Sussex Research Online

Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes

Article (Accepted Version)

Ghafourian, Taravat and Bozorgi, A.H.A. (2010) Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes. European Journal of Pharmaceutical Sciences, 40 (5). pp. 430-440. ISSN 0928-0987

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/64140/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Estimation of drug solubility in water, PEG 400 and their binary mixtures using the molecular structures of solutes

Taravat Ghafourian[*a], Atefeh Haji Agha Bozorgi* [b, c]

[a] Medway School of Pharmacy, University of Kent, Kent, UK; [b] Drug applied Research Centre, Tabriz University of Medical Sciences, Tabriz, Iran; [c] Current address: School of Pharmacy, Shahid Beheshti University, Tehran, Iran

*: Corresponding author at: Medway School of Pharmacy, Universities of Kent and Greenwich, Central Ave., Chatham Maritime, Kent ME4 4TB, UK; E-mail: t.ghafourian@kent.ac.uk; Tel: +44 1634 202952; Fax: +44 1634 883927

ABSTRACT

With the aim of solubility estimation in water, polyethylene glycol 400 (PEG) and their binary mixtures, Quantitative Structure – Property Relationships (QSPRs) were investigated to relate the solubility of a large number of compounds to the descriptors of the molecular structures. The relationships were quantified using linear regression analysis (with descriptors selected by stepwise regression) and Formal Inference – based Recursive Modeling (FIRM). The models were compared in terms of the solubility prediction accuracy for the validation set. The resulting regression and FIRM models employed a diverse set of molecular descriptors explaining crystal lattice energy, molecular size, and solute – solvent interactions. Significance of molecular shape in compound's solubility was evident from several shape descriptors being selected by FIRM and stepwise regression analysis. Some of these influential structural features, e.g. connectivity indexes and Balaban topological index, were found to be related to the crystal lattice energy. The results showed that regression models outperformed most FIRM models and produced higher prediction accuracy. However, the most accurate estimation was achieved by the use of a combination of FIRM and regression models. The results also showed that the use of melting point in regression models improves the estimation accuracy especially for solubility in higher concentrations of PEG. Aqueous or PEG/water solubilities can be estimated by these models with root mean square error of below 0.70.

KEYWORDS: solubility, PEG, Polyethylene glycol, QSPR, QSAR, cosolvent, ADME

1. Introduction

It is of tremendous benefit to the Pharmaceutical Industry to identify earlier in development those molecules that will eventually fail due to poor solubility, bioavailabity and pharmacokinetic issues. Therefore, properties such as Absorption, Distribution, Metabolism, and Excretion (ADME) need to be assessed early on during drug discovery. The properties also include aqueous solubility and permeability, as the two main determinants of intestinal absorption. Although, with the advent of new drug design technologies and high-throughput screening, many more hits and drug candidates have become available, properties of these candidates are becoming less favorable for development (Lipinski, 2002). Specifically, candidate drugs are becoming increasingly poorly water soluble (Lipinski, 2000). Since adequate aqueous solubility is a prerequisite for drug absorption from the gastrointestinal tract, it plays a major role in bioavailability of orally administered drugs. This has been recognized by the FDA (CDER, 2000) and European Medicines Agency (EMEA, 2007) on their Biopharmaceutics Classification System (BCS)-based biowaiver for waiving *in vivo* bioequivalence studies in favor of easier *in vitro* testing.

Rapid screening for solubility is possible through *in vitro* (experimental) and *in silico* (computational) screening approaches. The solubility of a compound depends on its degree of solvation in the solvent. The structural features in a solute molecule that improve the degree of solvation will result in a more soluble solute. This relationship between the molecular structure of compounds and their solubility has been extensively exploited and many different computational models have been constructed for the estimation of solubility. These include group contribution

methods such as the UNIFAC (Banerjee, 1985) and AQUAFAC (Lee et al., 1997), and statistically derived models based on molecular structures of compounds also known as Quantitative Structure – Property Relationships (QSPR) (Klopman and Zhu, 2001; Butina and Gola, 2003).

Yalkowsky's general solubility equation, GSE (Yalkowsky, 1999) is a well known model that simply employs two properties, melting point and octanol/water partition coefficient (log P), for the estimation of aqueous solubility of solids. On the other hand, more sophisticated QSPR models such as those using nonlinear statistical techniques have been proposed that rely solely on calculated molecular descriptors, without the need for the experimentally determined melting point (Palmer et al., 2007; Johnson and Zheng, 2006; Schroeter et al., 2007; Zhou et al., 2008; Huuskonen et al., 1998; Cheng and Merz, 2003; Wassvik et al., 2006; Hou et al., 2004; Bergström, 2005). Melting point is a measure of the crystal lattice energy that needs to be overcome during dissolution, hence the significance in solubility models. However, as the lattice energy depends on the strength of intermolecular interaction forces, it can be stipulated that the effect of these intermolecular forces on solubility can be accounted for by the calculated molecular descriptors such as polarizability, dipole moment and hydrogen bonding descriptors. In fact, several models have been proposed for the estimation of melting point using calculated descriptors (O'Boyle et al., 2008; Bergström et al., 2003), which show some level of predictivity, at least for a qualitative classification of the melting point (Bergström et al., 2003).

Apart from solubility in water, of a further importance is drug solubility in pharmaceutical co-solvents and their mixtures with water. Co-solvents are commonly used in liquid drug

formulations in order to increase the solubility of poorly water soluble drugs. Estimation of drug solubility in mixtures of water and co-solvents not only is very useful for drug formulators, but also in a comparative study with aqueous solubility, it can provide valuable understanding about the factors controlling the solubility phenomenon. Effects of volume fractions of a co-solvent in the binary mixtures with water have been modeled mathematically. One of the earlier models is that of Paruta and co-workers (1964) which described the solubility behavior using the dielectric constant of the mixed solvents. Other models have simply employed the volume fractions of the solvents and the solubility in pure solvents to describe the solubility in the solvent mixtures (Yalkowsky and Roseman, 1981; Jouyban-Gharamaleki et al., 1999). The need of these models for one or two solubility measurements (in one or both pure solvents) limits their applicability for the rapid estimation of solubility. On the other hand, QSPR is expected to provide a rapid estimation procedure for solubility in different solvents and solvent mixtures. Furthermore, using QSPR a quantitative comparison between molecular properties controlling solubility in different solvents can be achieved. The present investigation focused on the development of models relating molecular structures of solutes to the solubility in water, polyethyleneglycol 400 (PEG) and PEG/water binary solvent mixtures. Drug solubilities in these solvents were available through Rytting et al (2005). The models were developed using linear regression analyses and non-linear method of Formal Inference – based Recursive Modeling (FIRM). The structural properties responsible for solubility in different solvents were identified and comparisons were made between different solvents and the linear/ nonlinear models.

2. Materials and Methods

2.1. Solubility Data: The solubility dataset of Rytting et al (2005) was used in this study. The dataset consisted of equilibrium solubility of 122 compounds in 0%, 25%, 50%, and 75% (V/V) aqueous PEG and that of 94 compounds in pure PEG. Only the free forms of each drug were used in the solubility determination. The compounds represent a broad range of log P values (-2.4 to 7.5), molecular weights (111 to 614 Da) and melting points (53.5-360 °C). Melting points of the drugs were obtained from the literature and databases (Ryting et al., 2005; SRC, 2010; Wishart et al 2008). Melting point was not available for two compounds in the dataset, as they decompose before melting (see supporting material).

2.2. Structural descriptors: Structural descriptors were calculated using ACD-Labs LogD Suite, version 11 (Advanced Chemistry Development, Inc., Toronto ON, Canada) and Tsar 3D, version 3.3 (Accelrys Inc., USA). For each compound, 25 descriptors were obtained using Advanced Chemistry Development (ACD) Lab/ LogD Suite. These included logarithm of the octanol-water partition coefficient (log P), logarithm of apparent partition coefficient (log D) at different pH values of 1, 7.4 and 13, percentage weight of each atom type in the molecular structure, molar refractivity and density. After minimization of the molecular energies by COSMIC force field, a total of 90 descriptors were calculated for each compound using TSAR 3D software. The descriptors were deleted if the values for 98% of the compounds were identical. Furthermore, where there was a high intercorrelation between a pair of descriptors ($R > 0.99$), one of the descriptors was discarded.

2.3. Development of QSPR models:

a. FIRM analysis

FIRM is a non-linear method (a form of decision tree) that selects variables for classification of data. In this method a large set of data is split into subgroups based on important predictor variables (variable selection procedure). The response data are split by each variable into subgroups and a p-value is computed for each possible split. The p-value is the probability that the subgroups are homogeneous. The different possible splits are optimised for the lowest p-value, and the predictor variable with the lowest p-value is used to split the data into the optimal subgroups. The analysis stops when a subgroup can no longer be split.

To prepare the dataset for FIRM analysis, compounds were divided into three sets of training, test and validation with the ratio of 3:2:1 respectively. The procedure used for the allocation of compounds into groups ensured that each group contained a good spread of different ranges of aqueous solubility data. To this end, compounds were sorted according to their aqueous solubility value and from each set of 6 drugs, the first, the third, and the fifth were allocated into training set, the second and the sixth into test set and the fourth into the validation set. While the validation set remained the same, the remaining compounds (training and test sets) were randomly sampled 20 times to allocate compounds randomly into further 20 sets of training and test. Twenty-one FIRM models were built using the 21 training sets and the models were used to predict the solubility of compounds in test and validation sets. FIRM analyses were performed using TSAR 3D with solubility in various solvent systems as the dependent variable, and all of the calculated descriptors as independent variables. The predicted and experimental solubility data were used for the calculation of Root Mean Squared Error (RMSE) and the best FIRM models were selected based on the prediction accuracy for the test set.

$$RMSE = \sqrt{\frac{\sum(\log S_{obs} - \log S_{pred})^2}{n}} \qquad (1)$$

In equation 1, S_{pred} is the predicted solubility, and S_{obs} is the observed solubility.

b. Stepwise Regression Analysis

MINITAB Statistical Software (version 13) was used for stepwise regression analyses between solubility in different solvents and structural descriptors. The maximum p-value for a parameter to be included in the equations was set at 0.05 and the maximum number of parameters allowed in the equation was eight. The following statistical criteria of the models were noted: N the number of observations, R^2 the correlation coefficient, s the standard deviation, F the Fisher statistic and the p-value. Stepwise regression analyses were performed on the training set. The solubility of the validation set comprising 1/6 of the total number of compounds was calculated using the resulting regression equation. Calculated and experimental solubilities were used for the calculation of RMSE value using equation 1.

3. Results

3.1. FIRM Models

FIRM analysis was performed on drug solubility data in water, PEG and various water/ PEG mixtures. The analysis was performed on 20 randomly selected training sets and on the training set that was manually selected to cover all ranges of aqueous solubility. The best model was selected based on the RMSE value of the solubility predicted for the test sets. Presented in Figures 1 - 4 are the selected FIRM trees for the solubility of training and test set compounds in pure water, 25% PEG, 50% PEG, and 75% PEG, respectively. Table 1 gives a brief description

of the descriptors used in the selected FIRM models. FIRM was unable to partition the solubility of training set compounds in 100% PEG.

Apart from the selected FIRM trees, a second procedure for the prediction of solubility involved averaging of the estimated solubility by all the 21 FIRM models. The prediction of solubility by this method, consensus FIRM, will be compared with other models in Discussion section.

3.2. Regression Models

The regression equations resulting from stepwise regression of log solubility (log S) as the dependent variable and all the molecular descriptors and melting point (mp) as the independent variables resulted in equations 2-6 (Table 2). Stepwise regression was also performed with mp excluded from the analysis, resulting in equations 7-11 (Table 3). It must be noted that the p-values for the equations were less than 0.0005. The t-test p-values for the coefficients of all the descriptors in the QSPRs were less than 0.05. The descriptors used in equations 2-11 have been explained in Table 1.

4. Discussion

The seemingly simple process of dissolution of solid chemicals is a complex phenomenon which includes destruction of crystal lattice, creation of a cavity in the solvent to accommodate

the solute molecule, hydration of solute molecules and optimization of the 3D structure of dissolved molecules by intermolecular interactions of solutes with solvent molecules. In this study, solubility of solid compounds in water, PEG and water/PEG mixtures were analyzed in order to develop relationships with the molecular properties. Regression and FIRM models were developed using only calculated molecular properties. Moreover, in order to assess the value of melting point in the estimation accuracy of the QSPR models, this property was used as one of the descriptors of stepwise regression analysis and the resulting models were compared with models without mp in terms of the estimation accuracy and the selected descriptors.

The resulting FIRM and regression models employed a wide range of descriptors (Table 1) comprising calculated partition coefficients, atom and molecular attributes, molecular connectivity indexes, and fraction of compounds ionized at specific pH values. This should be advantageous for the prediction accuracy in comparison with models that are limited in terms of the range of the descriptors used. For example, Rytting et al (2004) used only molecular weight and volume, number of rotatable bonds, numbers of hydrogen bonding donor and acceptor groups, molecular density and radius of gyration, and Abraham et al (1999) employed only the five Abraham descriptors.

FIRM analysis used in this study has the advantage that it can take the nonlinear effects of structural properties into account (Hawkins et al., 1997; Blower et al., 2002) and has proven useful in classifying pharmaceutical data by discrete or continuous descriptors (Ghafourian and Cronin, 2006; Godden et al., 2003). For example, suppose hydrogen bonding groups aid aqueous solubility of a certain group of drugs to a certain extent. In this case FIRM can classify the

solubility data into several bins based on several ranges of the hydrogen bonding descriptor. However, this level of flexibility can result in over-fitting and poor generalization of the model. Although the FIRM tree starts with large number of compounds, successive classification into groups means that it will be using fewer data points at later stages of partitioning, increasing the probability of over-fitting. One solution to this problem resides in the construction of several differing FIRM trees and using the average of the predicted values (consensus modeling). Consensus QSPR models have been widely used to improve prediction abilities of the models (Santos and Hopfinger, 2008; Asikainen et al., 2004; Votano et al., 2004). However, consensus modeling adds to the complexity of the QSPR (Hewitt et al., 2007), rendering the interpretation difficult or even unmanageable.

In this investigation twenty-one FIRM trees were generated for randomly sampled training sets and the best tree was selected based on the prediction accuracy for the test set compounds. This procedure allows interpretation of the model and thereby provides some insight into the factors governing the process of dissolution. On the other hand, the average of calculated log solubility using these 21 trees was also calculated and compared with the selected tree in terms of the prediction accuracy. Moreover, two sets of regression models were constructed for solubility in each solvent system. These were the QSPR with or without incorporation of melting point in the descriptor list. In the discussion below, the prediction accuracy of different models will be examined and then the components of the models and the relation of the molecular descriptors with the solubility in different solvent systems will be discussed. The correlation matrix of all the selected descriptors is provided in Supporting Information.

4.1. Prediction accuracy of the models

Errors of solubility estimation using different methods have been presented in supporting material for all the drugs. The residuals of log solubility showed normal distributions with skewness values not greater than around 2. The RMSE values and percentages of drugs showing residuals below 0.5, below 1 or greater than 1.5 for all compounds and for the validation set have been presented in Table 4 and Table 5. In average, percentage of validation set compounds with predicted solubility within 1.0 log unit of the observed value is 88.9% for regression models (with mp), 91.2% for regression models (without mp), 82.1% for consensus FIRM and 80.4% for the selected FIRM models. These values are encouraging when comparing with the QSPR models presented by Rytting et al (2004) for this same dataset, where 78.1% of the predictions were within 1.0 log unit of the observed values. The QSPR model reported by Rytting et al involved splitting of the dataset into two groups of structurally similar compounds and development of separate regression equations for each group. They also compared the prediction accuracy of the QSPR with that of the classic log-linear model (Yalkowsky et al., 1972), where solubility of each compound in water is required for the estimation of solubility in the solvent mixtures. This procedure led to the solubility prediction of 84.2% of compounds within 1 log unit of the observed values. Therefore, the calculated regression models presented here (equations 7-11) are more accurate than the more strenuous log-linear method.

Similarly, RMSE values in Table 4 and 5 also show that the regression models provide the most accurate estimation of solubility in all the solvent systems. The RMSE values of the regression models with or without mp for all solvent systems are, respectively, 0.62 and 0.59 for

validation set and 0.61 and 0.65 for all drugs. This shows that melting point can comfortably be excluded from solubility estimation protocols, with only a modest reduction in the estimation accuracy. By comparing the reported statistics for the validation set, it can be seen that the calculated regression models outperform FIRM or consensus FIRM models in most solvent systems. It must be noted that for solubility in 25% PEG, FIRM model is exceptionally good with RMSE of 0.49 for the validation set. Consensus FIRM models are generally the third best in terms of RMSE for the validation set, followed by the selected FIRM models (Table 4). On the other hand, FIRM and consensus FIRM models show a better fit to the training set as it can be seen from the reported statistics for all the drugs (Table 5).

In comparing the error levels it must be born in mind that in a small number of cases, the descriptor values were not available for some drugs due to limitations of the software used for the calculation of the descriptors. Also, melting point was not available for two compounds as they decompose. Moreover, FIRM models failed to predict for those drugs in the validation set whose descriptor values fall outside the range defined in the model. From this perspective, the consensus FIRM models have the advantage that the predicted values are available for larger numbers of drugs Therefore the error is calculated for the remaining dataset (see the 'percentage of drugs calculated' in Table 5).

Table 4 indicates higher error levels for the prediction of solubility in water or pure PEG in comparison with the solvent mixtures. In fact, FIRM models could not be constructed for solubility in pure PEG. For majority of the drugs in the dataset, solubility is higher in PEG than in water. PEG has lower values of surface tension, dielectric constant and solubility parameter

than water (Yurquina et al., 2007) and when added to water such properties for the mixtures will be lower than those of the pure water. In order to describe the behavior of solvents and solvent mixtures, cohesive and adhesive forces between molecules have been considered. The Hildebrand – Scatchard equation (Hildebrand and Scott, 1950) employs solubility parameters of the solute (δ_B) and the solvent (δ_A) as measures of volume specific cohesion/ adhesion energy. The equation implies that the highest solubility occurs when solubility parameters of the solute and the solvent are equal (Stengele et al., 2001). Inspection of the plots between solubility and the solutes' melting point (Figure 5) shows that, unlike aqueous solubility, the solubility in PEG is highly controlled by the cohesion energy of the solute molecules reducing the contribution of the solute-solvent adhesion energy. On the other hand, aqueous solubility is more related to hydrophobicity than is the solubility in PEG (see graphs of solubility vs log P in Figure 5). This is probably the reason for the poor correlation of the aqueous solubility with the solubility in PEG ($R^2 = 0.04$).

Several chemicals showed high average estimation errors. These compounds in descending order of average absolute error were guanine, xanthine, folic acid, nalidixic acid and uric acid. The average solubility of these compounds were considerably lower than the average solubility of the whole dataset and, almost in all cases, their solubility values were overestimated. All these chemicals have high melting points or, in case of xanthine and uric acid, they decompose before melting. This may indicate an incomplete description of the crystal lattice for these compounds by the molecular descriptors used, or not enough weighting of melting point in the models using this descriptor. In fact, the average estimation error for high melting point chemicals with mp of

200.5 - 360 °C was found to be generally higher than that for low melting point compounds with mp of 53.5 - 192 °C (Table 6).

A consensus model based on the two different methodologies of linear regression and non-linear FIRM may result in more robust predictions, as the inclusion of FIRM may allow better prediction for some of the compounds for which the non linear behavior may be more pronounced. To examine this, the average predictions by regression (without mp) and the selected FIRM models were calculated. Comparing Table 5 and Table 7 shows that when comparing RMSE values for all drugs, the combined regression and FIRM models are more accurate than any individual models for aqueous solubility and solubility in 25% PEG. For solubility in 50% or 75% PEG, the regression models incorporating melting point (equations 4 and 5) are of superior or equal accuracy to the combined FIRM and regression models. Therefore, for future estimation of solubility in water and 25% PEG, a combination of FIRM and calculated regression models can be recommended; this is FIRM model in Figure 1 and equation 7 for aqueous solubility, and FIRM model in Figure 2 and equation 8 for solubility in 25% PEG. For solubility in 50%, 75% and 100% PEG, provided the availability of mp, equations 4, 5 and 6 (respectively) can be recommended. When mp for a drug is not available, a combination of FIRM and regression will be a suitable alternative for the estimation of solubility in 50% and 75% PEG; that is FIRM model in Figure 3 and equation 9 for solubility in 50% PEG, and FIRM model in figure 4 and equation 10 for solubility in 75% PEG. Equation 11 can be used for the estimation of solubility in pure PEG when mp is not available.

4.2. The Selected FIRM Models

Melting point was not used as a descriptor in FIRM analyses. Table 1 gives a brief description of all the selected descriptors. All the FIRM models for aqueous solubility including the selected model (Figure 1) classified the compounds by the log P value in the first step. This is expected from mechanistic point of view and also suggested by Yalkowsky (1999) in the General Solubility Equation (GSE). In the second step, J (Balaban topological index (Devillers and Balaban, 2000)) and MW (molecular mass) were selected for the classification. This shows the importance of molecular shape and size in the solubility process. J is a highly discriminating topological index whose values do not substantially increase with the molecular size and represents extended connectivity and the shape of molecules (Thakur et al., 2004). It appears that the group of compounds with low J values have a considerably higher average mp (236 °C) than the other group with average mp of 208 °C, which could explain their lower aqueous solubilities. The electronic parameter of SEI (sum of electrotopological state indexes) appeared in the third step of the classification. This is despite the important effect of electronic interactions (such as H-bonding) between a solute and water on the water solubility. This can be attributed to the fact that intermolecular electrostatic interactions can both reduce or increase solubility, depending on whether they are formed between the solute molecules leading to high crystal structure energy, and/ or between the solute and the solvent molecules resulting in the heat release. Moreover, electrostatic interactions can be formed intra-molecularly. Although there have been attempts to approximate the intra-molecular interactions for example by the use of product terms of hydrogen bonding acceptor and donor descriptors (Abraham and Le, 1999), the extent of such interactions can only be estimated by rigorous conformational analysis of the molecules.

In the FIRM tree (Figure 1) it can be seen that numbers of ethyl groups and heteroatoms have been used at later stages of partitioning. Ethyl groups are present in only 3 out of the 30 compounds in this bin. These compounds have a lower average solubility than those without ethyl groups, probably due to the hydrophobicity of the hydrocarbon groups. Contrary to the expectations, compounds with high number of heteroatoms (7 or 9) have lower average solubility. An inspection of this bin shows that several sulphonamides i.e. hydrochlorothiazide, hydroflumethiazide and acetazolamide, are amongst the compounds in this bin which has a higher average melting point (222 vs. 200) and molecular weight than the other bin. Finally, number of phenyl groups has been selected showing somehow contentious effect on aqueous solubility, as compounds with one phenyl group (n = 7) have higher solubility than those with 0, 2, 3, or 4 such groups (n = 18).

In the FIRM models for the solubility in water / co-solvent mixtures (Figures 2-4), log P was not selected in the early stages of partitioning. Lower polarity of the solvent mixtures as indicated by their reduced dielectric constant in comparison with water (Sengwa and Sankhla, 2007) is the probable cause of this, leading to the less negative effect of the solute lipophilicity on the solubility in such solvents.

For solubility in 25%PEG, Parachor (Pa) was the first descriptor selected by FIRM analysis. This descriptor represents the molecular size, with the partitioning showing the lower solubility of high molecular size drugs. For the large molecular size compounds, the next selected descriptor is log D_1, the distribution coefficient at pH 1, which is expectedly higher for the low-solubility drugs. For acidic drugs, the log D measured at pH 1 is expected to be higher than that

measured at basic pH values, and the opposite is true for the basic drugs, due to the lower percentage of an acidic drug ionized at pH 1 than that of a basic drug. Therefore, the lipophililic drugs that have been separated are more likely to be acidic as well. This expectation is confirmed by considering the types of drugs that reside in this bin that are mostly NSAIDS and different classes of steroids. The lipophilic, large molecular-size drugs have been finally partitioned by the number of six-membered aliphatic rings ($rings_{6aliph}$) which shows a nonlinear effect on the solubility. On the other hand, small compounds (Pa<513) have been partitioned according to their shape descriptor, J, with molecules containing high J values being more soluble in this solvent mixture, presumably due to the lower mp (average of 178 vs. 231 °C for low J group) as explained for aqueous solubility model. This trend was also the case for aqueous solubility (Figure 1). Compounds with high J values have been classified by V/SA (volume divided by the surface area). V/SA is larger for more spherical molecules (as they have minimum surface area for the volume) and it is smaller for planar or elongated compounds. According to this FIRM tree, such spherical molecules have a lower average solubility than the planar molecules. Both groups of compounds have conjugated planar rings which, in the high V/SA compounds, it is mostly attached to chlorine atoms and flexible/ branched chains rendering them less soluble; whereas in the low V/SA compounds, it is mostly attached to small rigid groups. Cytosine, caffeine, salicylic acid and benzoic acid are examples of low V/SA, while linuron, butylparaben, ibuprofen and 1,2,3-trichlorobenzene are examples of high V/SA drugs. This FIRM model shows good prediction accuracy as explained earlier, confirming the reliability of the model and the selected descriptors.

For solubility in 50% PEG, the descriptor $^{9}\chi_{ring}$ was selected by FIRM in the first step. This connectivity index has a value greater than zero, for compounds containing nine-membered rings. An example of such molecular structures can be seen in allopurinol, azathioprine and strychnine where a six-membered ring is fused to a five-membered ring (see Figure 6). Out of 102 compounds in the training/test sets, 30 contain this structural characteristic with a lower average solubility in this solvent. Based on the planar ring structures with several π-bonds, these compounds can be expected to have high melting points, indicating strong lattice energy. Indeed, the average melting point for the drugs with $^{9}\chi_{ring} > 0$ is 230 °C, whereas the remaining compounds have an average mp of 173 °C. Interestingly, majority of these chemicals have low J values (<2.2) which according to Figures 1 and 2 are grouped as having low solubility in water and 25% PEG, probably due to high melting points. Compounds containing 9-membered ring systems have only been partitioned once more and this has been done based on the ADME violations (AV) from Lipinski's rule of five (Lipinski et al., 1997) with only one compound (Diosgenin with log P of 5.84) showing a violation and having a very low solubility. The compounds in the other group have been classified according to the $^{1}\varkappa$ value with the compounds having higher $^{1}\varkappa$ value showing lower solubility. Information in $^{1}\varkappa$ is composed of molecular size and cyclicity (Hu et al., 2004). In the next step the larger molecules (with high $^{1}\varkappa$ values) have been classified according to their log P values. On the other hand, the group with smaller molecules has been classified according to a calculated descriptor, SEI/n (Sum of electrotopological state indexes divided by the number of heavy atoms), an indicator of the availability of electrons for electrostatic interactions (Hu et al., 2004). Finally the compounds with higher SEI/n have been grouped into those containing one aromatic ring and those containing zero or two such rings.

For solubility in 75% PEG, the number of 6-membered aliphatic rings was selected at the first step, with drugs having one or more such rings showing lower average solubility. These two groups of compounds have similar lipophilicities (average log P of 1.76 vs. 1.77), but significantly different average mp values of 167 vs. 222 °C for high and low solubility groups respectively. In the next step, compounds have been classified according to the $^{1}\chi$ value with the compounds with higher $^{1}\chi$ value showing lower solubility, which can be attributed to the large molecular size of such compounds. From the group of compounds with high $^{1}\chi$ values, three have been separated with relatively low atom-level molecular connectivity index, a descriptor related to the number of atoms (Hall and Kier, 2001). These three compounds have a lower solubility than the remaining compounds. Studies have shown that although the Kappa index represents significantly the molecular size, the size information of this index is different from that of the Chi index, as it combines the cyclicity information with the size (Hu et al., 2004).

For solubility in 100% PEG, FIRM was unable to split the dataset using any of the descriptors. In other words none of the descriptors were significant for the classification of PEG solubility data.

4.3. Linear Regression Models

The results of stepwise regression analysis showed that melting point was one of the first three descriptors of solubility models in all solvent systems (see equations 2-6 in Table 2). This finding is in agreement with the GSE and the hypothesis that the work required for the breakdown of

solute crystal structure is one of the most important contributors to the overall free energy of dissolution (Yalkowsky, 1999). When mp was not used in stepwise regression analysis, the resulting QSPRs had reduced R^2 values (compare equations in Table 1 and Table 2). Apart from melting point, QSPR equations 2-6 contain several other molecular descriptors which can be regarded as those describing other contributing factors such as the solute-solvent interaction and cavity formation in the solvent (Hermann, 1997).

Equation 2 shows that the first two descriptors selected by stepwise regression analysis are melting point and octanol/ water partition coefficient which, in accordance with the GSE, represent the effects of the two main factors, crystal lattice energy and solute-solvent interaction energies, respectively. Moreover, the work of cavity formation in the solvent for the solute molecule is represented by the first order molecular connectivity index, a size descriptor, with a negative coefficient indicating the effect of molecular size. This is in accordance with the model suggested by Meylan et al (1996) which, in addition to the first three descriptors of equation 2, incorporated 12 independent correction factors.

In equation 3 for the solubility in 25%PEG, melting point and octanol/ water partition coefficient are still the main descriptors. The third-order path molecular connectivity index is mostly an indicator of the molecular size and adjacency of branching (Hall and Kier, 2001) probably indicating the work required for cavity formation. FiB, the fraction of basic compounds ionized at pH 7.4, shows the favorable effect of the electrostatic interaction between the ionized solutes and water. The two connectivity indexes of 4^{th} order cluster and 10th order path (Hall and Kier, 2001) can indicate the higher solubility of specific molecular topologies having a relatively

high solubility despite the large molecular size. Examples of these molecules are strychnine and dexamethasone with large $^{10}\chi_p$, and ampicillin with high $^{4}\chi^{v}_{cluster}$.

Similarly, in equation 4 for the solubility in 50% PEG, mp, log P and FiB are selected. The other descriptors of the equation are zero-order valence connectivity index with a negative coefficient indicating the negative effect of molecular size and the number of phenyl groups (N_{Phenyl}) with positive coefficient describing certain type of the solute-solvent interaction energy for the compounds containing phenyl groups. As N_{Phenyl} is also a positive contributor to the solubility in 75% PEG (equation 5), it can be assumed that there is a favorable interaction between the phenyl group and the solvent, at higher concentrations of PEG. Contrary to the solubility in the PEG/ water solvent mixtures, the highly hydrophobic aromatic rings have been shown to reduce the aqueous solubility (Huuskonen et al., 2008).

In equation 5, a major change is observed in that the octanol /water partition coefficient is no longer significant. This follows the reducing trend of (absolute values of) log P coefficients from equation 2 to 4 with the reduction of water content of the solvent mixture. Number of phenyl rings (N_{phenyl}) and the hydrogen bonding acceptor atoms (N_{HA}) represent the favorable electrostatic interactions between the solute and the solvent molecules. Ratio of the number of flexible bonds to the total number of heavy atoms (F/N) shows a positive effect on the solubility in 75%PEG. The favorable effect of molecular flexibility on the aqueous solubility is also documented previously (Huuskonen et al., 2008; Bergström et al., 2002). Finally, the second order molecular shape index ($^{2}\kappa$) with a negative coefficient could be a shape descriptor or

simply an indicator of the lower solubility of compounds with larger molecular size (Hall and Kier, 2001).

In equation 6, number of six-membered aliphatic rings, number of four-membered rings, and the first order molecular connectivity index have negative coefficients indicating the negative effects of molecular size, and presence of four-membered and six-membered systems on the solubility in PEG. FIRM model 4 for solubility in 75% PEG (Figure 4) also indicated the lower solubility of the group of compounds containing six-membered rings, which could be attributed to the corresponding crystal lattice as explained before. Weight percentage of nitrogen atoms in the molecules is the other descriptor with a negative effect on PEG solubility. The combination of N% and the number of amino groups with negative and positive coefficients respectively can also be seen in equations 9 and 11. The polar hydrogen bonding amino group can be an obvious promoter of solubility in a polar solvent mixture. The negative coefficient of N% on the other hand shows the lower solubility of drugs containing a large number of non-amine nitrogen atoms such as that in xanthine, guanine and uric acid (see figure 6).

On the other hand, molecular descriptors in equations 7-11 are expected to represent the melting point as well as the solute-solvent interaction energy and the energy required for the creation of a cavity in the solvent. Molecular density (Dm), calculated by ACD labs software as molecular weight divided by molar volume, is the common descriptor in equations 7-10. Dm can be representing the solid state intermolecular interaction energy replacing the melting point descriptor in equations 2-5. Similar to equations 2-4, equations 7-9 involve partition coefficient with reducing coefficient. Log P is absent from equations 10 and 11. In equation 7, dipole

moment with a negative coefficient suggests involvement of electrostatic interaction between the solute molecules, and ninth order ring molecular connectivity index with a negative coefficient, as explained before, reveals the lower solubility of certain molecular structures containing two fused rings of six and five atoms, most probably due to their strong crystal lattice. Apparent partition coefficient in pH 7.4 indicates the lower solubility of lipophilic drugs especially those without acidic or basic groups. In equation 8, molecular connectivity index for five-membered rings could be an indicator of molecular size, ratio of the number of flexible bonds to the number of atoms is a descriptor of flexibility with a positive coefficient as seen previously in equation 5. Flexible molecules have lower melting points (Bergström et al, 2003) and this could be the reason for the higher solubility of these molecules. Equation 9 is very similar to equation 4 for the solubility in 50%PEG solvent mixture, with molecular density replacing melting point, and log D_1 and number of amino groups selected instead of FiB. Here negative coefficient of log D_1 indicates higher solubility of less lipophilic basic drugs (which have a lower apparent partition coefficient at pH 1 in comparison with acids or neutral drugs). This is clearly linked with FiB values. Likewise, in equation 10, apart from those descriptors that are also present in equation 5, number of six-membered aliphatic rings and ellipsoidal volume can be molecular size descriptors. The number of hydrogen bonding acceptor and amino groups with positive coefficients, and the number of nitrogen atoms with negative coefficient are explained earlier to be related to the solute-solvent interaction energy.

Finally, in equation 11, percentage weight of nitrogen atoms, number of six-membered aliphatic rings and number of four-membered rings are the same as equation 6. The other descriptors, number of hydrogen bonding donor ability, number of amino and hydroxyl groups

can represent the solute-solvent or the solute-solute interaction energies. In order to identify the descriptor(s) representing the solute crystal lattice energy, a regression analysis was performed which incorporated melting point in addition to the descriptors of equation 11 as the independent variables. The result showed that number of hydrogen bonding donor groups (N_{HD}) and number of hydroxyl groups (N_{OH}) were no longer significant. This shows that N_{HD} and N_{OH} describe crystal lattice energy in the absence of melting point or molecular density.

Conclusion

In this investigation non-linear and linear methods of FIRM and stepwise regression were used for the development of QSPR models for solubility in water, PEG or PEG/water mixtures. The results of stepwise regression analysis showed that melting point was a significant contributor to solubility estimation in all solvent systems. However, with the exclusion of melting point the resulting models were still able to estimate the solubility of the external validation set with only a slight decrease in the estimation accuracy. The accuracy of these models (with melting point excluded) were better than the log-linear model (Yalkowsky et al., 1972) which requires the solubility of each drug in water as the input for solubility estimation in the binary mixtures.

Regression models outperformed FIRM or consensus FIRM models for solubility estimation in most solvent systems. However it was shown that an estimation made by a combination of FIRM and regression models gives the most accurate estimation of solubility in water and 25% PEG. Estimation of solubility in higher concentrations of PEG is most accurate using mp-included regression models followed by a combination of FIRM and calculated regression models.

Stepwise regression analysis, in comparison with FIRM, employs a more diverse set of descriptors. With only a few exceptions, the descriptors selected by stepwise regression analysis are different from those selected by FIRM, despite the fact that they are both explaining similar effects of crystal lattice, lipophilicity, molecular size and solute-solvent interactions on solubility. This is due to the linear nature of regression in comparison with non-linear, classification type approach of FIRM. Log P is one of the descriptors that is selected by both FIRM and stepwise regression for solubility in water and lower concentrations of PEG.

For water solubility, the first property used for classification by FIRM is lipophilicity, followed by size and shape descriptors, with the latter reflecting the crystal lattice energy, and then electrostatic parameters. The trend changes for the solubility in PEG/water solvent mixtures with size and shape descriptors becoming more prominent (selected in earlier stages of classification in Figures 2-4). This could be due to higher significance of crystal lattice energy in non-aqueous solubility than in aqueous solubility, as the shape descriptors such as J and $^{9}\chi_{ring}$, were found to be related to the melting point of the compounds.

All regression models employed at least one descriptor for crystal lattice strength. Melting point is the descriptor of choice for crystal energy in regression models. When melting point is not used, the resulting regressions employ molecular density, $^{9}\chi_{ring}$, flexibility descriptor (F/N), N%, or in case of solubility in pure PEG, the number of hydrogen bonding donor groups and number of six-membered aliphatic rings to account for solid state energies. Similar to the FIRM models, the importance of lipophilicity descriptor in regression models declines with the

increasing fraction of PEG in the solvent mixture. The size descriptors are present in all the solubility models (FIRM and regression); these are molecular connectivity indexes, molecular weight, and Kappa indexes. Electrostatic descriptors are also evident in the regression and FIRM models. Examples are fraction of compounds ionized as base (FiB), the number of hydrogen bonding donor and acceptor groups and number of amino groups. Electrostatic features of a molecule can lead to increased or reduced solubility, as they control both favorable solute – solvent interactions and the work of crystal structure break down related to solute – solute interactions. Specific molecular features such as presence of 6-membered rings or fused five- and six-membered rings appeared in some models, indicating low solubility of the compounds containing such structures. These features can be linked to crystal energy as compounds with these characteristics show high melting points as well.

Supporting Information Available

1. Table of observed and predicted solubility values in different solvent systems and the estimation errors for training, test and validation sets
2. The correlation matrix for the molecular descriptors selected by FIRM and stepwise regression

References

Abraham, M.H., Le, J., 1999. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. J. Pharm. Sci. 88, 868-880.

Asikainen, A.H., Ruuskanen, J., Tuppurainen, K.A., 2004. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. J. Environ. Sci. Tech. 38, 6724-6729.

Banerjee, S., 1985. Calculation of water solubility of organic compounds with UNIFAC-derived parameters. Environ. Sci. Technol. 19, 369-370.

Bergström, C.A.S., 2005. In silico predictions of drug solubility and permeability: Two rate-limiting barriers to oral drug absorption. Basic Clin. Pharmacol. Toxicol. 96, 156-161.

Bergström, C.A.S., Norinder, U., Luthman, K., Artursson, P., 2002. Experimental and computational screening models for prediction of aqueous drug solubility. Pharm. Res. 19, 182-188.

Bergström, C.A.S., Norinder, U., Luthman, K., Artursson, P., 2003. Molecular Descriptors influencing melting point and their role in classification of solid drugs, J. Chem. Inf. Comput. Sci., 43, 1177- 1185.

Blower, P., Fligner, M., Verducci, J., Bjoraker, J., 2002. On combining recursive partitioning and simulated annealing to detect groups of biologically active compounds. J. Chem. Inf. Comput. Sci. 42, 393-404.

Butina, D., Gola, J.M.R., 2003. Modeling Aqueous Solubility. J. Chem. Inf. Comput. Sci. 43, 837-841.

CDER, 2000. (Center for Drug Evaluation and Research), Guidance for Industry, Rockville, MD: CDER/FDA.

Cheng, A., Merz, K.M., 2003. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure–property relationships. J. Med. Chem. 46, 3572-3580.

Devillers, J., Balaban, A.T., 2000. Topological Indices and Related Descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, pp. 117-119.

EMEA, 2007. Committee for Medicinal Products for Human Use, Concept Paper on BCS-Based Biowaiver, EMEA, London, EMEA/CHMP/EWP/213035/2007.

Ghafourian, T., Cronin, T.D.M., 2006. The effect of variable selection on nonlinear modelling of oestrogen receptor binding. QSAR Comb. Sci. 25, 824-835.

Godden, J.W., Furr, J.R., Bajorath, J., 2003. Recursive median partitioning for virtual screening of large databases. J. Chem. Inf. Comput. Sci. 43, 182-188.

Hall, L.H., Hall, L.M., 2005. QSAR modeling based on structure-information for properties of interest in human health. SAR QSAR Environ. Res. 16, 13–41.

Hall, L.H., Kier, L.B., 2001. Issues in representation of molecular structure: the development of molecular connectivity. J. Mol. Graph. Model. 20, 4–18.

Hawkins, D.M., Young, S.S., Rusinko, A., 1997. Analysis of a large structure-activity data set using recursive partitioning. Quant. Struct.-Act. Relat. 16, 296-302.

Hermann, R.B., 1977. Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions, Proc. Natl. Acad. Sci. USA 74, 4144-4145.

Hewitt, M., Cronin, M.T.D., Madden, J.C., Steven, J. Enoch., 2007. Consensus QSAR models: Do the benefits outweigh the complexity? J. Chem. Inf. Model. 47, 1460-1468.

Hildebrand, J.H., Scott, R.L., 1950. The solubility of nonelectrolytes, 3rd ed., Reinhold, New York.

Hou, T.J., Xi, K., Zhang, W., Xu, X.J., 2004. ADME evaluation in drug discovery. 4. prediction of aqueous solubility based on atom contribution approach. J. Chem. Inf. Comput. Sci. 44, 266-275.

Hu, Q.N., Liang, Y.-Z., Yin, H., Peng, X.-L., Fang, K.-T., 2004. structural interpretation of the topological index. 2. the molecular connectivity index, the kappa index, and the atom-type E-state index. J. Chem. Inf. Comput. Sci. 44, 1193-1201.

Huuskonen, J., Livingstone, D.J., Manallack, D.T., 2008. Prediction of drug solubility from molecular structure using a drug-like training set. SAR QSAR Environ. Res. 19, 191–212.

Huuskonen, J., Salo, M., Taskinen, J., 1998. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. J. Chem. Inf. Comput. Sci. 38, 450-456.

Johnson, S.R., Zheng, W., 2006. Recent progress in the computational prediction of aqueous solubility and absorption. AAPS J. 8, 4 (http://www.aapsj.org).

Jouyban-Gharamaleki, A., Valaee, L., Barzegar-Jalali, M., Clark, B.J., Acree, W.E., 1999. Comparison of various cosolvency models for calculating solute solubility in water-cosolvent mixtures. Int. J. Pharm. 177, 93-101.

Karthikeyan, M., Glen, R.C., Bender, A., 2005. General melting point prediction based on a diverse compound data set and artificial neural networks. J.Chem. Inf. Comput. Sci. 45, 581-590.

Klopman, G., Zhu, H., 2001. Estimation of the aqueous solubility of organic molecules by the group contribution approach. J.Chem. Inf. Comput. Sci. 41, 439-445.

Lee, Y.C., Pinsuwan, S., Yalkowsky, S.H., 1997. A comparison of AQUAFAC group q-values to their corresponding CLOGP f-values. Chemosphere. 35, 775-782.

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

Lipinski, C.A., 2002. Poor aqueous solubility-an industry wide problem in drug delivery. Am. Pharm. Rev. 5, 82–85.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Del. Rev. 23, 3–25.

Meylan, W.M., Howard, P.H., Boethling, R.S., 1996. Improved method for estimating water solubility from octanol water partition coefficient. Environ. Toxicol. Chem. 15, 100-106.

Modarresi, H., Dearden, J.C., Modarress, H., 2006. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. J. Chem. Inf. Model. 46, 930-936.

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

Palmer, D.S., O'Boyle, N.M., Glen, R.C., Mitchel, J.B.O., 2007. Random forest models to predict aqueous solubility. J. Chem. Inf. Model. 47, 150-158.

Paruta, A.N., Sciarrone, B.J., Lordi, N.G., 1964. Solubility of salicylic acid as a function of dielectric constant. J. Pharm. Sci. 53, 1349-1353.

Rytting, E., Lentz, K.A., Chen, X.Q., Qian, F., Venkatesh, S., 2004. A quantitative structure – property relationship for predicting drug solubility in PEG 400/ water cosolvent systems. Pharm. Res. 21, 237-244.

Rytting, E., Lentz, K.A., Chen, X.Q., Qian, F., Venkatesh, S., 2005. Aqueous and co-solvent solubility data for drug-like organic compounds. AAPS J. 7, Article 10. (http://www.aapsj.org).

Santos, O.A., Hopfinger, A.J., 2008. Combined 4D-fingerprint and clustering based membrane-interaction QSAR analyses for constructing consensus Caco-2 cell permeation virtual screens. J. Pharm. Sci. 97, 566-583.

Schroeter, T.S., Schwaighofer, A., Mika, S., Ter Laak, A., Suelzle, D., Ganzer, U., Heinrich, N., Mueller, K.R., 2007. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. J. Comput. Aided Mol. Des. 21, 485-498.

Sengwa, R.J., Sankhla, S., 2007. Characterization of heterogeneous interaction in binary mixtures of ethylene glycol oligomer with water, ethyl alcohol and dioxane by dielectric analysis. J. Mol. Liq. 130, 119–131.

SRC (Syracuse Research Corporation), 2010. Interactive PhysProp Database, website: http://www.syrres.com/what-we-do/databaseforms.aspx?id=386, accessed on March 2010.

Stengele, A., Stephanie, R., Leuenberger, H., 2001. A novel approach to the characterisation of polar liquids Part 1: pure liquids, Int. J. Pharm. 225, 123-134.

Thakur, A., Thakur, M., Khadikar, P.V, Supuran, C.T., Sudele, P., 2004. QSAR study on benzenesulphonamide carbonic anhydrase inhibitors: topological approach using Balaban index. Bioorg. Med. Chem. 12, 789–793.

Votano, J.R., Parham, M., Hall, L.H., Kier, L.B., Oloff, S., Tropsha, A., Xie, Q., Tong, W., 2004. Three new consensus QSAR models for the prediction of Ames genotoxicity. Mutagenesis 19, 365-77.

Wassvik, C.M., Holmen, A.G., Bergstrom, C.A.S., Zamora, I., Artursson, P., 2006. Contribution of solid-state properties to the aqueous solubility of drugs. Eur. J. Pharm. Sci. 29, 294-305.

Wishart D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M., 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res., 36 (Database issue), D901-6. The DrugBank database website accessed on March 2010.

Yalkowsky, S.H., 1999. Solubility and Solubilization in Aqueous Media. Oxford University Press, New York.

Yalkowsky, S.H., Flynn, G.L., Amidon, G.L., 1972. Solubility of nonelectrolytes in polar solvents. J. Pharm. Sci. 61, 983-984.

Yalkowsky, S.H., Roseman, T., 1981. Solubilization of Drugs by Cosolvents. Marcel Dekker, New York, pp. 91- 134.

Yurquina, A., Manzur, M.E., Brito, P., Manzo, R., Molina, M.A., 2007. A. Physicochemical studies of acetaminophen in Water-PEG 400 systems. J. Mol. Liq. 133, 47–53.

Zhou, D.S., Alelyunas, Y., Liu, R.F., 2008. Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility. J. Chem. Inf. Model. 48, 981-987.

Table 1. Descriptors selected by the models

Descriptor	Description	Number of times selected	
		FIRM	Stepwise Regression
$^{1}\kappa$	1rd order kappa shape index	2	
$^{2}\kappa$	2rd order kappa shape index		1
$^{0}\chi$	0rd order molecular connectivity index	1	
$^{1}\chi$	1st order molecular connectivity index		2
$^{3}\chi^{v}_{p}$	3rd order valence path molecular connectivity index		2
$^{4}\chi^{v}_{c}$	4rd order valence cluster molecular connectivity index	1	
$^{5}\chi_{ring}$	5th order ring molecular connectivity index		1
$^{9}\chi_{ring}$	9th order ring molecular connectivity index	1	1
$^{10}\chi_{p}$	10th order path molecular connectivity index		
AV	ADME Violations according to Lipinski's rule of five	1	
Dm	Molecular density		4
F/N	Flexible bonds/number of heavy atoms		2
FiB	Fraction of compound ionized as base ($1/(1+10^{7.4\text{-}pKa_{base}})$)		1
J	Balaban Topological index	2	
Log D_{1}	Apparent partition coefficient at pH 1	1	1
Log $D_{7.4}$	Apparent partition coefficient at pH7.4		1
Log P	Octanol-water partition coefficient	2	6
mp	Melting point measured in °C		5
MW	Molecular weight	1	
N%	Percentage of nitrogen atom		3
N_{Amino}	Number of amino groups		3
N_{Ethyl}	Number of ethyl groups	1	1
N_{HA}	Number of hydrogen bond acceptor groups		2
N_{HD}	Number of hydrogen bonding donor groups		1
N_{hetero}	Number of Heteroatom	1	
N_{N}	Number of nitrogen atoms		1
N_{OH}	Number of hydroxyl groups		1
N_{Phenyl}	Number of Phenyl groups	1	2
Pa	Parachor	1	
$Rings_{4}$	Number of 4 membered rings		2
$Rings_{6ali}$	Number of 6 membered aliphatic rings	2	3
$Rings_{6arom}$	Number of 6 -membered aromatic rings	1	
SEI	Sum of electro topological state indexes	1	
V/SA	Volume/Surface area	1	
V_{Ellip}	Ellipsoidal Volume; the volume defined by moments of inertia		1
SEI/n	Sum of electrotopological state indexes / number of heavy atoms	1	
μ	Dipole moment calculated by the AM1 Hamiltonian		1

Table 2. QSPRs obtained from stepwise regression analysis using all the molecular descriptors and melting point as the independent variables

Equation No.	Equation	N	R^2	s	F
2	Log S (Water) = -0.054 - 0.00674 mp - 0.596 Log P - 0.0906 $^{1}\chi$	119	0.69	0.73	86.2
3	Log S (25%PEG) = 0.323 - 0.385 Log P - 0.00639 mp - 0.293 $^{3}\chi^{v}_{p}$ + 0.327 FiB + 3.17 $^{4}\chi^{v}_{c}$ + 0.412 $^{10}\chi_{p}$	119	0.64	0.63	33.5
4	Log S (50%PEG) = 0.426 - 0.344 Log P - 0.00584 mp – 0.0717 $^{0}\chi^{v}$ + 0.365 N_{Phenyl}	119	0.59	0.59	40.7
5	Log S (75%PEG) = 0.591 + 0.457 N_{Phenyl} - 0.00509 mp - 0.422 $^{2}\kappa$+ 0.201 N_{HA} + 3.76 F/N	120	0.54	0.60	27.2
6	Log S (100%PEG) = 0.810 - 0.00493 mp - 0.246 $Rings_{6ali}$ - 2.03 $Rings_{4}$ - 0.0182 N% - 0.0536 $^{1}\chi$	92	0.71	0.46	43.0

N: number of compounds, R2: the square of the correlation coefficient, s: standard deviation, F: Fisher statistic.

Table 3. QSPRs obtained from stepwise regression analysis using all the molecular descriptors with the exclusion of melting point as the independent variables

Equation No.	Equation	*N*	R^2	*S*	F
7	Log S (Water) =0.619 - 1.48 Dm- 0.504 log P - 32.7 $^{9}\chi_{ring}$ - 0.176 Log $D_{7.4}$ - 0.103 μ	121	0.69	0.74	51.1
8	Log S (25%PEG) =0.238 - 1.32 Dm-0.432 log P - 4.88 $^{5}\chi_{ring}$ - 0.569 N_{Ethyl} + 2.35 F/N	121	0.58	0.68	31.8
9	Log S (50%PEG) = 0.823 - 0.846 Dm- 0.176 log P - 0.182 $^{3}\chi^{v}$ - 0.0399 N% - 0.0999 log D_1 + 0.341 N_{Amino}	121	0.52	0.65	20.6
10	Log S (75%PEG) =0.504 -0.444 $Rings_{6ali}$ - 0.184 N_N -0.000445 V_{Ellip} - 0.969 Dm+ 0.129 N_{HA} + 0.241 N_{Amino}	122	0.51	0.65	20.2
11	Log S (100%PEG) = 0.004 - 0.0366 N% - 0.433 $Rings_{6ali}$ - 1.79 $Rings_4$- 0.4478 N_{HD} + 0.435 N_{Amino} + 0.232 N_{OH}	95	0.69	0.54	32.2

N: number of compounds, R2: the square of the correlation coefficient, s: standard deviation, F: Fisher statistic.

Table 4. Comparison of different models for the prediction of solubility of validation set drugs in different solvents validation set

	Percentage of drugs with residual in log units			RMSE
	<±0.5	<±1.0	>±1.5	
Log S (water)				
Regression with mp	63.2	78.9	5.3	0.71
Regression	57.9	89.5	0.0	0.61
Consensus FIRM	36.8	68.4	15.8	0.93
FIRM	28.6	57.1	14.3	1.03
Log S (25% PEG)				
Regression with mp	63.2	89.5	5.3	0.62
Regression	57.9	100	0.0	0.55
Consensus FIRM	45.0	90.0	5.0	0.70
FIRM	60.0	100	0.0	0.49
Log S (50% PEG)				
Regression with mp	84.2	94.7	0.0	0.44
Regression	73.7	94.7	0.0	0.51
Consensus FIRM	60.0	85.0	5.0	0.68
FIRM	66.7	86.7	0.0	0.62
Log S (75% PEG)				
Regression with mp	63.2	94.7	5.3	0.52
Regression	75.0	85.0	5.0	0.59
Consensus FIRM	65.0	85.0	5.0	0.68
FIRM	50.0	77.8	0.0	0.68
Log S (100%PEG)				
Regression with mp	80.0	86.7	6.7	0.80
Regression	80.0	86.7	6.7	0.71

Table 5. Comparison of different models for the calculation of solubility of all drugs in different solvents

	Percentage of drugs with residual in log units			RMSE	% drugs calculated
	<±0.5	<±1.0	>±1.5		
Log S (water)					
Regression with mp	55.5	84.9	4.2	0.72	97.5
Regression	47.1	82.6	3.3	0.72	99.2
Consensus FIRM	54.5	85.1	5.8	0.73	99.2
FIRM	53.0	80.9	6.1	0.75	94.3
Log S (25% PEG)					
Regression with mp	68.1	92.4	2.5	0.61	97.5
Regression	49.6	87.6	1.7	0.67	99.2
Consensus FIRM	55.7	90.2	1.6	0.62	100
FIRM	64.6	92.0	1.8	0.57	92.6
Log S (50% PEG)					
Regression with mp	68.9	92.4	0.8	0.57	97.5
Regression	52.9	91.7	1.7	0.64	99.2
Consensus FIRM	60.7	91.8	1.6	0.60	100
FIRM	60.9	87.8	2.6	0.68	94.3
Log S (75% PEG)					
Regression with mp	65.5	94.1	3.4	0.59	97.5
Regression	55.7	88.5	3.3	0.63	100
Consensus FIRM	63.1	91.0	2.5	0.62	100
FIRM	59.0	88.0	6.0	0.67	95.9
Log S (100%PEG)					
Regression with mp	73.1	94.6	1.1	0.55	97.9
Regression	68.4	92.6	2.1	0.58	100

Table 6. RMSE values for the high and low melting point drugs in the dataset with 54 and 68 drugs in each group, respectively.

	RMSE	
	high mp	low mp
Log S (water)		
Regression with	0.71	0.73
mp Regression	0.75	0.70
Consensus FIRM	0.67	0.77
FIRM	0.74	0.76
Log S (25% PEG)		
Regression with	0.68	0.56
mp Regression	0.74	0.60
Consensus FIRM	0.67	0.57
FIRM	0.62	0.53
Log S (50% PEG)		
Regression with	0.60	0.55
mp Regression	0.73	0.55
Consensus FIRM	0.69	0.52
FIRM	0.83	0.54
Log S (75% PEG)		
Regression with	0.61	0.57
mp Regression	0.64	0.62
Consensus FIRM	0.76	0.48
FIRM	0.81	0.53
Log S (100%PEG)		
Regression with	0.64	0.47
mp Regression	0.65	0.51

Table 7. RMSE of log solubility predicted by combined regression and FIRM models.

	RMSE	
	Validation set	All drugs
Log S (water)	0.70	0.63
Log S (25% PEG)	0.49	0.52
Log S (50%	0.54	0.61
PEG) Log S	0.62	0.59
(75% PEG)		

Figure Captions

Figure 1. The selected FIRM tree for the classification of drugs into different aqueous solubility groups in water; N is the number of compounds, MLogS is the mean log solubility.

Figure 2. The selected FIRM tree for the classification of drugs into different solubility groups in 25% PEG; N is the number of compounds, MLogS is the mean log solubility

Figure 3. The selected FIRM tree for the classification of drugs in different solubility groups in 50% PEG; N is the number of compounds, MLogS is the mean log solubility.

Figure 4. The selected FIRM tree for the classification of drugs into different solubility groups in 75% PEG; N is the number of compounds, MLogS is the mean log solubility.

Figure 5. Plots of aqueous solubility and solubility in PEG 400 against melting point (mp) or log P.

Figure 6. Molecular structures of drugs with fused 5- and 6-membered rings

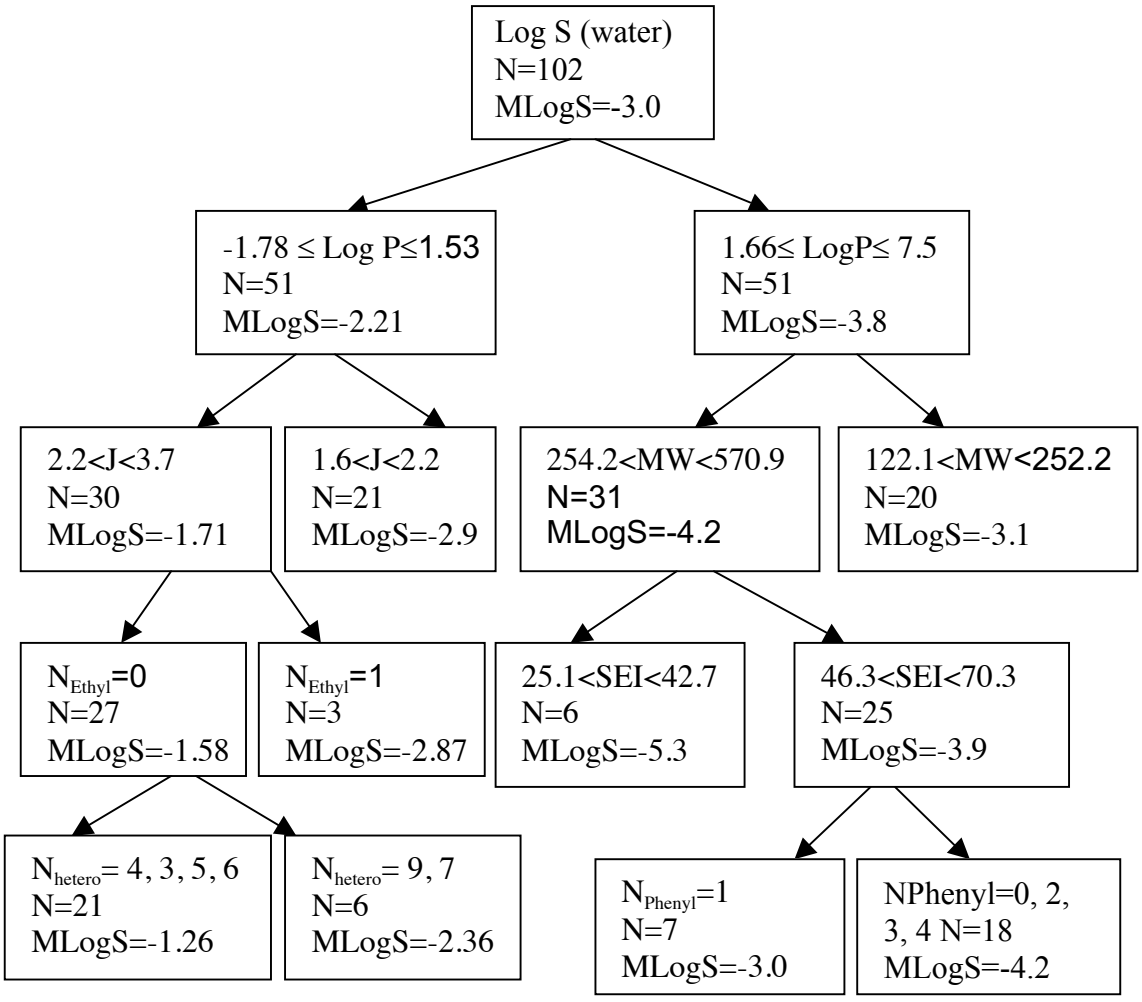
Log S (water)
N=102
MLogS=-3.0
-1.78 ≤ Log P≤1.53
N=51
MLogS=-2.21
1.66≤ LogP≤ 7.5
N=51
MLogS=-3.8
2.2<J<3.7
N=30
MLogS=-1.71
1.6<J<2.2
N=21
MLogS=-2.9
254.2<MW<570.9
N=31
MLogS=-4.2
122.1<MW<252.2
N=20
MLogS=-3.1
N_Ethyl=0
N=27
MLogS=-1.58
N_Ethyl=1
N=3
MLogS=-2.87
25.1<SEI<42.7
N=6
MLogS=-5.3
46.3<SEI<70.3
N=25
MLogS=-3.9
N_hetero= 4, 3, 5, 6
N=21
MLogS=-1.26
N_hetero= 9, 7
N=6
MLogS=-2.36
N_Phenyl=1
N=7
MLogS=-3.0
NPhenyl=0, 2, 3, 4 N=18
MLogS=-4.2

Figure 1

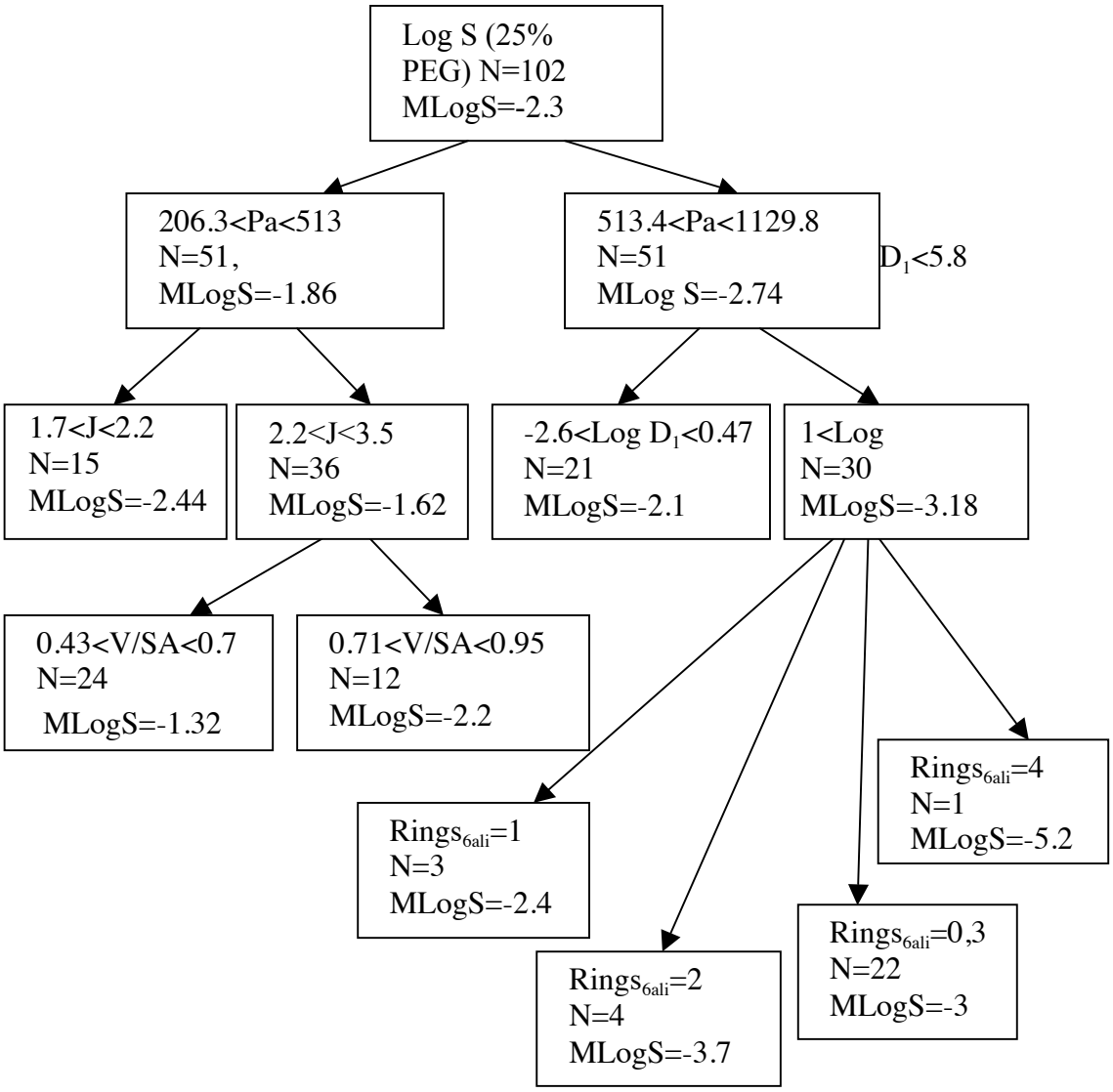
Log S (25% PEG) N=102 MLogS=-2.3
206.3<Pa<513 N=51, MLogS=-1.86
513.4<Pa<1129.8 N=51 MLog S=-2.74
D1<5.8
1.7<J<2.2 N=15 MLogS=-2.44
2.2<J<3.5 N=36 MLogS=-1.62
-2.6<Log D1<0.47 N=21 MLogS=-2.1
1<Log N=30 MLogS=-3.18
0.43<V/SA<0.7 N=24 MLogS=-1.32
0.71<V/SA<0.95 N=12 MLogS=-2.2
Rings6ali=4 N=1 MLogS=-5.2
Rings6ali=1 N=3 MLogS=-2.4
Rings6ali=0,3 N=22 MLogS=-3
Rings6ali=2 N=4 MLogS=-3.7

Figure 2

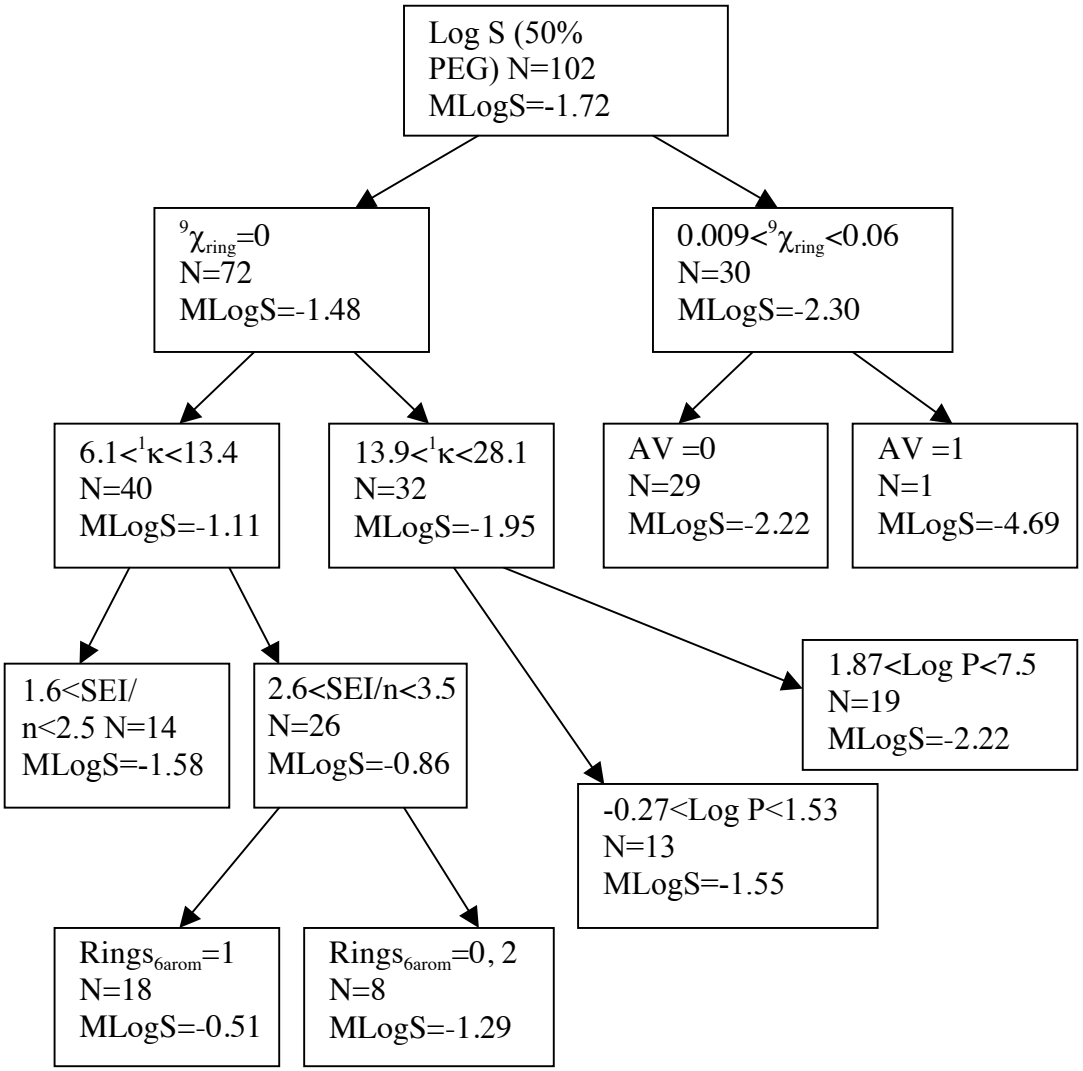
Log S (50% PEG) N=102 MLogS=-1.72
$^{9}\chi_{ring}$=0 N=72 MLogS=-1.48
0.009<$^{9}\chi_{ring}$<0.06 N=30 MLogS=-2.30
6.1<$^{1}\kappa$<13.4 N=40 MLogS=-1.11
13.9<$^{1}\kappa$<28.1 N=32 MLogS=-1.95
AV =0 N=29 MLogS=-2.22
AV =1 N=1 MLogS=-4.69
1.6<SEI/n<2.5 N=14 MLogS=-1.58
2.6<SEI/n<3.5 N=26 MLogS=-0.86
1.87<Log P<7.5 N=19 MLogS=-2.22
-0.27<Log P<1.53 N=13 MLogS=-1.55
$Rings_{6arom}$=1 N=18 MLogS=-0.51
$Rings_{6arom}$=0, 2 N=8 MLogS=-1.29

Figure 3.

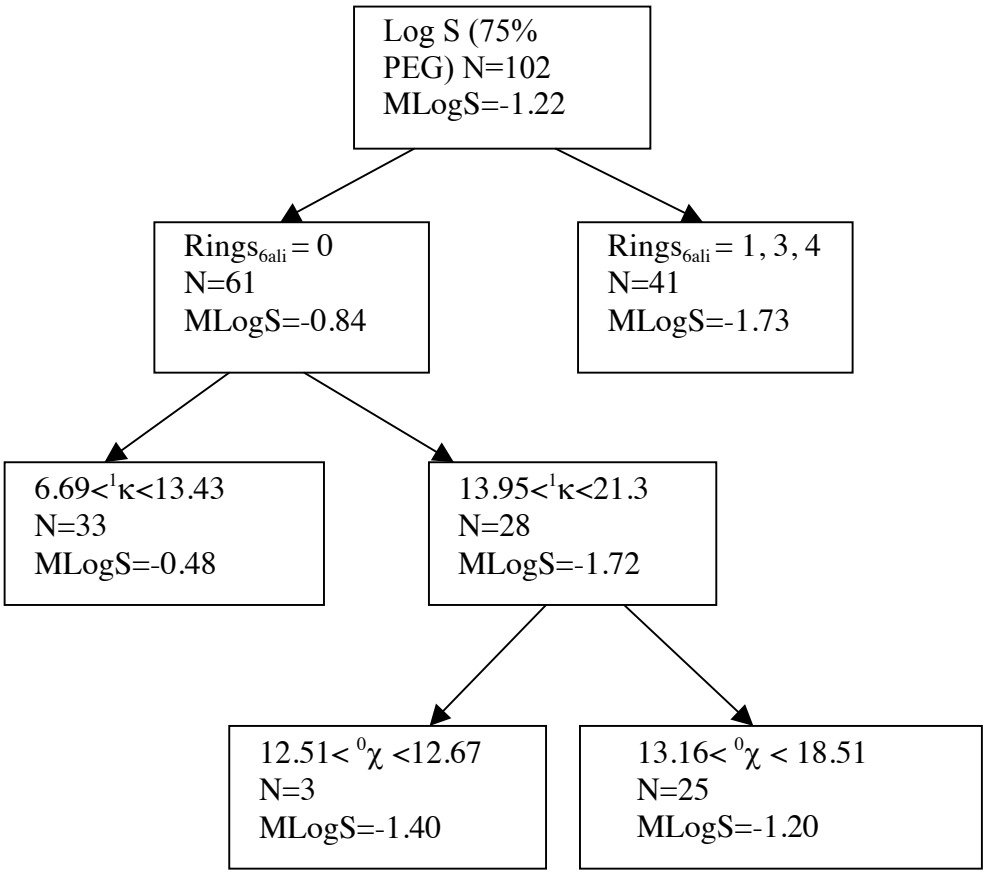
Log S (75% PEG) N=102 MLogS=-1.22
Rings6ali = 0 N=61 MLogS=-0.84
Rings6ali = 1, 3, 4 N=41 MLogS=-1.73
6.69<1κ<13.43 N=33 MLogS=-0.48
13.95<1κ<21.3 N=28 MLogS=-1.72
12.51< 0χ <12.67 N=3 MLogS=-1.40
13.16< 0χ < 18.51 N=25 MLogS=-1.20

Figure 4.

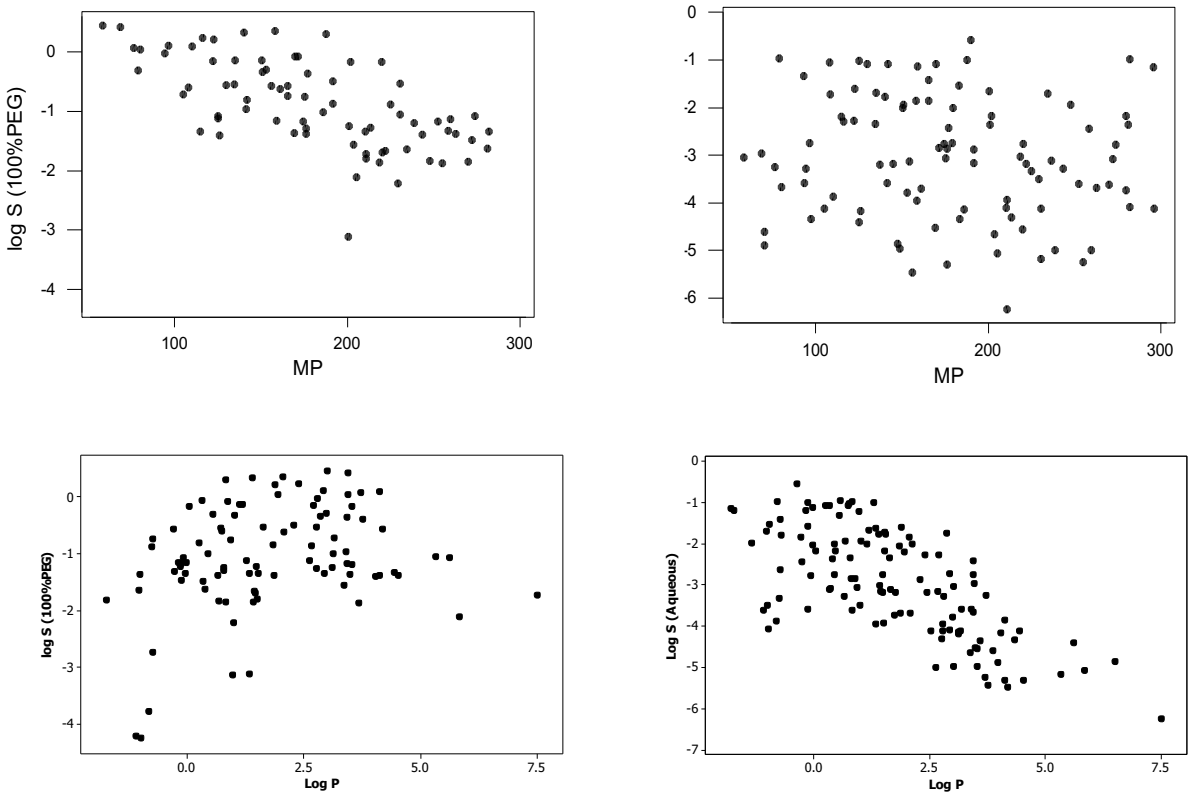

Figure 5.

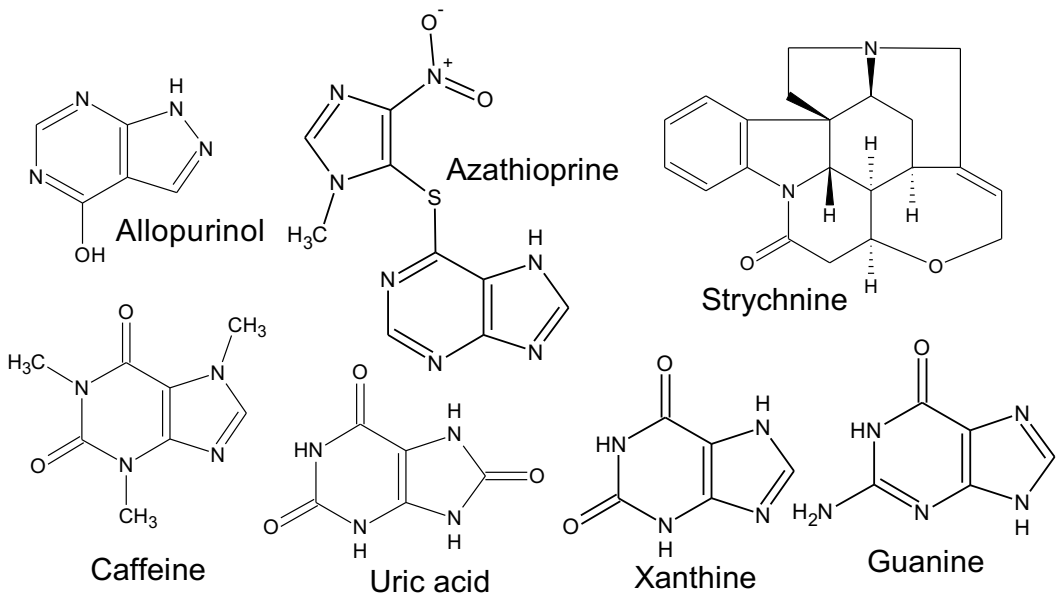
Allopurinol
Azathioprine
Strychnine
Caffeine
Uric acid
Xanthine
Guanine

Figure 6.