JOURNAL OF CHEMICAL INFORMATION AND MODELING
Article
pubs.acs.org/jcim

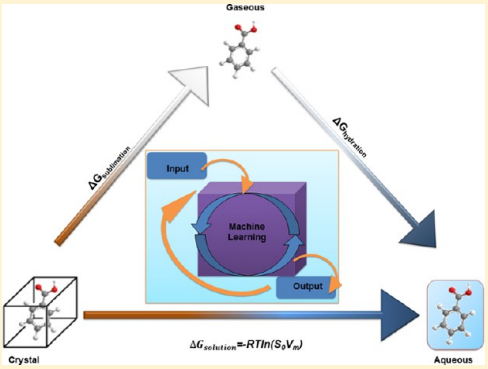
Uniting Cheminformatics and Chemical Theory To Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules

James L. McDonagh,[†,‡] Neetika Nath,[†,‡] Luna De Ferrari,[†,‡] Tanja van Mourik,[‡] and John B. O. Mitchell*[,†,‡]

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

Supporting Information

ABSTRACT: We present four models of solution free-energy prediction for druglike molecules utilizing cheminformatics descriptors and theoretically calculated thermodynamic values. We make predictions of solution free energy using physics-based theory alone and using machine learning/quantitative structure−property relationship (QSPR) models. We also develop machine learning models where the theoretical energies and cheminformatics descriptors are used as combined input. These models are used to predict solvation free energy. While direct theoretical calculation does not give accurate results in this approach, machine learning is able to give predictions with a root mean squared error (RMSE) of ∼1.1 log *S* units in a 10-fold cross-validation for our Drug-Like-Solubility-100 (DLS-100) dataset of 100 druglike molecules. We find that a model built using energy terms from our theoretical methodology as descriptors is marginally less predictive than one built on Chemistry Development Kit (CDK) descriptors. Combining both sets of descriptors allows a further but very modest improvement in the predictions. However, in some cases, this is a statistically significant enhancement. These results suggest that there is little complementarity between the chemical information provided by these two sets of descriptors, despite their different sources and methods of calculation. Our machine learning models are also able to predict the well-known Solubility Challenge dataset with an RMSE value of 0.9−1.0 log *S* units.



INTRODUCTION

Poor aqueous solubility remains a major cause of attrition in the drug development process. Despite theoretical developments, the solubility of druglike molecules still eludes truly quantitative computation. In recent work,[1] we have shown that accurate first-principles calculation is now becoming possible, provided that both the crystalline and solution phases are described by accurate theoretical models. Before this, energy terms from a computed thermodynamic cycle (see Figure 1) had been used as descriptors in a multilinear regression model for intrinsic solubility, delivering accuracy much better than from direct computation and comparable with the leading informatics approaches.[2]

Since then, sophisticated machine learning techniques have been applied to many problems in the chemical sciences, while, as we have shown,[2,3] the accuracy of direct computation of hydration energies and solubilities has improved significantly. This led us to revisit the idea of hybrid informatics-theoretical models for solubility.

Cheminformatics methods have seen widespread use for property prediction, particularly in the pharmaceutical industry where they have been applied to; aqueous solubility, melting point, boiling point, log *P* (where *P* is the partition coefficient between octanol and water), binding affinities, and toxicology predictions.[4] Such methods are usually much quicker than pure chemical theory calculations, making high throughput virtual screening (HTVS) a possibility. Some methods have become accessible and easy-to-use web-based tools.[5] However, informatics methods suffer from the difficulty of decomposing the results into intuitive, physically meaningful understanding and cannot reflect the physical details of the system. To understand the underlying physics and chemistry, it is necessary to carry out an atomistic physics-based calculation.

Many chemical theory methods have been developed to specifically address one phase. The exact nature of the theory varies between these methods and the phase being studied. Crystal structures are often modeled using one of the lattice energy minimizing simulation methods,[6] plane-wave density functional theory (DFT) methods,[7] or periodic DFT using atom-centered basis sets.[8] The latter two methods come from a quantum-chemical standpoint. The results are often very good but have a high computational cost. The simulation methods often contain empirical parameters, which lowers the cost of these methods significantly, compared to DFT.

Popular solution-phase models include atomistic simulation methods based on molecular mechanics and dynamics,[9] quantum-mechanical implicit solvation methods (such as the

Received: October 7, 2013

   dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

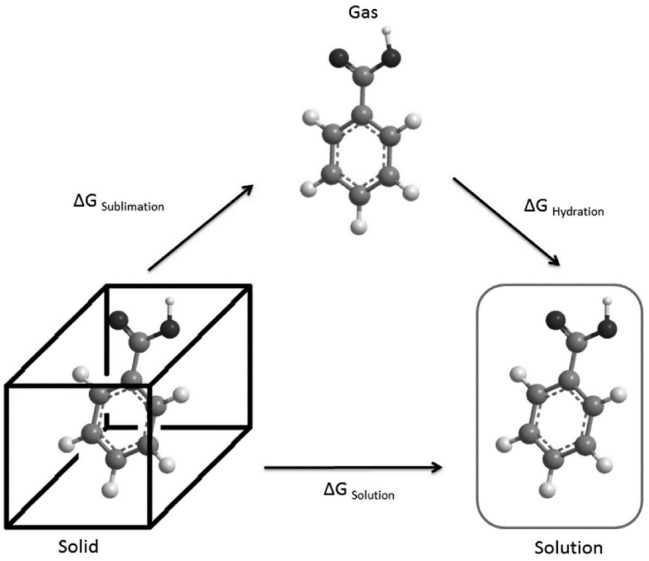


Figure 1. Thermodynamic cycle.

polarizable continuum model (PCM)),[10] and "hybrid" models (such as the classical statistical mechanics-based reference interaction site model (RISM)[11] or hybrid quantum mechanics/molecular mechanics (QM/MM) methods[12]). These methods have the inherent problem for industrial and drug discovery applications of being significantly more computationally intensive than cheminformatics models, which makes high-throughput computation infeasible. The closest thing to an exception among contemporary theoretical models may be 1D RISM, which requires only a few minutes of calculation time per compound and has been previously combined with cheminformatics to build the 1D-RISM/SDC method.[13]

By combining lower levels of theoretical chemistry with cheminformatics, we hope to produce results in good agreement with experiment, but at a lower cost than higher-level theoretical methods, and with higher accuracy than using cheminformatics descriptors alone.

METHODS

Molecules and Solubility. A set of 100 broadly druglike organic molecules was assembled with the prerequisites that each molecule should have an available crystal structure in the Cambridge Structural Database (CSD)[14] and a well-documented aqueous intrinsic solubility in the literature. Where possible, we prefer experimental solubilities obtained with the CheqSol method,[15] which has been shown to give reproducible results with only small random errors. The possibility of significant systematic errors between different experimental methodologies remains an issue and may possibly limit the accuracy with which modeling-based studies can be validated.

A total of 122 potentially useful CheqSol solubilities were obtained from the two Solubility Challenge papers[16] and downloaded from the Web.[17] While noting that several corrections had previously been made, we also corrected or disambiguated the following names: amitriptyline, 5-bromogramine, 5,5-diphenylhydantoin, 4-hydroxybenzoic acid, nortriptyline, and phenanthroline. Of the 122 compounds, 38 had corresponding crystal structures and could be included in our DLS-100 dataset. Where a choice existed, we selected the solubility and crystal structure of the least soluble and, therefore, most stable polymorph. For druglike compounds with known crystal structures, one further CheqSol solubility was available from Palmer et al.[2] and two from Narasimham et al.[18] We sourced solubility data for an additional 59 compounds from other experimental methods.[19] This gave us a total data set of 100 molecules.

The crystal structures were obtained using either the CrystalWeb[20] interface or the ConQuest[21] interface. Crystal structures were selected on the basis of stability, preferring the polymorph with the lowest literature solubility or the lowest lattice energy according to our computations where polymorph-specific experimental information was not available. We also applied the additional pragmatic selection criterion that the asymmetric unit cell should contain only one molecule. Once structures were identified, they were downloaded in either the SHELX format (.res) or CSD legacy format (.dat).

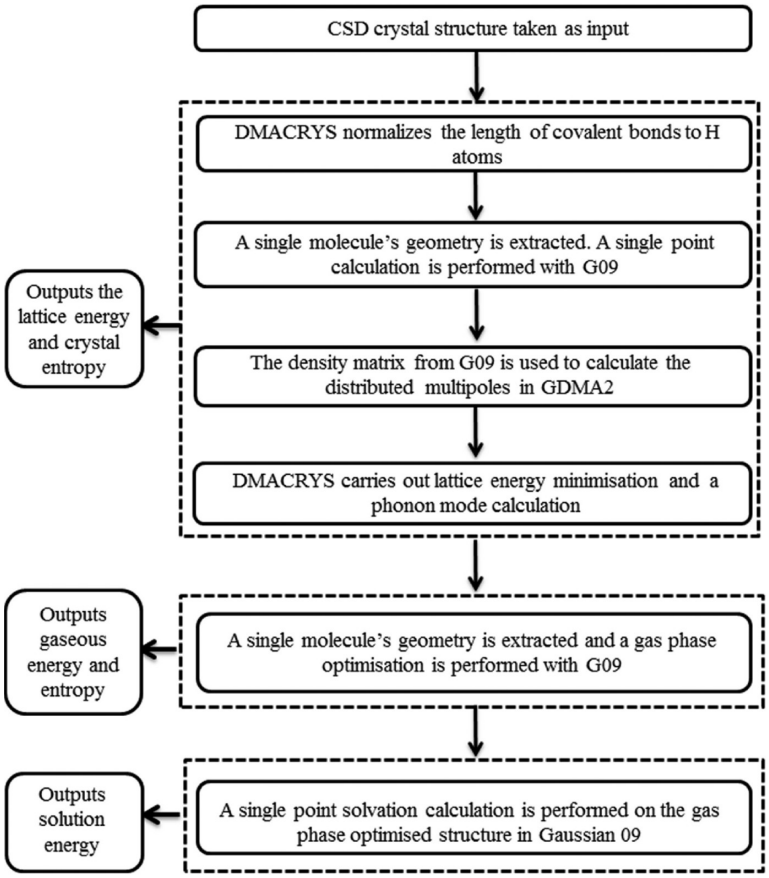
We chose to use Chemistry Development Kit (CDK)[22] molecular descriptors in this study, because these descriptors do not require proprietary software and are applicable to solubility prediction.[23] The CDK is an open source cheminformatics Java library. In order to use the CDK molecular descriptors,[22] we required each of our chemical structures in SMILES format. As noted by O'Boyle,[24] SMILES can be ambiguous. We thus decided to use one principal source for SMILES records, selecting the well-annotated database ChemSpider.[25] Since we are modeling intrinsic solubility, we wish to describe the neutral form of the druglike compound. This remains the case even if a protonated or deprotonated charged form dominates at neutral pH or across the pH range of the CheqSol (or other) experiment. To obtain a SMILES string for each molecule in the DLS-100 dataset, we wrote a Taverna workflow,[26] which uses web services provided by the ChemSpider database.[25,27] The workflow is freely available on the MyExperiment[28] repository at the following reference.[29] In five cases, we found the ChemSpider SMILES to correspond to an undesirable protonation state. Thus, we instead took the SMILES from the solubility challenge Web site[17] for cimetidine, pindolol, and phenobarbital, and from Wikipedia for griseofulvin[30] and glipizide.[31] Using the resulting 100 SMILES, we initially calculated all 268 available nonprotein CDK descriptors for each compound. We found that 145 of these descriptors were either undefined for 2D structures, or had the same value for all 100 compounds; their deletion left 123 remaining descriptors.

Crystal Structure and Gas-Phase Calculations. We took experimentally determined crystal structures of the compounds in our DLS-100 dataset as the initial input to our calculations. DMACRYS,[6] a periodic lattice simulation program, was used to perform the crystal structure minimizations and calculate vibrational contributions arising from the crystal. DMACRYS works in conjunction with the GDMA2[32] and Gausssian 09 (G09) programs.[33] The output of these calculations gives us the enthalpy of sublimation and crystal portion of the entropy of sublimation.

The selected crystal structures were input into DMACRYS, which was used to standardize the covalent bond lengths between hydrogens and heavy atoms, as the experimentally determined bond lengths are not accurate, because of the uncertainty in the hydrogen positions obtained by X-ray diffraction, before any calculations were run. Electrostatic interactions were calculated by multipole expansions[34] (obtained using GDMA2) of molecular charge distributions calculated at the MP2/6-31G** level using G09. Multipolar

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Scheme 1. DMACRYS-G09 Workflow[a]



[a]This scheme typically takes a few hours of calculation time per molecule on two 2.8-GHz 6-core Intel Xeon X5660 processors.

expansions up to hexadecapole were calculated. Intermolecular repulsion and dispersion were calculated by a Buckingham potential.[6,35]

DMACRYS carries out a rigid-body minimization of the crystal structure, hence arriving at minimized lattice energies. This lattice energy can be converted to an enthalpy of sublimation by the following formula:

Enthalpy of sublimation:

$$\Delta H_{sub} = -U_{latt} - 2\mathbf{R}T \quad (1)$$

where U_{latt} is the lattice energy (energy of the crystal assuming the crystal is static and at 0 K relative to infinitely separated molecules) and the $-2\mathbf{R}T$ term arises from lattice vibrational energy.[2,36]

The entropy of sublimation was calculated by:

Entropy of sublimation:

$$\Delta S_{sub} = (S_{rot} + S_{trans}) - S_{crys} \quad (2)$$

where S_{rot} is the rotational entropy in the gas phase and S_{trans} is the entropy of translation in the gas phase. S_{crys} is the entropy of phonon vibrations within the crystal. The use of eq 3 makes these assumptions: (i) the rotational and translational entropy of the crystal is minimal, (ii) there is no change in electronic entropy between phases, and (iii) the intramolecular entropy is constant between the two phases. The crystal entropy is calculated by locating the frequencies of the phonon normal modes (lattice vibrations) at the gamma point. This is achieved using lattice dynamics, the results of which are used to calculate the Helmholtz free energy (see eqs S2 and S3 in the Supporting Information).

Gibbs free energy:

$$\Delta G_{sub} = \Delta H_{sub} - T\Delta S_{sub} \quad (3)$$

The coordinates of a single molecule were extracted from the minimized lattice and used as input for the gaseous optimization with G09. Optimizations were carried out at the M06-2X and HF levels of theory with a 6-31G* basis set. The gas-phase entropy values were calculated from statistical thermodynamics in G09. Finally, ΔG_{sub} is calculated from the enthalpy and entropy of sublimation.

Solution-Phase Calculations. All solution-phase calculations were carried out with G09 using the Self-Consistent Reaction Field (SCRF) protocol. We selected the SMD (Solvation Model based on Density)[37] implicit solvent model based on previous work.[1] Although RISM yielded more-accurate absolute hydration energies than SMD in our recent work,[1] SMD generated a higher correlation coefficient against experimental results for hydration free energy prediction (R = 0.97 vs R = 0.93). Given the parametrized nature of our present model, correlation is more important than absolute agreement, and, hence, SMD is a suitable solvation model. Solution-phase calculations were carried out with the same methodologies as used in the gas-phase calculations, M06-2X/6-31G* and HF/6-31G*. Geometry optimization was again carried out, this time taking the gas-phase optimized structure as the starting point.

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

The SMD model is a parametrized implicit solvation model. SMD solves for the free energy of solution (ΔG_{hyd}) as a sum of the electrostatic contributions and nonelectrostatic contributions. The electrostatic contributions are calculated by the solution of the nonhomogeneous Poisson equation;[23,37] this equation is a second-order differential equation linking the electrostatic potential, dielectric constant, and charge distribution. The nonelectrostatic contributions of cavitation, dispersion, and solvent structure are calculated as a sum of atomic and molecular contributions using parameters inherent to the SMD method. SMD has been shown to provide significant improvements over some other implicit solvent models for datasets containing molecules similar to those used in this study.[1] The hydration free energy is given by eq 4,

Gibbs free energy of hydration:

$$\Delta G_{hyd} = E_{solution} - E_{gaseous} \quad (4)$$

where $E_{solution}$ is the total energy of the system in the SMD solvation model and $E_{gaseous}$ is the total energy of the system in a vacuum. Scheme 1 represents the workflow for making such predictions.

Standard States. Sublimation energies were calculated in the 1 atm standard state, which is the conventional standard for experimental sublimation energies to be quoted. However, solvation free energies are usually quoted in the Ben-Naim standard state of 1 mol/L. In this work, ΔG° corresponds to the 1 atm standard state, while ΔG^{*} corresponds to the Ben-Naim 1 mol/L standard state (see Figure 2).[38] The difference

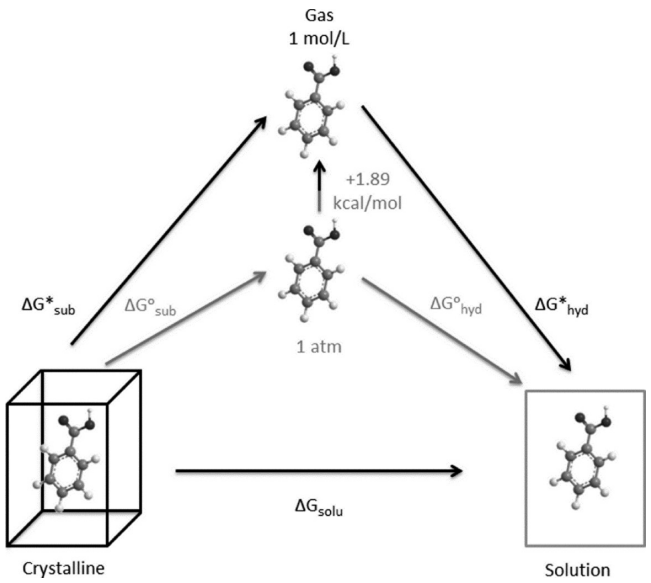


Figure 2. Thermodynamic cycle showing standard state corrections. The presence of an asterisk (*) denotes that the values refer to the standard state of 1 mol/L.

between these two standard states is a constant energy value of 1.89 kcal/mol (7.91 kJ/mol). In this work, we calculate the sublimation free energy in the 1 atm standard state and then apply the correction to 1 mol/L in order to be consistent with the hydration free energy calculations; hence, ΔG_{solu} is in the 1 mol/L standard state for all predictions in this work.

Theoretical Log S Prediction. Our final solution free-energy prediction is then given as the sum of the predicted sublimation and hydration free energies:

Gibbs free energy of solution:

$$\Delta G^{*}_{solu} = \Delta G^{*}_{sub} + \Delta G^{*}_{hyd} \quad (5)$$

Therefore, we have two predictions for each molecule: The first method couples DMACRYS with G09 and the SMD solvation model at the HF/6-31G* level of theory. This model will be referred to as SMD(HF). The second method is DMACRYS coupled with G09 and the SMD solvation model at the M06-2X/6-31G* level of theory. This will be referred to as SMD(M06-2X).

For convenience of comparison with experimental values of solubility, we convert the free energy of solution to log S values, and all experimental solubility values to log S values:

$$\log S = \frac{\Delta G_{solu}}{-2.303\mathbf{R}T} \quad (6)$$

Here, $\mathbf{R}$ is the universal gas constant and T is the absolute temperature (in Kelvin).

The conversion of experimental solubility to log S can be found in the Supporting Information (eq S7). Values for the full DLS-100 dataset, including SMILES and InChI, can be found in the Supporting Information (see zip file and dataset).

Informatics Models. To model the data, we use linear and machine learning regression models: partial least-squares regression, random forest and support vector regression. For reporting the predictive accuracy of these models, we averaged the RMSE of log S over a 10-fold cross-validation of the DLS-100 dataset. The cross-validation fulfils two purposes in this study: parameter optimization and evaluation of the accuracy of the models on unseen data. To ensure that each test fold of data is truly unseen, the parameter optimization is carried out in a separate layer of cross-validation within the training folds, as we will discuss below. In order to avoid overfitting, the data are preprocessed before building the predictive models.

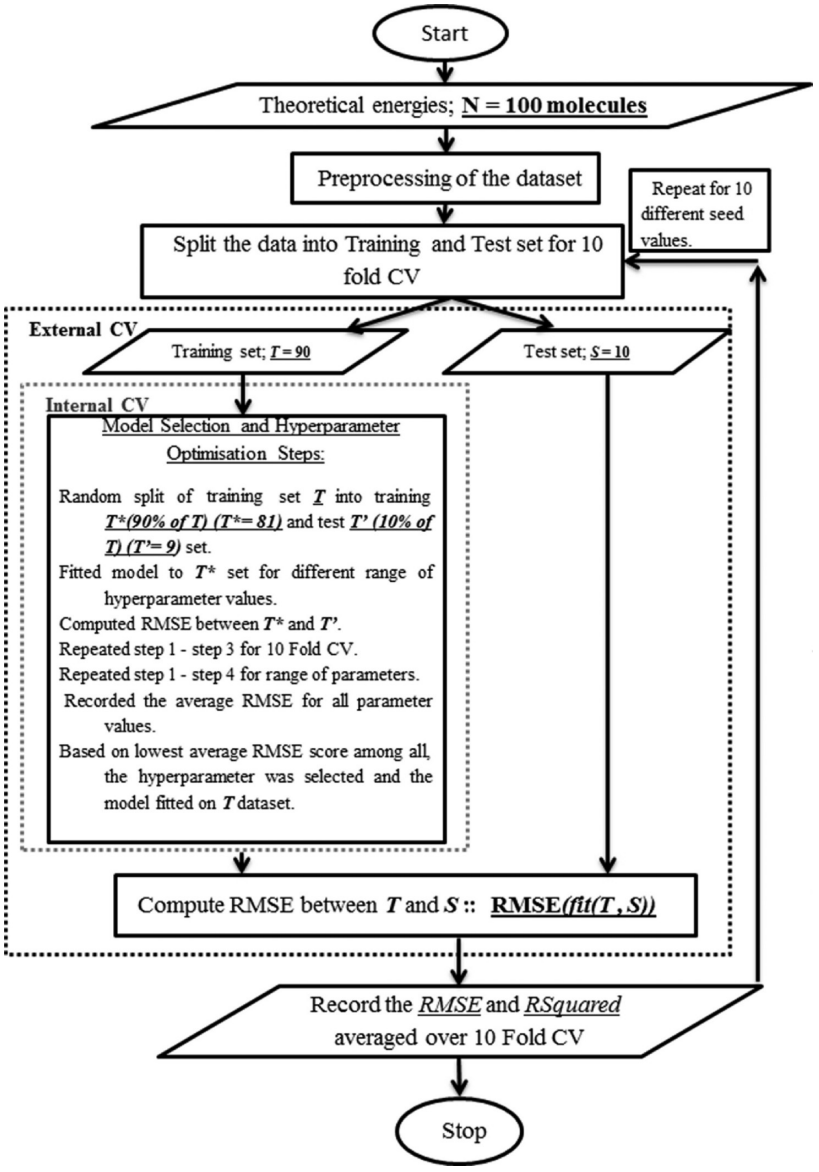
Data Preprocessing. The use of multivariate data presents a danger of overfitting machine learning regression models; moreover, redundancy of attributes and correlation within the data add to the risk of reaching misleading conclusions.[39] To avoid such issues, we have used two normalization methods. One is the commonly used standardization method of variable scaling, equalizing the distributions of the variables by normalizing the mean and standard deviation of each column (variable).[40] The advantage of using this method is that it equalizes the prior importance of all the attributes. The second normalization method is principal component analysis (PCA), transforming the data into a smaller subspace where the new variables are uncorrelated with each other.[39] The PCA data transformation method deals with the redundancy of the data, and places emphasis on the variance of the data. The ability of each principal component to explain the data is measured according to the variance accounted for. Third, we have also fitted each model on the nonpreprocessed raw dataset, for comparison with the results of the two different scaling methods.

Machine Learning Regression Models. In this section, a summary of the regression models are presented; detailed explanations can be found in the Supporting Information.

Partial Least Squares Regression. The Partial Least Squares Regression (PLSR) model design is appropriate in a situation where there is no limit to the X variables or predictors, or where the sample size is small. Moreover, the PLSR model is also beneficial for analyzing strongly colinear and noisy data.

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Scheme 2. Machine Learning Regression Workflow



The goal of a PLSR model is to predict the output variable Y from the input variables X and to describe the structure of X. For this, PLSR finds a set of components from X that are relevant to Y; these components are known as latent variables. The intention of PLSR is to capture the information in the X-variables that is most useful to predict Y.[41] A graphical representation is supplied in the Supporting Information Figure S1(A).

Random Forest Regression. Random Forest (RF), a method for classification and regression analysis, has very attractive properties that have previously been found to improve the prediction of quantitative structure−activity relationship (QSAR) data.[42] An ensemble of many decision trees constitutes a random forest, and each is tree constructed using the Classification and Regression Trees (CART) algorithm.[43] The RF method is efficient in handling high-dimensional data sets and is tolerant of redundant descriptors.

Support Vector Regression. The main idea in Support Vector Regression (SVR) is to minimize the risk factor based on the structural risk minimization[44] from structure theory, to obtain a good generalization of the limited patterns available in the given data. First, the given data D are mapped onto a higher dimensional feature space, using the kernel function $k(x_i,x_j)$ and then a predictive function is computed on a subset of support vectors. Here, we have used the radial basis kernel function (eq 7) to map the data onto a higher dimensional space. A graphical representation is supplied in the Supporting Information (Figure S1(B)).

SVR mapping on radial basis kernel function:

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\delta^2}\right) \tag{7}$$

Statistical Measures. To evaluate the performance of various machine learning models, we report two statistics: the root mean squared estimate (RMSE) and squared Pearson correlation coefficient R^2 (not to be confused with the coefficient of determination).[45] Formulas for these are given in the Supporting Information (eq S5). We have also assessed statistical significance using Menke and Martinez's method,[46]

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

which we have used previously for similar analysis[47] (see Supporting Information (eq S6, Tables S3−S9 for R2, and Boxes S1−S3) for statistical significance). We also analyzed the variable importance for the RF method (see Table S17 in the Supporting Information). Variable importance was calculated in the CART program as implemented in R.[42b,48,49]

10-Fold Cross-Validation. In order to compute and compare the performance of the various regression models, we consider RMSE scores averaged over a 10-fold cross-validation.[50] In the 10-fold cross-validation, the dataset is randomly split into 10 partitions, where the training set consists of 90% of the data and the test set consists of 10% of the data. A predictive regression model is fitted on the training set. The predictivity on the test fold is considered as an external measure to compute the accuracy of the fitted model. The entire process is repeated 10 times in order to cover the entire dataset, with each fold forming the test set on one occasion, and we record the average RMSE. The complete design of the workflow is represented in a flowchart (Scheme 2); similar workflows have been used for classification in other studies.[47,51] The complete workflow of this analysis was written in R[52] using the CARET package;[53] all scripts are available in the Supporting Information.

In out-of-bag validation, one evaluates the performance of the model by separating training and test data through bootstrap sampling; this is convenient only for the RF method. It is not appropriate to compare RF out-of-bag predictions with other models such as PLS and SVR, which are not based on bootstrap sampling. So, we used 10-fold cross-validation to evaluate the performance of our various models.

10-Fold Cross-Validation for Parameter Tuning. For each model, we use 90% of the total data designated as the training set in order to find the optimum values for these parameters. We selected a range incorporating 20 different possible values for each model parameter, in order to select its best value. For each parameter, a further level of 10-fold cross-validation is carried out in order to retrieve the RMSE of the models using each possible parameter value. Here, the training portion of 90% of the original data is further split into 10 new folds of 9%, with nine (81% of the original data) being used to build each model and one (9%) as an internal validation; this process of model building and internal validation is repeated to predict each of the 10 possible internal validation folds. This internal cross-validation step is repeated 20 times, once for each possible value of the parameter being assessed. Then, based on the value giving the lowest average RMSE score in the internal validation folds, the optimum parameter value is selected. Finally, the model is fitted on the complete training set of 90% of the original data using the selected parameter values.

Assessing the Final Models by 10-Fold Cross-Validation. The given 90%:10% split of the data into training and test sets was used to fit the final model for each fold of the main 10-fold cross-validation, once the optimum parameter values have been selected. The average RMSE and R^2 values over the 10 folds were considered in order to compare the usefulness of different descriptor sets and to evaluate the performance of the fitted models.

Dataset. The full DLS-100 dataset, with the experimental log *S* values, can be found as Supporting Information or downloaded from the Mitchell group web server (http://chemistry.st-andrews.ac.uk/staff/jbom/group/Informatics_Solubility.html; see the Supporting Information (csv_smiles_-SI.csv and Table S1)), which is consistent with the excellent suggestions from Walters.[54] The dataset includes CSD refcodes, Chemspider numbers, SMILES, experimental log *S* values and InChI for all molecules. The log *S* values in this work come from refs 2, 16, 18, and 19. Where possible, we have selected data obtained from the CheqSol method; where this was not available, we have selected reliable sources using different determination techniques. A good solubility prediction can be considered as a prediction of approximately the same error as that of the experiment. The experimental values have been shown in a number of previous papers to vary considerably.[55] Here, we consider the experimental accuracy limit to be between 0.6 and 1 log *S* unit (where 1 log *S* unit represents 5.7 kJ/mol at 298 K). Previous work has reported the experimental error in solubility prediction to be as great as 1.5 log *S* units and, on average, the error to be at least 0.6 log *S* units.[56] In 2006, Dearden[55a] noted, as was later reiterated in the Solubility Challenge, that models with RMSE predictions of <0.5 log *S* units are likely to be overfitted.[16b,55a] For a prediction to be useful, it must have an RMSE within the standard deviation of the experimental data; otherwise, a trivial prediction using the mean of the experimental data is a more accurate prediction of the log *S* value.[1] For the DLS-100 dataset, the experimental standard deviation is 1.71 log *S* units.

RESULTS AND DISCUSSION

We have compiled four sets of results for our DLS-100 dataset. First, a purely theoretical prediction, in which no machine learning is used and where predictions are made using only physics-based calculations. Second, theoretical energies are used as the sole descriptors in machine learning models. Third, cheminformatics descriptors, calculated using the CDK, are used as the sole input to machine learning methods. Finally, cheminformatics descriptors and theoretically computed energies are combined as input to machine learning methods. For each of these methods, we present the results and discussion, with comparison between the methods made on the basis of RMSE and R^2 (correlation coefficients for cheminformatics and combined models can be found in the Supporting Information (Tables S3−S9); RMSE values can be found in the Supporting Information (Tables S10−S16)). In addition to these results, we have replicated the solubility challenge using 2D molecular descriptors alone.

Table 1. RMSE and R^2 Values for Theoretical Energy Calculation[a]

	DMACRYS + SMD(M06-2X)	DMACRYS + SMD(HF)
RMSE (log *S* units)	4.045	2.946
R^2	0.252	0.327

[a]Linear regression calculated from eq 1.

Theoretical Predictions. The theoretical methodologies described earlier utilize a thermodynamic cycle to access the free energy of solution. Table 1 shows the R^2 correlation coefficient and the RMSE for the predictions made by these methods. Chart 1 shows the linear fit to the data from the SMD(HF) method, which has the lower RMSE and the higher R^2 correlation coefficient of the two purely theoretical methods.

Chart 1 shows that the data are poorly explained by a linear model. The RMSE for the SMD(HF) method is nearly three times the suggested criterion of 1 log *S* unit of error. The situation is even worse for the SMD(M06-2X) method for

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Chart 1. SMD(HF) Predictions Linear Model

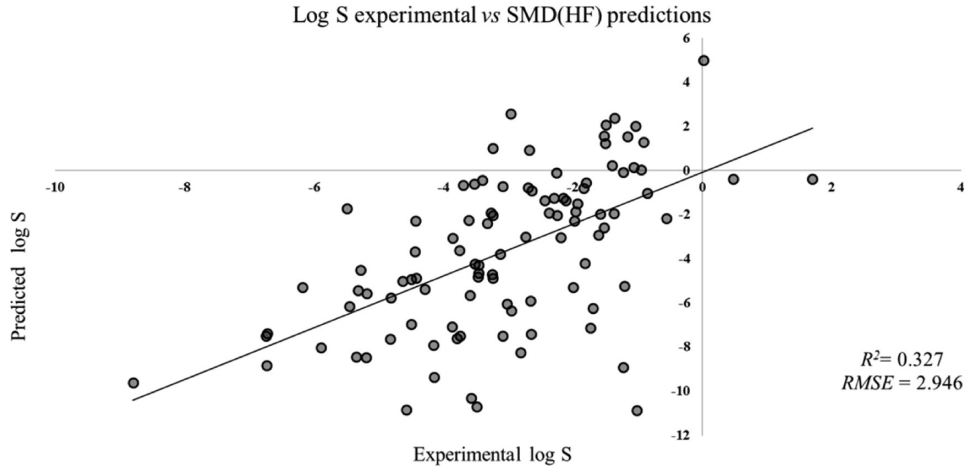


Chart 2. Model HF: Plot Representing the Average RMSE Scores over 10-Fold Cross-Validation for Different Scaling Methods of Various Machine Learning Regression Models

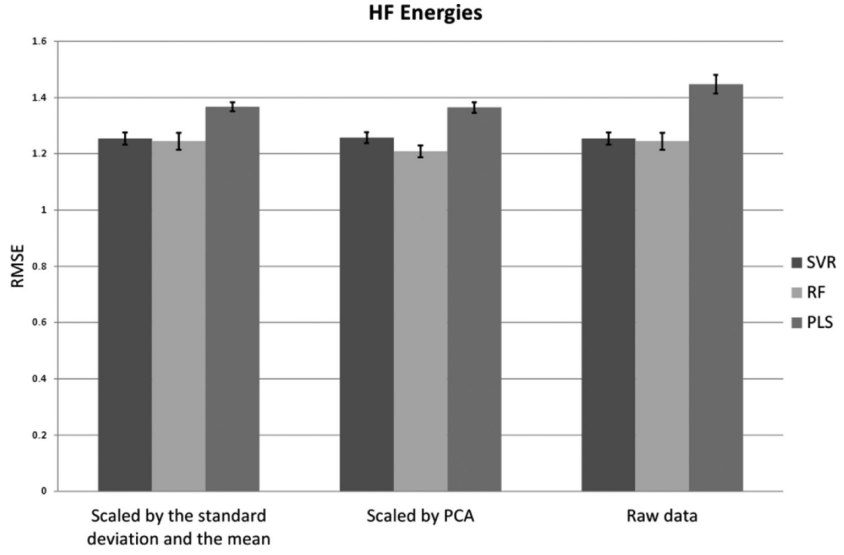
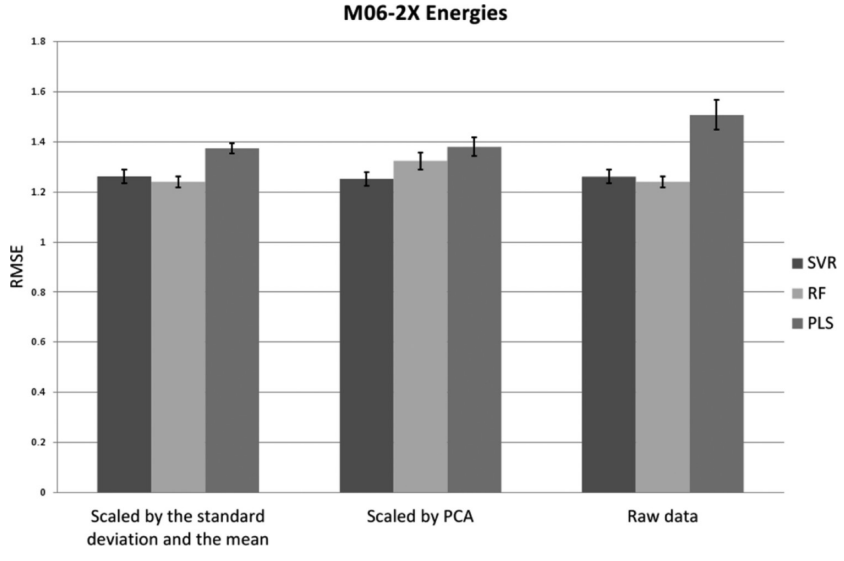


Chart 3. Model M06-2X: Plot Representing the Average RMSE Scores over 10-Fold Cross-Validation for Different Scaling Methods of Various Machine Learning Regression Models



dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Chart 4. Cheminformatics Model: Average RMSE Scores for 10-Fold Cross-Validation

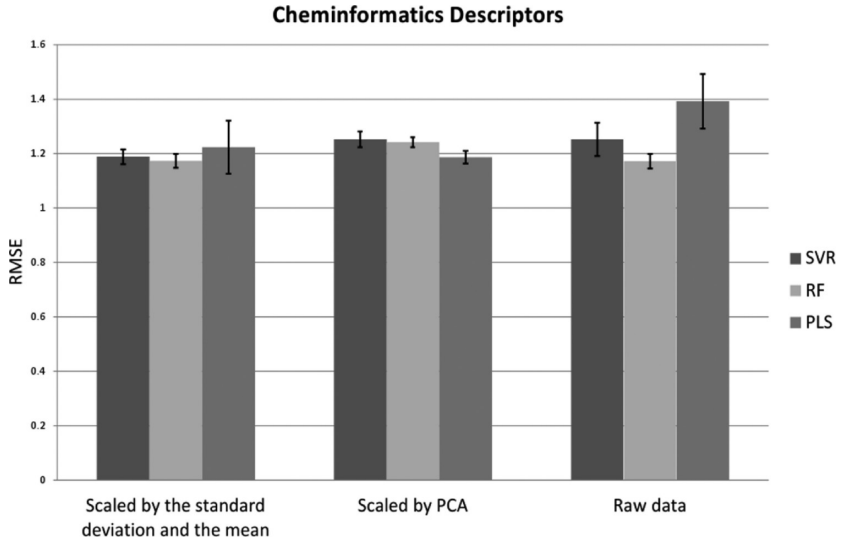


Chart 5. Cheminformatics Model and Model M06-2X: Average RMSE Scores for 10-Fold Cross-Validation

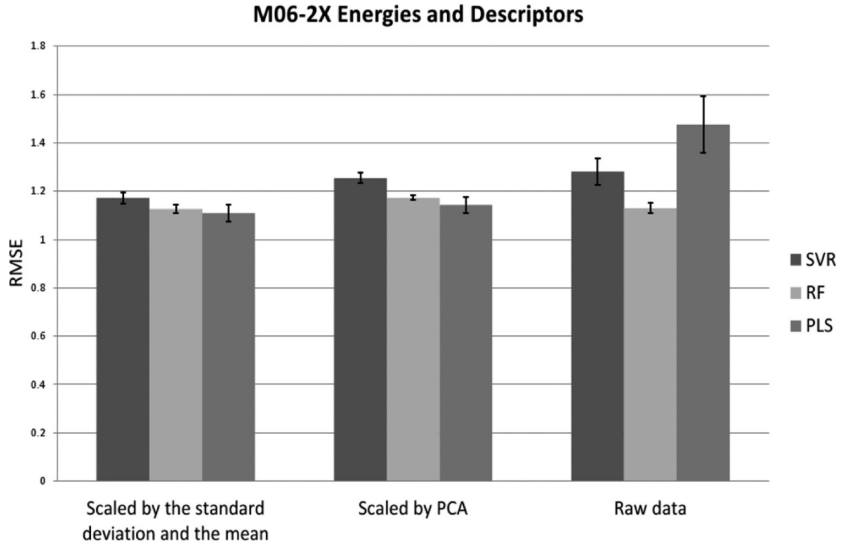
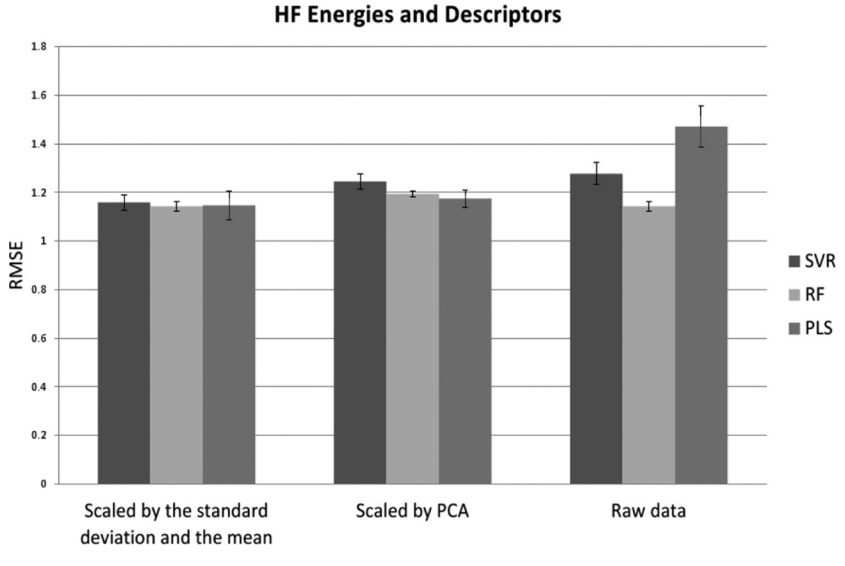


Chart 6. Cheminformatics Model and Model HF: Average RMSE Scores for 10-Fold Cross-Validation



dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 2. Solubility Challenge Dataset: Average over Ten Repetitions of 10-Fold Cross-Validation of RMSE (Standard Deviation) for the Log *S* Calculation

Machine Learning Solubility Challenge	raw data ± stdev	scaled by mean/stdev ± stdev	scaled by PCA ± stdev
PLS	1.08 ± 0.04	1.03 ± 0.02	1.15 ± 0.01
RF	0.93 ± 0.01	0.93 ± 0.01	1.12 ± 0.01
SVR	1.17 ± 0.04	0.93 ± 0.02	0.95 ± 0.02

Table 3. RMSE for the Log S Calculation Using the Solubility Challenge Dataset with Its Original Training:Test Split

Solubility Challenge	raw data	scaled by mean/stdev	scaled by PCA
PLS	0.89	0.91	0.91
RF	0.93	1.03	1.02
SVR	1.08	1.07	1.08

which the RMSE is just over four times this criterion (see Charts S4−S6 in the Supporting Information). Both methods produce results outside the useful prediction criterion of 1.71 log *S* units. From these results, we can draw a couple of conclusions. First, it is clear that the given methodologies do not adequately quantify the physics occurring in the solution process (i.e., solid to solution). Second, we can conclude that, if it is possible to explain the underlying structure of these data using a general model, based on the predicted log *S* values, such a model will be inherently nonlinear.

Compared with our previous work,[1] in which theoretical models provided a good prediction of log *S*, our theoretical methodology here differs only marginally, in the use of MP2 multipoles, and still produces good results (see Supporting Information (Chart S1 and Table S2)) for the same 25 molecules in this work (dataset DLS-25). The predictions for the additional 75 molecules alone show worse predictions than for the full 100-molecule set presented above (see Charts S2 and S3 in the Supporting Information). The additional 75 molecules therefore appear to form a more difficult dataset to predict. It is likely that improved results can be obtained from purely theoretical calculations, if some of the approximations made here are improved; for example, improved modeling of the solvated phase to more accurately describe the solvent and its effects on the solute could increase accuracy. Also, we note that the intramolecular degrees of freedom are neglected in the DMACRYS calculations, and further assumptions are made by using eqs 2 and 3 in the Methods section.

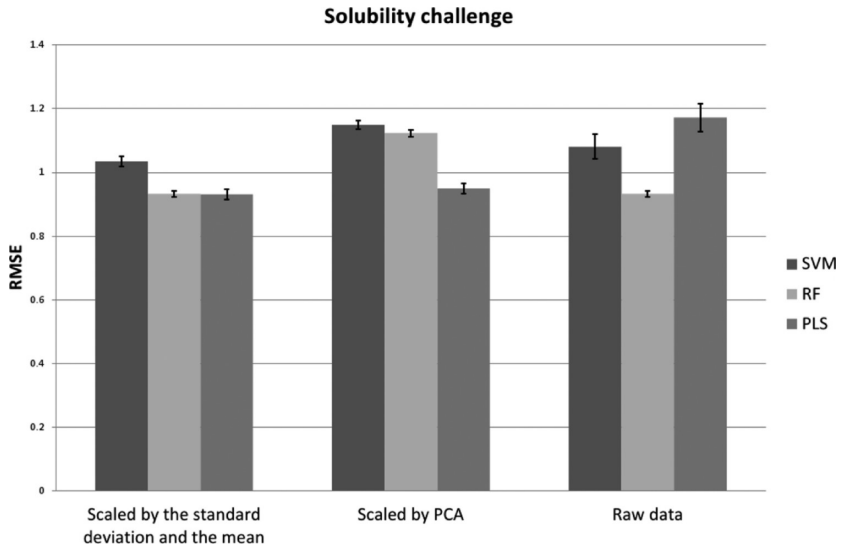
We subsequently applied machine learning methods to the theoretical energies in order to carry out nonlinear regression analysis. The average RMSE scores over 10-fold cross-validation (see the Methods section for details) is represented as two-dimensional (2-D) column charts (see Charts 2 and 6). Different grayscale column bars represent the different machine learning methods used in this study. The standard deviation is shown as an error bar (black line).

Theoretical Energies as Sole Descriptors in Machine Learning. The use of the calculated energies as descriptors in the machine learning models yields considerably improved results, compared to those from the predictions made without machine learning. The results now, while still missing the 1 log *S* unit error criterion, do make useful predictions in which the RMSE is within the standard deviation of the experimental data (1.71 log *S* units). The RF and SVR models produce notably better results than PLS. Charts 2 and 3 show that the method minimizing the RMSE (1.21 log *S* units) is RF with HF when scaled with PCA.

Cheminformatics Descriptors as the Sole Input to Machine Learning. An additional point of interest is that the chemical descriptors alone using RF or SVR can provide a marginally better prediction of log *S* than the machine learning methods with only the energies as descriptors. In particular, we noticed that fitting the RF model on data that are scaled to a given mean and standard deviation produces a statistically significant improvement in its prediction with cheminformatics descriptors alone rather than theoretical energies (see the Supporting Information (Boxes S1−S3)). In all other cases, the changes are not significant. This suggests that slightly more useful information about the molecules' log *S* values is conveyed by the cheminformatics descriptors than by the theoretical energies alone (see Chart 4).

Theoretical Energies and Cheminformatics Descriptors as Input to Machine Learning. When the descriptors and energies are combined as input for the machine learning

Chart 7. Solubility Challenge: Average RMSE Scores for 10-Fold Cross-Validation



dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

methods, we obtain results that are generally only very slightly better than those obtained from cheminformatics descriptors alone. This implies that the theoretical energies contain very little extra useful information not already present in the descriptors. The joint results do present a statistically significant improvement for PLS and RF, once scaled by the mean/standard deviation, compared to those for the theoretical energies alone. In light of this, and given that the descriptors alone produce a marginally improved result compared to chemical theory, it is fair to say the cheminformatics descriptors are seen to contain a modest amount of additional information not incorporated in the theoretical energy terms. This suggests that the 123 descriptors of the cheminformatics descriptors and the 10 theoretical energy descriptors convey similar information, with only a small amount of additional information being conveyed by adding the descriptors to the energies and almost no information gained by adding the energies to the set of descriptors. We can conclude that these two sets of features are not generally complementary.

Interestingly, the best result in terms of RMSE is from the descriptors with the M06-2X energies, which, on their own, produced the worse of the two pure theory results in this work (see Charts 5 and 6). The RF model performs particularly well over all descriptor sets, even without any type of scaling, the best RMSE result being only 0.13 units outside the 1 log S unit target. The best single prediction, in terms of the RMSE, was made by the PLS model, using descriptors and the M06-2X energies scaled by the standard deviation and the mean, with an RMSE of 1.11 $\pm$ 0.04 log S units. All of these methods make predictions inside the standard deviation of the experimental data; therefore, all of the predictions are useful. We also note that the RF model shows small but statistically significant improvements with all scaling methods (using the theoretical energies and cheminformatics descriptors combined) when compared to some models trained on the theoretical energies only (see Supporting Information (Boxes S1−S3)). This is the only model to show such improvements with all scaling methods in the present work.

We analyzed the relative variable importance (see Table S17 in the Supporting Information) and found that X log P (from ref 57) was consistently rated as the most important feature. X log P is a computed estimate of the base-10 logarithm of the octanol:water partition coefficient (the ratio of concentrations of solute solvated in the two different solvents). This has been seen in many previous studies and is not so surprising given that it provides information specifically about the solvated phase.[4,56] X log P uses an atom additive model for the prediction of log P. In the Supporting Information, we include tables (Table S17 in the Supporting Information) displaying the 10 most important descriptors; here, we will briefly comment upon these. We find Kier and Hall's χ path and chain indices[58] to be of importance; these quantify the degree of bonding to heavy atoms within a given path or chain length. In addition, the Moreau−Broto autocorrelation,[59] which describes the charge and mass distribution along a given path length, is found in the top 10. Finally, we also note Randic's weighted path descriptors,[60] which are used to account for molecular branching. Once the theoretical energies are added to the descriptor set, the free energies of hydration and solution are ranked in the top 10, along with the theoretical log S prediction. Explanations of the molecular descriptors used in this work can be found in ref 61.

Chemically, we can see logic in the most important descriptors. One may expect that molecular branching would play an important role, because it gives information on the extent and flexibility of the molecule, hence contributing some entropic information. Coupling this descriptor with the Kier Hall descriptors, information can be acquired on the composition of such chains, in terms of heavy atoms. The autocorrelation descriptor provides charge and mass distribution information. Again, here, information is imparted concerning the distribution of heavy atoms and electronic factors. For example, the degree of charge separation across a molecule and the localization of charges are important factors in determining particularly enthalpic but also entropic contributions. The theoretical energies in the top 10 are all closely related quantities; it is not surprising that the (purely theoretical) prediction of log S is found in the top 10: since this is the quantity we are trying to predict, it is expected to provide sufficient information to the model to be found in the top 10. The free energies of solution and hydration provide direct information from electronic structure theory and statistical thermodynamics on the interactions of a given molecule, in a given conformation, within its environment, and on the energetics of phase transitions.

As a benchmark, we also present our method's predictions of the solubility challenge set based solely on cheminformatics descriptors (see Table 2). As suitable crystal structures are not available for all molecules in the solubility challenge, we could not calculate the theoretical energies.

Tables 2 and 3, and Chart 7, demonstrate that our method can make predictions for the solubility challenge dataset within the coveted 1 log S unit RMSE error and, in fact, makes predictions that are consistent with some commercially available methods and deep-learning methods. A recent publication[56] reported RMSE scores of 0.95 log S units[56] for the commercially available package MLR-SC[62] and 0.90 log S units for a deep-learning method.[56] However, these results are not directly comparable with ours, for two reasons. First, our results have been calculated for a 10-fold cross-validation and for the canonical training:test split (see Tables 2 and 3). Second, the deep-learning result (RMSE = 0.90) given by Lusci et al.[56] is contingent on correcting eight putative errors in the CheqSol solubility data, the most substantial of which is for indomethacin, a compound that has been shown to hydrolyze under alkaline conditions.[63] While we have corrected names and SMILES for the solubility challenge set, we have not adjusted any solubility values therein. It is also reasonable to suggest that, using the solubility challenge set as a benchmark, our 100-molecule set could be considered as a "difficult set", given the improved prediction offered by our method when the solubility challenge set is used instead.

■ CONCLUSIONS

Our current work shows that accurate solution free energies are not calculable via the simple theoretical procedure that we present here. A significant portion of the important physics in the solution process is not captured using the approximate methodologies that we utilize in this work. This reaffirms that, currently, QSPR methodologies are the most-accurate and time-efficient methods for accurate solution free energy predictions. In addition, we show that state-of-the-art machine learning methods, with a modest number of cheminformatics descriptors, are capable of making solution free-energy predictions that are consistent with those of commercially

dx.doi.org/10.1021/ci4005805 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

available programs and newer deep-learning approaches. Here, theoretical energies and cheminformatics descriptors are generally shown to not be complementary for such predictions. Since both sets of descriptors (theoretical energies and cheminformatics descriptors) produce a similar level of accuracy when used alone in the machine learning methods, and little improvement is seen when they are combined, we can conclude that the information conveyed is of a similar nature and that the theoretical energies are, for this reason, a more efficient form of information storage, as 10 descriptors contain equivalent information to 123 molecular descriptors. However, in terms of time, the molecular descriptors are much less expensive to calculate and their use is therefore more time-efficient. Additionally, we note that the RF method has produced promising predictions in this work, with relatively low RMSE. This method has consistently produced good results and would be our recommended method to make solubility predictions.

■ ASSOCIATED CONTENT

Supporting Information

Informatics_Solubilty_datasets_and_scripts.zip, including R codes, Bash scripts, Python scripts, macro (.xlsb), DLS-100.csv and Solubility_Challenge_dataset.xlsx. DLS-100.csv contains experimental log *S* values, references, SMILES, sources of smiles, CSD refcodes, molecules names, InChI and Chemspider numbers. SI_document.pdf: Structure data, 2D images of the molecular structures, experimental log *S* values, CSD refcodes, R^2, statistical significance, variable importance. This material is available free of charge via the Internet at http://pubs.acs.org. All scripts and datasets used in this work are available for download from the Mitchell Group web server (http://chemistry.st-andrews.ac.uk/staff/jbom/group/Informatics_Solubility.html, as well as in the Supporting Information.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jbom@st-andrews.ac.uk.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. Datasets were curated by J.L.M. and J.B.O.M. Machine learning R scripts were produced by N.N. and L.D.F. The Taverna workflow was produced by L.D.F. Bash scripts, Excel macros, and the R script to run over multiple directories were produced by J.L.McD. DMACRYS and Gaussian calculations were run by J.L.McD. Advice on computational chemistry and machine learning methods was provided by T.v.M. and J.B.O.M. R calculations were run by J.L.McD. and N.N.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Scottish Universities Life Science Alliance (SULSA), this work was partly supported by Biotechnology and Biological Sciences Research Council (BBSRC) (No. BB/I00596X/1), Scottish Funding Council (SFC). We thank EaStCHEM for access to the EaStCHEM Research Computing Facility, and Dr. Herbert Früchtl for its maintenance. We are grateful to Dr. Graeme Day (University of Southampton) for providing additional scripts for DMACRYS. We thank Dr. David Palmer (University of Strathclyde) for a script to help automate running DMACRYS. We also thank our colleagues at the University of St. Andrews for useful discussions, particularly Dr. Lazaros Mavridis and Rachael Skyner. We thank the BBSRC for Grant No. BB/I00596X/1 to J.B.O.M., which supports L.D.F.'s research. We thank the Scottish Universities Life Sciences Alliance (SULSA) for supporting J.B.O.M., J.L.McD., and N.N., and we also thank the Scottish Overseas Research Student Awards Scheme of the Scottish Funding Council (SFC) for financial support of N.N.'s studentship.

■ ABBREVIATIONS

CDK = Chemistry Development Kit; HTP = high-throughput screening; DFT = density functional theory; HF = Hartree–Fock; PCM = polarizable continuum model; RISM = reference interaction site model; CSD = Cambridge Structural Database; G09 = Gaussian 09; SCRF = self-consistent reaction field; SMD = Solvation Model based on Density; PLSR = partial least squares regression; RF = random forest; SVR = support vector regression; RMSE = root mean squared error; PCA = principal component analysis; CART = classification and regression trees; QSAR = quantitative structure–activity relationship; QSPR = quantitative structure–property relationship; SRM = structural risk minimization; SMILES = Simplified Molecular-Input Line-Entry System; DLS-100 = Drug-Like-Solubility-100

■ REFERENCES

(1) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 3322–3337.

(2) Palmer, D. S.; Llinas, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting intrinsic aqueous solubility by a thermodynamic cycle. *Mol. Pharm.* **2008**, *5* (2), 266–279.

(3) Mitchell, J. B. O. Informatics, machine learning and computational medicinal chemistry. *Future Med. Chem.* **2011**, 3 (4), 451–67.

(4) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model* **2008**, *48* (1), 220–232.

(5) (a) Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **2005**, *10* (22), 1497–1500. (b) 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. Aid. Mol. Des* **2005**, *19*, 453–63.

(6) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **2010**, *12* (30), 8478–8490.

(7) Segall, M. D.; Lindan, P. J. D.; Probert, M. J.; Pickard, C. J.; Hasnip, P. J.; Clark, S. J.; Payne, M. C. First-principles simulation: Ideas, illustrations and the CASTEP code. *J. Phys. Condens. Matter* **2002**, *14* (11), 2717–2744.

(8) Dovesi, R.; Orlando, R.; Civalleri, B.; Roetti, C.; Saunders, V. R.; Zicovich-Wilson, C. M. CRYSTAL: a computational tool for the ab initio study of the electronic properties of crystals. *Z Kristallogr.* **2005**, *220* (5-2005–6-2005), 571–573.

(9) (a) Luder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Kjellander, R. In silico prediction of drug solubility: 2. Free energy of solvation in pure melts. *J. Phys. Chem. B* **2007**, *111* (7), 1883–1892. (b) Luder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Kjellander, R. In silico prediction of drug solubility. 3. Free energy of solvation in pure amorphous matter. *J. Phys. Chem. B* **2007**, *111* (25), 7303–7311.

(c) Luder, K.; Lindfors, L.; Westergren, J.; Nordholm, S.; Persson, R.; Pedersen, M. In Silico Prediction of Drug Solubility: 4. Will Simple Potentials Suffice? *J. Comput. Chem.* **2009**, *30* (12), 1859−1871. (d) Westergren, J.; Lindfors, L.; Hoglund, T.; Luder, K.; Nordholm, S.; Kjellander, R. In silico prediction of drug solubility: 1. Free energy of hydration. *J. Phys. Chem. B* **2007**, *111* (7), 1872−1882.

(10) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999−3094.

(11) (a) Ten-no, S. Free energy of solvation for the reference interaction site model: Critical comparison of expressions. *J. Phys. Chem.* **2001**, *115* (8), 3724−3731. (b) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a universal method for calculating hydration free energies: A 3D reference interaction site model with partial molar volume correction. *J. Phys.: Condens. Matter* **2010**, *22* (49), 492101.

(12) Stanton, R. V.; Hartsough, D. S.; Merz, K. M. Calculation of solvation free energies using a density functional/molecular dynamics coupled potential. *J. Phys. Chem.* **1993**, *97* (46), 11868−11870.

(13) Ratkova, E. L.; Fedorov, M. V. Combination of RISM and Cheminformatics for Efficient Predictions of Hydration Free Energy of Polyfragment Molecules: Application to a Set of Organic Pollutants. *J. Chem. Theory Comput.* **2011**, *7* (5), 1450−1457.

(14) Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* **2002**, *B58*, 380−388.

(15) Box, K.; Comer, J. E.; Gravestock, T.; Stuart, M. New Ideas about the Solubility of Drugs. *Chem. Biodiversity* **2009**, *6* (11), 1767−1788.

(16) (a) Hopfinger, A. J.; Esposito, E. X.; Llinàs, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *49* (1), 1−5. (b) Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48* (7), 1289−1303.

(17) The Goodman group. http://www-jmg.ch.cam.ac.uk/data/solubility/ (accessed Feb. 8, 2013).

(18) Narasimham, L.; Barhate, V. D. Kinetic and intrinsic solubility determination of some β-blockers and antidiabetics by potentiometry. *J. Pharm. Res.* **2011**, *4* (2), 532−536.

(19) (a) Bergström, C. A. S.; Luthman, K.; Artursson, P. Accuracy of calculated pH-dependent aqueous drug solubility. *Eur. J. Pharm. Sci.* **2004**, *22* (5), 387−398. (b) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1477−1488. (c) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (2), 354−357. (d) Rytting, E.; Lentz, K.; Chen, X.-Q.; Qian, F.; Venkatesh, S. Aqueous and cosolvent solubility data for drug-like organic compounds. *AAPS J.* **2005**, *7* (1), E78−E105. (e) Shareef, A.; Angove, M. J.; Wells, J. D.; Johnson, B. B. Aqueous Solubilities of Estrone, 17β-Estradiol, 17α-Ethynylestradiol, and Bisphenol A. *J. Chem. Eng. Data* **2006**, *51* (3), 879−881.

(20) CrystalWeb unfortunately withdrawn in 2013. http://cds.dl.ac.uk/cds/datasets/crys/cweb/cweb.html.

(21) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. New software for searching the Cambridge Structural Database and visualizing crystal structures. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58* (3 Part 1), 389−397.

(22) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493−500.

(23) Gupta, R. R.; Gifford, E. M.; Liston, T.; Waller, C. L.; Hohman, M.; Bunin, B. A.; Ekins, S. Using Open Source Computational Tools for Predicting Human Metabolic Stability and Additional Absorption, Distribution, Metabolism, Excretion, and Toxicity Properties. *Drug Metab. Dispos.* **2010**, *38* (11), 2083−2090.

(24) O'Boyle, N. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **2012**, *4* (1), 22.

(25) RSC ChemSpider. (accessed Feb. 8, 2013).

(26) Wolstencroft, K.; Haines, R.; Fellows, D.; Williams, A.; Withers, D.; Owen, S.; Soiland-Reyes, S.; Dunlop, I.; Nenadic, A.; Fisher, P.; Bhagat, J.; Belhajjame, K.; Bacall, F.; Hardisty, A.; Nieva de la Hidalga, A.; Balcazar Vargas, M. P.; Sufi, S.; Goble, C. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **2013**, *41* (W1), W557−W561.

(27) Little, J. L.; Williams, A. J.; Pshenichnov, A.; Tkachenko, V. Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider. *J. Am. Soc. Mass. Spectrom.* **2012**, *23* (1), 179−185.

(28) Goble, C. A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P.; De Roure, D. myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Res.* **2010**, *38* (suppl 2), W677−W682.

(29) De Ferrari, L. Workflow Entry: From molecule name to SMILE and InchI using ChemSpider. http://www.myexperiment.org/workflows/3603.html. (accessed 10th February 2014).

(30) Griseofulvin. http://en.wikipedia.org/wiki/Griseofulvin (accessed 11th December 2012. SMILES source).

(31) Glipizide. http://en.wikipedia.org/wiki/Glipizide (accessed 11th December 2012. SMILES source).

(32) Stone, A. Distributed Multipole Analysis of Gaussian wavefunctions GDMA version 2.2.02. http://www-stone.ch.cam.ac.uk/documentation/gdma/manual.pdf (accessed Feb. 10, 2014).

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; shida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Gaussian, Inc: Wallingford, CT, 2009.

(34) Stone, A. J. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chem. Phys. Lett.* **1981**, *83* (2), 233−239.

(35) Buckingham, R. The classical equation of state of gaseous helium, neon and argon. *Proc. R. Soc. Lon. Ser-A* **1938**, *168* (933), 264−283.

(36) Gavezzotti, A.; Filippini, G. *Theoretical Aspects and Computer Modeling*.; Gavezzotti, A., Ed. Wiley and Sons: Chichester, 1997; pp 61−97.

(37) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378−6396.

(38) (a) Ben-Naim, A. Standard thermodynamics of transfer. Uses and misuses. *J. Phys. Chem.* **1978**, *82* (7), 792−803. (b) Ben-Naim, A.; Marcus, Y. Solvation thermodynamics of nonionic solutes. *J. Phys. Chem.* **1984**, *81* (4), 2016−2027.

(39) Howley, T.; Madden, M. G.; O'Connell, M.-L.; Ryder, A. G. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. *Knowl.-Based Syst.* **2006**, *19* (5), 363−370.

(40) Wold, H. *Partial Least Squares (PLS) Regression* **2003**, 1−7.

(41) (a) Abdi, H. *Partial Least Squares (PLS) Regression* **2003**, 1−7. (b) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemometr. Intell. Lab.* **2001**, *58* (2), 109−130. (c) Mevik, B.; Wehrens, R. The pls Package: Principal Component and Partial Least Squares Regression in R. *J Stat Softw.* **2007**, *18* (2), 1−24.
(42) (a) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model* **2006**, *47* (1), 150−158. (b) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947−1958.
(43) Breiman, L. Random Forests. *Mach. Learning* **2001**, *45* (1), 5−32.
(44) (a) Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z.; Chen, X.; Li, H.-D. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J. Chemometr.* **2010**, *24* (9), 584−595. (b) Vapnik, V. N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10* (5), 988−999.
(45) Hu, S. In *R2 Vs. r2*, SCEA/ISPA Conference, 2008; pp 1 - 15.
(46) Menke, J.; Martinez, T. R. In Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons, *IEEE IJCNN*, July 25−29, 2004; **2004**; Vol. 2, pp 1331-1335.
(47) Nath, N.; Mitchell, J. B. O. Is EC class predictable from reaction mechanism? *BMC Bioinformatics* **2012**, *13* (1), 60.
(48) Kuhn, M. Variable Importance Using The caret Package 2012; Available via the Internet at http://cran.open-source-solution.org/web/packages/caret/vignettes/caretVarImp.pdf, accessed Feb. 10, 2014.
(49) Kuhn, M. *Variable Importance Using The caret Package* **2010**, 1−7.
(50) Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7* (1), 91.
(51) Simon, R. M.; Subramanian, J.; Li, M.-C.; Menezes, S. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief. Bioinform.* **2011**, *12* (3), 203−214.
(52) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing, Vienna, Austria, 2011.
(53) (a) Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Software* **2008**, *28*, 1−26. (b) Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z. R Core Team. caret: Classification and Regression Training. *R Package "caret"*. http://CRAN.R-project.org/package=caret.
(54) Walters, W. P. Modeling, Informatics, and the Quest for Reproducibility. *J. Chem. Inf. Model* **2013**, *53* (7), 1529−1530.
(55) (a) Dearden, J. C. In silico prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1* (1), 31−52. (b) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 355−366.
(56) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, *53*, 1563−1575.
(57) Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Perspect. Drug Discovery Des.* **2000**, *19* (1), 47−66.
(58) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(59) Moreau, G.; Broto, P. The autocorrelation of a topological structure: A new molecular descriptor. *New J. Chem.* **1980**, 359−360.
(60) Randic, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24* (3), 164−175.
(61) CDK Descriptor Summary (2011−05−28). http://pele.farmbio.uu.se/nightly-1.2.x/dnames.html, accessed Feb. 10, 2014.
(62) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49* (11), 2572−2587.
(63) Tsvetkova, B.; Pencheva, I.; Zlatkov, A.; Peikov, P. High Performance Liquid Chromatographic Assay of Indomethacin and its Related Substances in Tablet Dosage Forms. *Int. J. Pharm. Pharm. Sci.* **2012**, *4* (Supplement 3), 549−552.