility Equation (GSE) proposed by Yalkowsky and Jain[8] and the Estimated SOLubility (ESOL) model proposed by Delaney.[9]

Both the GSE and ESOL models exploit the octanol−water partition coefficient of un-ionized molecules (log P) in their respective QSPR equations. The GSE model also incorporates an expression derived from the melting point of the compound expressed as m.p. (C)-25 in addition to logP making it a two-descriptor model 1.

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

In contrast, ESOL (2) is a 4-descriptor model which includes the molecular weight (MWT) of a compound, the number of rotatable bonds the molecule possesses (RB), the proportion of heavy atoms in the molecule which are directly associated with an aromatic system (AP), and the logP of the molecule.

$$\log S = 0.16 - 0.63\ \text{clogP} - 0.0062\ \text{MWT} + 0.066\ \text{RB} - 0.74\ \text{AP} \quad (2)$$

The predictive abilities of the GSE and ESOL models are comparable and both have been widely adopted by the industry for use in the estimation of aqueous solubility. The predictive power of the GSE equation has been reported with a coefficient of determination (r^2) = 0.96 and root-mean-square error (RMSE) = 0.53 for a data set containing 1026 organic compounds.[10] However, the experimental melting point descriptor used in the GSE can limit its applicability. For example, the GSE cannot be readily applied to libraries of

Received: September 19, 2012
Published: October 28, 2012

virtual compounds where experimental melting points have yet to be determined. It should be noted that a recent model proposed by Yalkowsky and co-workers (SCRATCH)[11] incorporates a complex algorithm to predict melting point. SCRATCH displays similar predictive ability when compared to the original GSE, thus circumventing the issues with virtual compounds.

The ESOL model (2) addresses the issue of reliance on experimentally determined properties for calculation of aqueous solubility, by using descriptors calculated directly from the molecular structure of a compound. The ESOL model showed good predictive ability when it was originally formulated (r^2 = 0.72 and average absolute error (AAE) of 0.75 for a data set containing 2874 organic compounds).[9]

However, the ESOL, GSE, and SCRATCH models do not explicitly account for the effect of polar and polarizable atoms on aqueous solubility. A recent study investigated modifications to the GSE using a simple, time effective descriptor for calculating polar surface area (TPSA).[12] The addition of TPSA was shown to be beneficial in the prediction of aqueous solubility as it was possible to use this descriptor as an alternative for the melting point term to generate a model with similar predictive ability to the original GSE (r^2 = 0.813 for a data set of 1256 compounds). However, a three-descriptor "TPSA" model (3) which incorporated logP, melting point, and TPSA performed even better (r^2 = 0.869 for a data set of 1256 compounds).

$$\log S = -1.0144 \log P - 0.0056 (\text{m.p. (C)-25}) - 0.0134 \text{ TPSA} + 0.5134 \quad (3)$$

Despite the improvements over the standard GSE, a number of compounds were still poorly predicted by the TPSA model. The investigations outlined below sought to identify common chemical scaffolds in the poorly predicted compounds as a route to identifying additional molecular descriptors which could be incorporated into and improve the QSPR models. In keeping with the original investigations, the focus for the present work was on identifying descriptors which were accessible and easily identifiable from chemical structure.

METHODS

Manipulation of the Data Set. A data set of 1256 structurally unique compounds was curated using the Molecular Operating Environment (MOE)[13] software. Generation and comparison of canonical SMILES[14] strings ensured all structures in the data set were unique.

The data set contained compounds with no formal electrostatic charge at physiological pH (7.4). Each compound has an experimentally determined thermodynamic aqueous solubility value (logS) with which it is associated. These values have been extracted from reliable literature sources.[12] In addition, each molecule also has a validated octanol−water partition coefficient (logP)[15] value and an experimental melting point (m.p. C) value. The TPSA descriptor for each compound was calculated in MOE using the fast method based on a sum of fragment based contribution.[16]

In order to reproduce the ESOL model (2), so it could be applied to the full and reduced data set, the molecular weight (MWT) and the number of rotatable bonds (RB) were calculated from molecular structure using MOE. The proportion of heavy atoms in the molecule which are directly associated with aromatic systems (AP) was also derived from chemical structure using MOE by establishing the ratio between the number of heavy aromatic atoms (a_aro) and the total number of heavy atoms (a_heavy) in a compound.

The data set compounds span a range of logS 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 the diverse subset tool within MOE. This delineation was achieved using MACCS Structural Keys as the discriminant.

In addition, a reduced data set was constructed to focus on the range of solubility values populated by the majority of compounds. This reduced data set was used as an additional test of model quality. Validation using the reduced data set guards against artificial inflation/depression with respect to the predictive ability of models generated by regression-based analyses. 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 removed resulting in a data set containing 1160 compounds (92.4% of the full data set). A more complete discussion behind the rationale of using a reduced data set as a quality control mechanism in QSPR models is presented in our previous study.[12]

Application of Existing QSPR Models to the Full and Reduced Data Sets. The GSE (1), ESOL (2), and the TPSA model (3) were applied to the full and reduced data sets. Linear regression analysis was used to correlate the measured logS values against the predicted value of each model and produce corresponding r^2 and RMSE values for both data sets. Poor predictions 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.[4]

Cluster Analysis. Cluster analysis utilizing QuaSAR-Cluster[13] was performed on the full data set using a fingerprint clustering algorithm. The Jarvis-Patrick clustering method using MACCS Structural Keys grouped compounds according to chemical structure in order to identify trends in prediction as well as highlight chemical moieties that were poorly predicted by each model. A similarity threshold of 70% and clustering utilizing a Tanimoto coefficient matrix also set at 70% were employed.

Determination of Molecular Descriptors: Phenolic Descriptor (aroOHdel). The phenol/phenol-like descriptor, (aroOHdel), was generated using SMART[17] coding in the MOE program. Initially a descriptor (aroOH) identifying the number of hydroxyl substituents attached to an aromatic carbon atom "c[OH]" was used to count all of the simple phenol moieties for each of the compounds in the data set. This group of compounds was combined with those highlighted from a second SMART match "[OH]C=C[#6i]" which focused on identifying hydroxyl groups which, although not directly connected to an aromatic carbon atom, were attached to conjugated π systems, and as such could be argued were at least "phenol-like" with respect to their electronic characteristics. This descriptor was termed (OHdel). The aroOH and OHdel descriptors were then combined to give a single descriptor (aroOHdel) which described the total number of "true phenol" and "phenol-like" hydroxyl moieties in a given compound. The functionality of this descriptor was supported by identifying 12 compounds from the data set containing different combinations of true phenol and phenol-like moieties (Figure 1) and ensuring that the expected number for aroOHdel was returned

in each case before this descriptor was used to characterize the compounds in the full and reduced data sets.

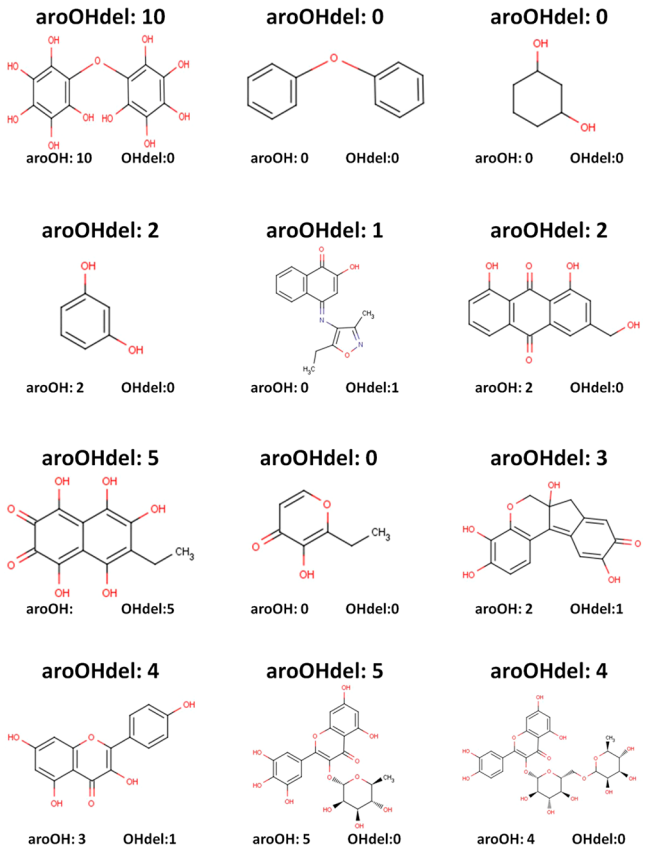


Figure 1. Representative structures to illustrate the derivation and application of the aroOHdel descriptor.

A four descriptor QSPR model exploiting logP, mp-25, TPSA, and aroOHdel (the "Phenol Model") was built using the training set of 1004 compounds used to construct the previous QSPR models[12] and MOE, specifically its QuaSAR-Model function with the Partial least-squares (PLS) regression method.[18] No limit was placed on the degree of the fit for these investigations. The maximum number of conditions permissible before the model finds the best fit was set at 10^5 for each iteration. Overall robustness of the model was assessed by considering r^2 and RMSE values i.e. leave-one-out cross-validated training set analysis.[19] In addition, the predictive ability of the models generated was quantified by assessing how well the model predicted the aqueous solubility of the test set compounds. The performance of the Phenol model on both the full and reduced data sets gives an indication of how robust the models are. Jarvis-Patrick cluster analysis was used as previously defined to identify trends in prediction and highlight chemical moieties that were poorly predicted.

RESULTS AND DISCUSSION

The Database. The definition of the training and test sets was achieved using the diverse subset tool within MOE which gives a representative distribution of the molecules in both test and training sets when compared to the data set as whole.[12] A subset of the full data set ($-6.00 < \log S < 0.00$) focusing on the region of dataspace occupied by the majority of the compounds (92.4%) was also defined (Figure 2). The full and reduced data sets were used to benchmark the predictive ability of the GSE, ESOL, and TPSA models.

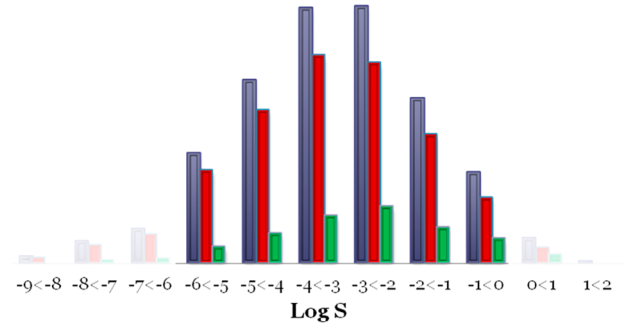


Figure 2. Distribution of compounds with respect to logS in the full (blue), training (red), and test (green) data sets. Those regions of the data set which are considered only in the reduced data set ($-6.00 < \log S < 0.00$) analyses are in bold.

Benchmarking Literature Models with the Full and Reduced Data Sets. The GSE (1) was applied to the full (1256 compounds) and reduced (1160 compounds) data sets. The correlation of predicted versus observed aqueous solubilities for the full data set (r^2 = 0.816, RMSE = 0.848) and the reduced data set (r^2 = 0.752, RMSE = 0.831) were recorded. This analysis suggests that for this data set the sparsely populated regions of the data set (logS −6.00 and logS > 0.00) inflate the predictive ability of the GSE by 6.4%, although this may be as a consequence of the fact that the GSE was not derived using this data set of compounds.

The correlation between predicted and observed aqueous solubilities after the ESOL model (2) was applied showed r^2 = 0.793, AAE = 0.649 for the full data set and r^2 = 0.717, AAE = 0.610 for the reduced data set. This improvement in predictive ability for our data set in comparison to the original data set used to derive the ESOL model (r^2 = 0.72, AAE = 0.75) is attributable to the fact that the data sets examined are different (the data set used to derive the ESOL model is much larger than the data set being examined in this study). The sparsely populated regions of the data set result in inflation of the predictive ability for the ESOL model, which is similar to the results observed for the GSE. Both the GSE and ESOL models demonstrate considerable predictive abilities when applied to the full data set, with both models able to accurately (i.e., to within ±1 log unit) predict the aqueous solubility of approximately 80% of the compounds in the catalogue.

In comparison, the TPSA model (3) shows improved predictions of aqueous solubility over both GSE and ESOL for the full (r^2 = 0.869, RMSE = 0.619, AAE = 0.491) and reduced data sets (r^2 = 0.818, RMSE = 0.604, AAE = 0.481). However, all three models show that there are a significant number of compounds in the data set for which the accurate prediction of aqueous solubility is poor.

The TPSA model fails to predict the aqueous solubility of 122 (9.7%) compounds to within ±1 log unit of their experimentally determined values, whereas the GSE and ESOL model failed to predict 304 (24.2%) and 260 (20.7%) compounds respectively. The nature of these poorly predicted compounds was examined further using cluster analysis.

Cluster Analysis. The QuaSAR-Cluster feature in MOE was used to cluster compounds in the data set according to their MACCS Structural Keys molecular fingerprint. 543

dx.doi.org/10.1021/ci300447c | J. Chem. Inf. Model. 2012, 52, 2950−2957

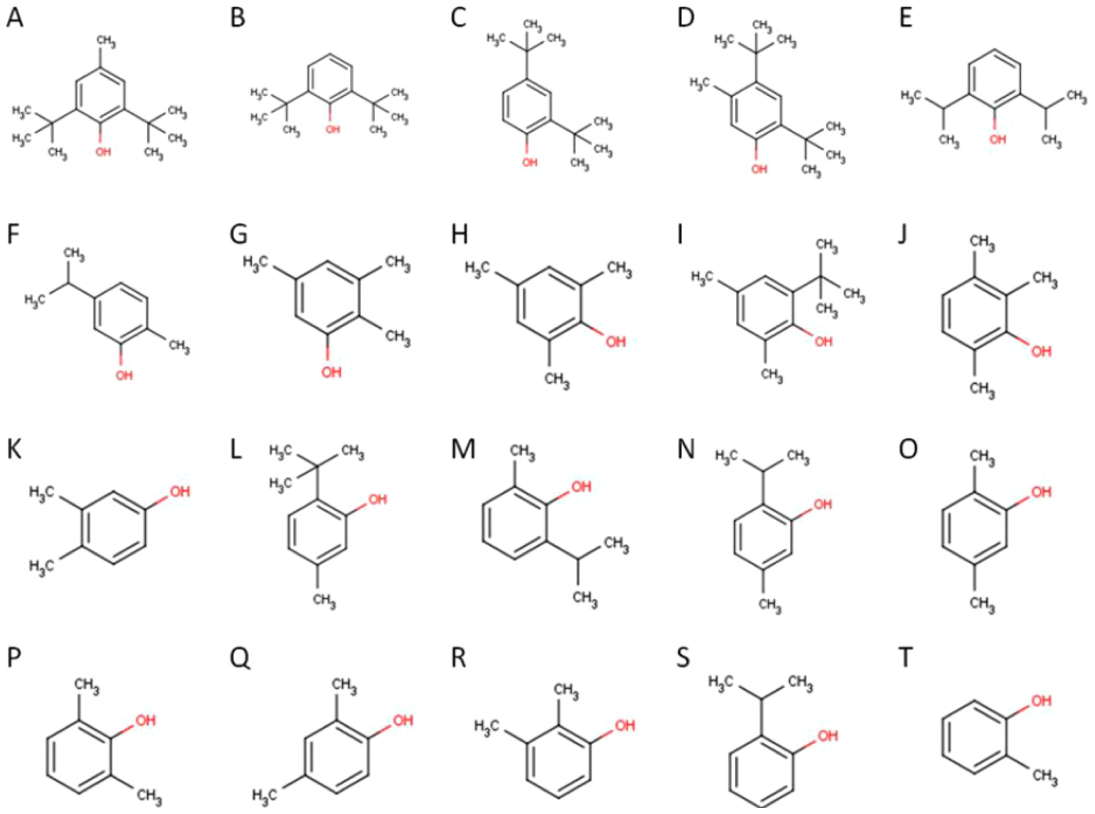


Figure 3. Structures of the 20 phenolic compounds in CLUSTER 01. Compounds R and T were poorly predicted by the GSE (1), compounds M, N, R, S, and T were predicted incorrectly by the TPSA model (3), and compounds N, O, P, Q, R, S, and T were predicted incorrectly by ESOL (2).

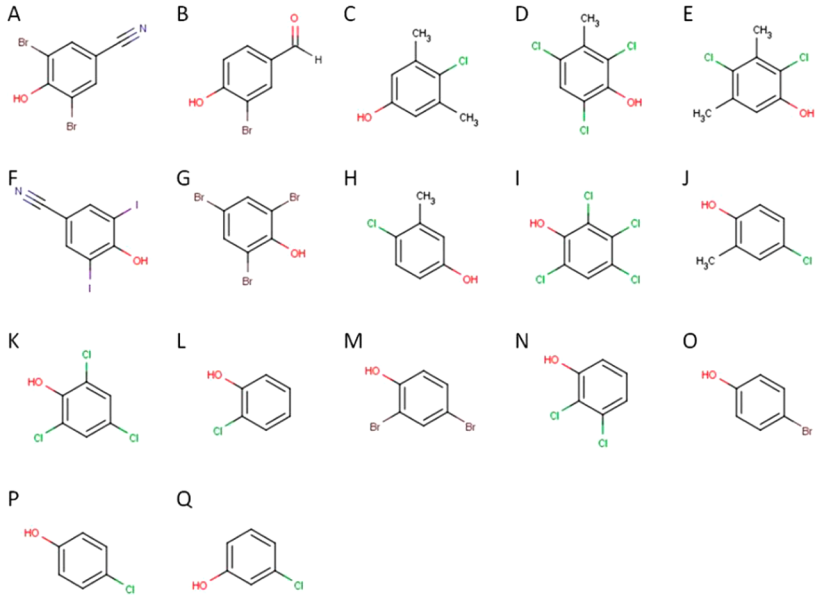


Figure 4. Structures of the 17 halogenated phenolic compounds comprising CLUSTER 02. Compounds C, I, J, K, N, O, P, and Q were predicted incorrectly by the GSE (1), compounds H, I, J, K, N, O, P, and Q were predicted incorrectly by the TPSA model (3), and compounds G, H, I, J, K, L, M, N, O, P, and Q were predicted incorrectly by ESOL (2).

(43.2%) of the 1256 compounds in the full data set clustered into 153 nonorphan clusters when similarity thresholds of 70% and a Tanimoto coefficient matrix set at 70% were applied. 508 (43.8%) compounds gave 138 nonorphan clusters when the same methodology was applied to the 1160 compound reduced data set.

Investigation into the results of the cluster analysis centered on the 122 poorly predicted compounds from the TPSA model (3) showed 29 (23.7%) of the poorly predicted compounds belonged to nine nonorphan clusters and that three of these clusters contained at least two poorly predicted molecules. The remaining 93 poorly predicted compounds were either the only

dx.doi.org/10.1021/ci300447c | J. Chem. Inf. Model. 2012, 52, 2950−2957

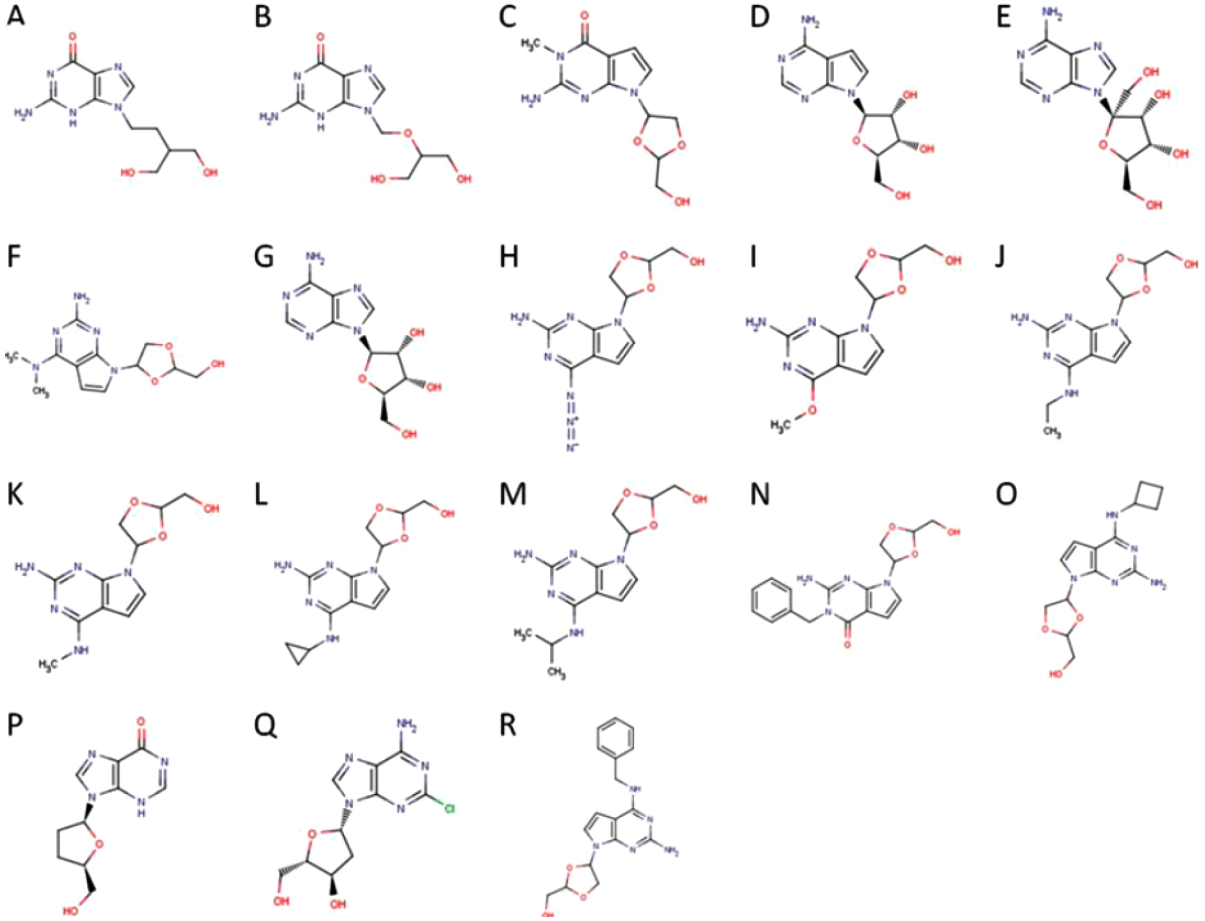


Figure 5. Structures of the 18 compounds (Cluster 03) which contain a large number of polar/polarizable atoms. Compounds A, B, C, and F were predicted incorrectly by the TPSA model (3), and compounds A, B, and C were predicted incorrectly by ESOL (2). The GSE model (1) only predicted O, P, Q, and R correctly.

molecule in an orphan cluster or the only poorly predicted compound in a cluster containing two or more compounds.

In comparison, the results highlighted by the GSE show 106 of the 304 poorly predicted compounds (34.9%) can be found in 35 clusters containing at least two compounds with 11 clusters having at least 2 poorly predicted compounds. ESOL has 75 (28.8%) of its poorly predicted compounds contained in 25 clusters. Eleven of these clusters contain at least two poorly predicted compounds. Arising from these separate analyses it was determined that three clusters contained at least two compounds which had been poorly predicted by all three models of aqueous solubility. These clusters were examined to investigate what the common cause of these failures might be.

"CLUSTER 01" consisted of 20 simple compounds each containing a phenol moiety (Figure 3). The GSE accurately predicted all but 2 of the 20 compounds (i.e., to within one log unit of the experimentally observed aqueous solubility) and was the best performing of the three models investigated. The TPSA model failed to accurately predict five of the compounds, and ESOL failed for 7 of the 20 structures.

CLUSTER 02 contained a number of compounds which were poorly predicted by each of the models (Figure 4). Of the 17 compounds in the cluster, 8 compounds were predicted poorly for both the GSE and TPSA models, and 11 compounds were predicted poorly by ESOL. The predominant structural feature relating these compounds to one another, but discriminating them from compounds contained in CLUSTER 01, was the presence of a halogenated phenol ring system (see Figure 4).

CLUSTER 03 (Figure 5) was the final cluster identified as having at least two compounds poorly predicted by all 3 models of aqueous solubility. The compounds in the cluster are characterized by being derivatives, or closely related analogues, of 2-aminopurine and 5H-pyrollo[2,3- d]pyrimidine-2-amine core scaffolds. Of the 18 compounds comprising CLUSTER 03, 14 compounds were poorly predicted by the GSE. This number was reduced to four for the TPSA model, an improvement which is consistent with the need to explicitly account for polar/polarizable atoms in order to accurately predict aqueous solubility.[12] The ESOL model was the best performing of the three models predicting all but 3 of the 18 compounds within ± one log unit.

Phenolic Descriptor. Given that both CLUSTERs 01 and 02 showed that all three models performed poorly on compounds containing phenolic moieties, further investigation with respect to the impact of phenols and phenol-like functionalities on the accurate prediction of aqueous solubility was warranted.

Analysis of the 122 poorly predicted compounds from the TPSA model which failed to predict to within ±1 log unit indicated 41 compounds contained either aromatic hydroxyl substituents (identified using the aroOH descriptor) and/or delocalized hydroxyl i.e. "phenol-like" substituents (identified using the OHdel descriptor). The two were combined to give

dx.doi.org/10.1021/ci300447c | J. Chem. Inf. Model. 2012, 52, 2950−2957

an overall number for phenolic and "phenol-like" hydroxyl groups in a compound (aroOHdel; see Figure 1). Analysis of the full data set of compounds from the TPSA model showed 209 compounds contained phenol and/or phenol-like functionality. Of these, 201 compounds contained phenolic hydroxyl groups only, 4 contained phenolic-like hydroxyl groups only, and the remaining 4 compounds contained both phenolic and phenol-like hydroxyl groups.

Comparing the predictive ability of the GSE, ESOL, and TPSA models for the subset of 209 compounds showed all three models had a tendency to underpredict aqueous solubilities for compounds containing phenolic and/or phenol-like groups. In detail, the TPSA model failed to predict a total of 41 phenol/phenol-like containing compounds to within ±1 log unit with 39 of these being underpredicted. The GSE model failed to accurately predict 42 compounds from the phenol subset, with 27 being underpredicted. Analysis of results from the ESOL model confirmed this trend with 56 phenol/phenol-like containing compounds being poorly predicted, with 41 of these compounds being underpredicted.

The rationale behind combining phenols and phenol-like moieties into a single descriptor arises from examination of structures in the database such as morin (Figure 6), the aqueous solubility of which was predicted poorly by all three models. The compound contains four discrete phenol substituents and one delocalized hydroxyl substituent. When the partial charges of the oxygen atoms were calculated, the values observed for the phenolic and delocalized hydroxyl substituents were very similar to one another. This similarity between partial charges is not unexpected, given that the extent of delocalization due to resonance will be essentially the same for "true" phenols and those hydroxyls attached to conjugated π systems. This offers the possibility that phenols and conjugated hydroxyl moieties may be treated as equivalent for the purposes of predicting aqueous solubility, and hence the combined aroOHdel descriptor is justified.

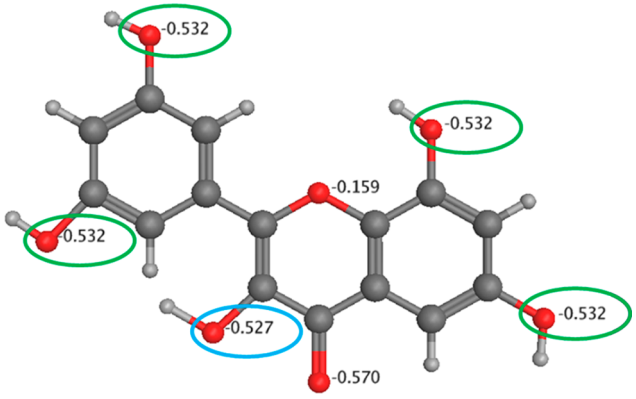


Figure 6. Structure of morin showing partial charges for the oxygen atoms the circled phenolic (green) and the conjugated hydroxyl moieties (blue) being counted by the aroOHdel descriptor. Partial charges were calculated using a parametrized all-atom forcefield (MMFF94x) in the MOE software.

A "Phenol" QSPR model exploiting logP, TPSA, melting point, and aroOHdel as molecular descriptors was created using the training set of 1004 compounds. This model (4) was validated using internal cross-validation and by assessing its predictive ability with respect to the test set of 252 compounds.

$$\log S = -1.0239\ \log P - 0.0148\ \mathrm{TPSA} - 0.0058\ (\mathrm{m.p.\ (C)\text{-}25}) + 0.3295\ \mathrm{aroOHdel} + 0.5337 \quad (4)$$

The Phenol model generated showed r^2 = 0.885 and RMSE = 0.574 (n = 1004). The model is robust as evidenced by the fact that the cross-validated (leave-one-out) regression statistics have very similar values when compared to the initial results (CV r^2 = 0.885, CV RMSE = 0.578). This new model showed an increase in predictive ability in comparison to the TPSA model (3) (r^2 = 0.866, RMSE = 0.622, n = 1004, CV r^2 = 0.865, CV RMSE = 0.624).

The Phenol model also demonstrated similar accurate predictive ability when applied to the test set molecules (r^2 = 0.876, RMSE = 0.599, n = 252) which was comparable to the performance for the TPSA model (r^2 = 0.873, RMSE = 0.611, n = 252).

Application of the Phenol model (4) to both the full data set of 1256 compounds (r^2 = 0.886) and reduced data set of 1160 compounds (r^2 = 0.842) showed the new model to have a greater predictive ability when compared with existing models. There is a 7−8% improvement over the GSE and ESOL models for the full compound data set and a 9−11% improvement over the GSE and ESOL models for the reduced data set, when considering the number of compounds accurately predicted to within ±1 log unit of their experimentally observed aqueous solubility. Comparison to the TPSA model shows slight improvement with a 1.7% increase with respect to the full data set but a greater than 2.3% improvement when the reduced data set is considered (Table 1).

Analysis of the poorly predicted compounds from the Phenol model indicated that only 91 compounds were incorrectly predicted compared to 122 from the TPSA model. Analysis of these 91 poorly predicted compounds showed 17 to contain either phenol or phenol-like moieties with 10 of the 17 compounds being underpredicted. This compares to 41 incorrectly predicted phenol/phenol-like containing compounds from the TPSA model with 39 compounds being underpredicted and is evidence to support the importance of considering the impact of phenol/phenol-like functionality when predicting aqueous solubility.

Table 1. Summary of the Physicochemical Parameters and Regression Statistics for the Predictive Quality of Previously Published Models[8,9,12] with the Developed Phenol Model

		full data set (n = 1256)			reduced data set (n = 1160)		
model	description	r^2	RMSE	AAE	r^2	RMSE	AAE
GSE	LogP, m.p.-25	0.816	0.848	0.677	0.752	0.831	0.665
ESOL	LogP, MWT, RB, AP	0.793	0.822	0.649	0.717	0.767	0.610
"TPSA" model	LogP, m.p.-25, TPSA	0.869	0.619	0.491	0.818	0.604	0.481
"Phenol" model	LogP, m.p.-25, TPSA, aroOHdel	0.886	0.580	0.463	0.841	0.564	0.452

dx.doi.org/10.1021/ci300447c | J. Chem. Inf. Model. 2012, 52, 2950−2957

The problematic simple phenol cluster which had previously been identified was less of an issue for the Phenol model, in that of the five compounds in CLUSTER 01 which were poorly predicted by the TPSA model only one (compound T, Figure 3) was still poorly predicted when the Phenol model was applied.

Furthermore the problematic CLUSTER 03, which the TPSA model had performed well with, was also improved after the Phenol model was applied with only 2 of the 18 compounds in the cluster (Compounds B and C, Figure 5) being poorly predicted. This improvement was unexpected as the compounds in CLUSTER 03 do not contain any phenol and/or phenol-like hydroxyl groups; therefore, the additional descriptor should have had zero effect on these compounds. This suggests that the change in the intercept value and the coefficients preceding the logP, TPSA, and (mp-25 C) in the Phenol model as a consequence of including the additional aroOHdel descriptor are responsible for this "improvement", which is therefore artifactual. Analysis of the compounds in CLUSTER 03 showed both compounds which were improved in the Phenol model were overpredicted by the TPSA model, but the predicted values were very close to the ±1 log unit boundary for a "good" prediction. The Phenol model lowered the predicted values for these compounds so that these values were now within 1 log unit of their experimentally observed aqueous solubility, although the actual difference between the TPSA and Phenol predicted values for both compounds was small. This is further evidence to support the improvement seen for this cluster is an artifact of the equation and not as a consequence of specifically incorporating a phenol count descriptor.

The only significant cluster which persisted after application of the Phenol model was the cluster containing the phenolic halide moieties (CLUSTER 02). Of the 17 compounds in the initial cluster, 8 were poorly predicted by the TPSA model, and only 2 of these poorly predicted compounds showed improvement and were predicted to within ±1 log unit after the Phenol model was applied (Figure 4 Compounds H and J). Six compounds were still poorly predicted (Compounds I, K, N, O, P, and Q, Figure 4).

The overall improvement in prediction after the incorporation of the aroOHdel descriptor suggests that despite the inclusion of TPSA in earlier models compounds with aromatic and conjugated hydroxyl substituents were not sufficiently described by the TPSA model. Analysis of the algorithm from which the TPSA values for the molecules in the data set were derived showed aliphatic hydroxyl moieties contribute a value of 20.23 per moiety to the overall TPSA value, and aromatic hydroxyl moieties contribute a value of 13.14 per moiety to the overall TPSA value. Hence, each aliphatic hydroxyl moiety contributes approximately 1/3 more to the overall TPSA value when compared to aromatic hydroxyl moieties. However, it is commonly accepted that a phenolic hydroxyl group can significantly alter the electronic properties of, for example, a benzene ring as evidenced by variation in Hammett substituent (σ) values for a hydroxyl substituent at the meta and para positions.[20] This implies, in the case of the TPSA model, the importance of the effect of phenolic substituents on the electronic characteristics of aromatic systems, and, as a consequence, the impact of phenolic moieties on the aqueous solubility of a compound is underestimated in the majority of cases.

As an example, a hypothetical molecule containing a single phenol group is considered, and the TPSA model calculates the contribution to overall solubility for the TPSA descriptor for such a molecule as −0.176 (−0.0134 × 13.14). The Phenol model would arrive at a value of −0.194 for the TPSA component of the equation, but this would then be increased via the aroOHdel descriptor (+0.33) to an overall contribution to aqueous solubility of +0.14. This indicates that the aroOHdel descriptor is essentially acting as a correction factor which acknowledges the positive contribution to aqueous solubility made by phenolic moieties, especially in molecules which have small numbers of polar and polarizable atoms such as those occupying CLUSTER 01. If TPSA increases, without a concomitant increase in the number of phenolic groups this effect is somewhat diluted, but again this is indicative of the importance of phenols and conjugated hydroxyl groups in increasing aqueous solubility. It is reasonable to assume, given the underestimation of the significance of phenolic groups in determining aqueous solubility, that there may be other atom types/functional moieties in the TPSA algorithm which contribute to the underestimation of solubility. However, cluster analyses did not immediately identify any such functionalities.

Halogen Descriptor. The success of incorporation of a simple phenolic/phenolic-like hydroxyl count to improve the prediction of aqueous solubility suggested a similar approach for other problematic functional groups highlighted by the cluster analysis. As such, a similar approach was employed in an attempt to improve the prediction of the problematic phenolic halides cluster (CLUSTER 02). However, investigations into the TPSA algorithm showed that it does not ascribe values to halide substituents, and hence application of a correction factor based on a halide count was meaningless and did not improve predictive ability. Further investigation into the contribution of aromatic halide moieties to aqueous solubility is required to improve the QSPR models still further, as this group of molecules remains problematic for a range of well-established equations. The effect of halides/phenolic halides on aqueous solubility remains an active area of research.

CONCLUSION

Analysis of established models which give a reasonably accurate prediction of aqueous solubility demonstrated that a number of compounds were consistently inaccurately predicted by a number of independently developed QSPRs.

Cluster analysis identified compounds containing phenolic moieties to be problematic. Incorporating a simple "phenol/phenol-like" hydroxyl count improved the predictive ability for both the full data set and a reduced data set. The additional descriptor is essentially acting as a correction factor which acknowledges the effect of phenolic substituents on the electronic characteristics of aromatic systems and as a consequence the positive contribution to aqueous solubility made by phenolic moieties. The previously best-performing model, the TPSA model, had underestimated the contribution of phenol and phenol-like moieties to increasing the aqueous solubility of compounds, particularly those compounds with small numbers of polar/polarizable atoms.

Further analysis showed compounds containing aromatic halide moieties are also poorly predicted by a range of models. Incorporation of simple counts of halide functionality, in a manner similar to that employed for phenols, does not improve the predictive ability of the TPSA model, due to the fact that

halides are not explicitly accounted for in the TPSA algorithm, and hence application of a halide correction factor is meaningless. As such, further investigation is required to identify descriptors which can account for the effect on aqueous solubility due to aromatic halide functionality in order to improve predictions for this class of compounds.

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.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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

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) Jouyban, A.; Fakhree, M. A. A. Experimental and Computational Methods Pertaining to Drug Solubility. In *Toxicity and Drug Testing*, 1.; Acree, W., Ed.; InTech: Croatia, 2012; p 187.

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

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

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

(7) Wang, J.; Hou, T. Recent advances on aqueous solubility prediction. *Comb. Chem. High Throughput Screening* **2011**, *14* (5), 328−338.

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

(9) Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1000−1005.

(10) 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.

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

(12) Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B. Revisiting the general solubility equation: In silico prediction of aqueous solubility incorporating the effect of topographical polar surface area. *J. Chem. Inf. Model.* **2012**, *52* (2), 420−428.

(13) Molecular Operating Environment (MOE), 2010.10; Chemical Computing Group Inc.: 2010.

(14) SMILES − A Simplified Chemical Language. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed October 18, 2012).

(15) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−63.

(16) 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.

(17) SMARTS − A language for describing molecular patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed October 18, 2012).

(18) Haenlein, M.; Kaplan, A. M. A beginner's guide to partial least squares analysis. *Understanding Stat.* **2004**, *3* (4), 283−297.

(19) Cross-validation. www.qsarworld.com/qsar-ml-cross-validation.php (accessed October 18, 2012).

(20) Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **1937**, *59* (1), 96−103.