Articles

Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure−Property Relationships

Ailan Cheng* and Kenneth M. Merz, Jr.‡

ADMET R&D, Accelrys, subsidiary of Pharmacopeia, Inc., CN5375, Princeton, New Jersey 08543-5375

Received June 21, 2002

"Fail early and fail fast" is the current paradigm that the pharmaceutical industry has adopted widely. Removing non-drug-like compounds from the drug discovery lifecycle in the early stages can lead to tremendous savings of resources. Thus, fast screening methods are needed to profile the large collection of synthesized and virtual libraries involved in the early stage. Solubility is one of the filters that are applied extensively to ensure that the compounds are reasonably soluble so that synthesis of the compounds and assay studies of pharmacokinetics and toxicity are feasible. To address this need, we have developed a fast quantitative structure−property relationship (QSPR) model for the prediction of aqueous solubility (at 298 K, unbuffered solution) from the molecular structures. Multiple linear regressions and genetic algorithms were used to develop the models. The model was based on a set of diverse compounds including small organic molecules and drug and drug-like species. The predicted solubility for the training and test sets agrees well with the experimental values. The coefficient of determination is R^2 = 0.84 for the training set of 775 compounds and the RMS error = 0.87. This model was validated on four sets of compounds. The RMS error for the 1665 compounds from the four validation data sets (including compounds from the Physician's Desk References and Comprehensive Medicinal Chemistry databases) is 1 log unit and the unsigned error is 0.77. This model does not require 3-D structure generation which is rather time-consuming. Using 2-D structure as input, this model is able to compute solubility for 90 000−700 000 compounds/h on a SGI Origin 2000 workstation. This kind of fast calculation allows the model to be used in data mining and screening of large synthesized or virtual libraries.

Introduction

Solubility of a compound is defined as the amount of solute dissolved in a saturated solution under equilibrium conditions. Dissolution is the process of approaching the equilibrium solubility.[1] Solubility is a property of interest to many areas of research, such as pharmaceutical, material, physical, and environmental sciences. It is particularly important to the pharmaceutical industry because solubility is relevant to pharmacokinetic properties (absorption, distribution, metabolism, and excretion) and toxicity. For example, a drug must be soluble so that it can be absorbed across the biological membrane to reach the target organ or issue. Solubility of a compound must be accurately determined to assess the concentrations that the drug will achieve in the target area, to establish the therapeutic level, and to prevent toxicity. Lower solubility can hinder the biological activity (for example, absorption and distribution) of a compound, and often a special formulation or modification is required to enhance the solubility. Drug modification can be complex, time-consuming, and sometimes lead to unexpected results.

A drug discovery lifecycle, a rather lengthy and costly process, typically spans 10−15 years. The worldwide R&D expenditures by the pharmaceutical industry have been increasing steadily during the past 30 years. The annual cost for U. S. pharmaceutical companies reached 26 billion in 2000.[2] However, 75% of the overall R&D cost is attributed to failures.[3] Among these, 29% is due to lack of efficacy, 40% is due to pharmacokinetics, and 11% is due to animal toxicity.[4] It is clear that an ideal drug is a balance of potency, selectivity, pharmacokinetics, and toxicity. Appropriate physicochemical properties (e.g., logP and solubility) together with pharmacokinetic properties and toxicity are the major determinants for progressing from a good lead to a good drug. Tremendous cost can be saved by weeding out the non-drug-like compounds in the early drug discovery lifecycle. However, experimental assays and animal or clinical tests are expensive and not practical to apply to the large collection of compounds in the early stage. Accurate predictive methods can be used to identify and prioritize candidates for development, to assist the rational design of compounds with desirable profiles, and to prioritize and even to reduce the experimental studies and animal tests. Many currently available ADME/Tox prediction tools have been applied in the

* To whom correspondence should be addressed. 548 Westgate Drive, State College, PA 16803. Telephone: (814)-231-8422. Fax: (814)-863−7846. E-mail: cheng189@adelphia.net.

‡ Current address: Department of Chemistry, 152 Davey Laboratory, Pennsylvania State University, University Park, PA 16802.

10.1021/jm020266b CCC: $25.00
Published on Web 07/23/2003

discovery settings by medicinal chemists to enhance the drug-like characteristics of lead compounds.[5−11]

Moreover, with the advent of combinatorial chemistry approaches it is possible to generate very large real or virtual libraries, which has required the development of library design tools that drive libraries toward more drug-like characteristics. Hence, fast property prediction tools are essential to ensure that the design of combinatorial libraries can be approved.

Significant effort has been put into the prediction of solubility for small organic compounds and environmentally important chemicals. However, only very few studies were focused on the prediction of solubility of drug-like molecules, primarily due to the lack of consistent experimental data in the published literature. Yalkowsky and Banerjee have summarized the various methods used to develop solubility models.[12] Below we list some of the more recent efforts along these lines. The interested readers are referred to the original articles and the references therein for earlier publications in this area.

Work by Jurs's group utilized topological, electronic, and 3-D geometrical descriptors to represent the structural features.[13−16] In these publications, the electronic properties such as atomic charges were computed using semiempirical molecular orbital calculations. Bodor and Huang used a semiempirical method to calculate various molecular properties to fit a linear equation to reproduce the experimental solubility.[17] Katritzky and co-workers developed solubility models for hydrocarbons and halogenated hydrocarbons based on the physicochemical properties obtained from semiempirical quantum calculations.[18] All these methods work reasonably well for the prediction of the solubility of small organic compounds. However, calculations involving quantum approaches are rather time-consuming tasks.

Yalkowsky has developed a method to predict the solubility for several series of structurally related compounds by correlating solubility with experimental water−octanol partition coefficients and melting points.[19] Meylan et al expanded this method by including molecular weight (MW).[20] The approaches based on experimental properties (such as logP and melting temperature) are only suitable for compounds for which the measured values are available. This method is not applicable to compounds yet to be synthesized, thus, making it impossible to use in the design of drug-like virtual libraries.

Group contribution methods have been employed for the prediction of solubility on several occasions.[21−23] This type of method often requires numerous parameters to achieve a good predictive model. For example, 40−200 parameters are not uncommon.[21−23] Kuhne et al employed a fragmentation method where the experimental melting point was also considered as a fragment.[24] Recently, Abraham and Le used linear solvation energy relationship for the prediction of solubility.[25] Ruelle and Kesselring applied the mobile order thermodynamics method to compounds with no hydrogen bond donor capacity.[26]

Huuskonen and Taskinen developed an artificial neural network model based on molecular topology and E-state keys for the prediction of solubility of drug and drug-like compounds.[27] However, the experimental solubility of the molecules included in the training set was measured under different experimental conditions, thereby making it difficult to interpret the predicted solubility. Recently, Huuskonen developed an artificial neural network model for a large diverse set of compounds.[28] The model used 30 descriptors based on 1-D and 2-D molecular information. The model gave excellent $R^2 = 0.94$ and SD = 0.47 log unit for a training set of 884 compounds. This model, using a 30−12−1 neural network architecture, has a total of 385 adjustable weights and is rather complex.

Tetko et al developed an artificial neural network model using a 33−4−1 architecture (a total of 141 adjustable weights).[29] Using a set of homogeneous descriptors, such as MW and electrotopological indices (E-states), the model gave comparable performance to that presented in Huuskonen's work.[28] Liu and So[30] used an even smaller neural network, 7−2−1 architecture (a total of 19 adjustable variables) on the same set of training set used by Huuskonen.[28] The model gave a reasonable $R^2 = 0.86$ and SD = 0.72.

More recently, Jorgensen and Duffy have developed an approach combining Monte Carlo simulation and QSPR method.[31] Separate regression equations are used for small organic compounds and drug molecules and, as a result, this scheme gives good prediction for a diverse set of molecules. However, Monte Carlo simulation is relatively time-consuming for the prediction of large collections of compounds. Recently, Ran and Yalkowsky[32] used a generalized solubility equation (GSE) to estimate the solubility of the same set of compounds studied by Jorgensen and Duffy. The inputs used in the GSE were the experimental melting points and calculated or experimental water−octanol partition coefficients**.** The method is simple, yet as accurate as the Monte Carlo approach.

In this work, we describe a fast quantitative−structure property relationship (QSPR) method for the prediction of aqueous solubility. We tried to focus on a small number of descriptors or parameters that have physical meaning and, as a result, are intuitive to bench chemists. Thus, our predictive model can be readily used in library design and optimization. Moreover, the models are fast enough to be used in data mining and in silico screening of large libraries.

The Data Set. There are fairly large collections of experimental solubility data for small organic compounds. However, only limited data were available for drug and drug-like compounds. Furthermore, the experimental conditions vary considerably from laboratory to laboratory, especially for the drug and drug-like compounds. For example, measurements at various pH's and temperatures were reported. These conditions affect the solubility of a compound. The solubility of a compound usually increases as the temperature is elevated and the solubility of many compounds can be easily altered by adjusting the pH of the solution.[1] Furthermore, for a given condition (T and/or pH), the solubility measurement may be influenced by the experimental protocol. For example, purity of the material, especially, the solute, and the length of equilibration time can be critical. Thus, experimental conditions and experimental protocols can all lead to inter-laboratory variation in the measurement. These issues are beyond the scope of this

paper and are thoroughly discussed in an excellent book by Yalkowsky.[1] The interested reader may consult the original book for further discussion. Any model developed based on the available experimental solubility data is limited by the accuracy of experimental measurements. Unfortunately, most of the experimentally determined solubility values do not report the standard error for the measurements. This made it impossible to assess the quality of the experimental data. Provided that reasonable experimental procedures are followed the experimental measurements from different laboratories should agree with each other within an acceptable accuracy. Comparison of the measured solubility from different research groups can give us a rough idea about the quality of the data. The comparison of a small collection of hydrocarbons by Huibers et al shows that the error is at least 0.16 log units.[18] Hence, for a collection of diverse compounds, like the one used here, the errors might be significantly greater.

The experimental conditions under which solubility has been determined vary widely (e.g., temperature, pH, etc. are experimental variables). Because of this, we decided upon one experimental set (T = 298 K, unbuffered solution) of conditions that would give us a clean and consistent data set. We focused on the thermodynamic solubility that is defined as the equilibrium concentration of the chemical in a saturated solution. For the solid, the thermodynamic solubility reflects the true solubility of the most stable crystal form; therefore, the solubility is not affected by the crystal form. Different crystal forms may have different dissolution rates which results in a temporary change in the apparent solubility only.[1] We only concentrated on the aqueous solubility of pure neutral species. We did not include mixtures and formulated drug products because, indeed, formulation ingredients can alter the solubility significantly.[33] Furthermore, the formulation design of drug products depends on the nature of the individual compound, the route of administration, and many other factors.[33] Thus, the highly variable nature of the formulation procedure reduces the value of the inclusion of formulation dependent solubility data in a general solubility model. Others have attempted to fit variable conditions to a QSPR model, but it is our opinion that this makes it even more difficult to model what is a very complex physical process.

We started with a literature data set that consisted of 330 small organic compounds.[13] This set was originally from Yalkowsky's AQUASOL database.[34] To obtain a more diverse data set, we expanded the set to 551 compounds by including compounds from several other sources.[15,17,25] Compounds with inconsistent information were eliminated. This small molecule data set covers a wide variety of functional groups (20% of the molecules had multiple functional groups) and solubility ranging from 10^{-12} to 10^{2} (mol/L). The highest MW is 500. However, MWs of only eight molecules are between 400 and 500. Examining the CMC and PDR compound collections showed that 12 and 8% of compounds, respectively, have MW over 500. To develop a model suitable for the prediction of a wider range of compounds, we added higher MW compounds (MW > 250) to the training set. Specifically, drug-like compounds and compounds with functionalities under-represented

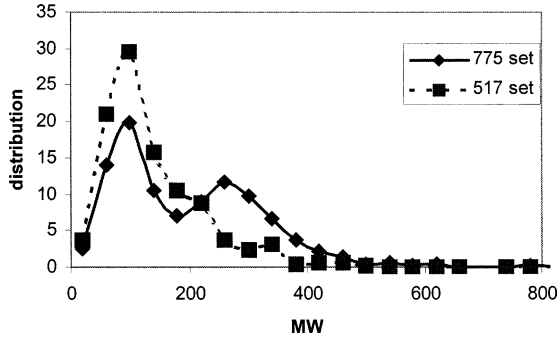

Figure 1. Molecular weight (MW) distribution for the two training sets. One contains 517 small molecules and the other has 775, including many molecules with multiple functional groups. The latter set includes more molecules with MWs greater than 250 and has 6 compounds with MWs between 600 and 800.

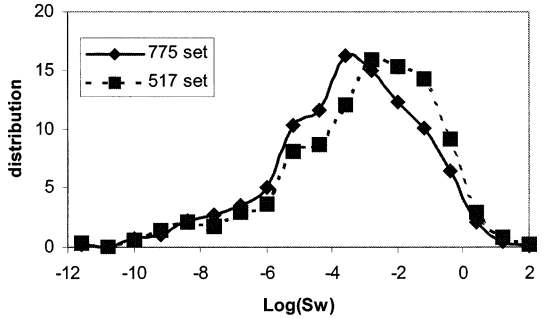

Figure 2. The distribution of aqueous solubility S_w (mol/L) for the two sets of data. The logarithmic scale is used to compress the data into a convenient range.

Table 1. Profile of Functional Groups in the Data Set

no. of compounds	functionality
28	alkanes
20	alkenes
9	alkynes
144	halogen derivatives
68	aromatic and cyclic
60	nitrogen-containing compounds: nitros, nitriles, amides
11	amines
57	alcohols
20	ketones
8	aldehydes
27	esters
14	ethers
20	acids
6	sulfur-containing compounds
59	drug and drug-like molecules
258	with multiple functional groups and MW > 250

in the small molecule data set were chosen to better cover the chemical space. Most of these compounds were from the PHYSPROP database.[35] We only included one enantiomer of a chiral compound since enantiomers have the same melting point, boiling point, spectroscopic properties, and the same solubility in water while only their interaction with plane-polarized light and their reactions and/or interactions with other chiral molecules are affected. The final data set (809 compounds) contained multiple functional groups, which are listed in Table 1. Many molecules with multiple functional groups were arbitrarily put in one of those categories since this table just gives us a rough idea about the distribution of compound functionality not an elaborate classification of all of the compounds. Figures 1 and 2

show the MW profile and the solubility distribution for the training set, respectively. The solubility is expressed in mol/L and a logarithmic scale was used to compress the solubility data into a more convenient range.

A small number of molecules (34) were randomly selected from 809 compounds and were kept outside the training set as a test set. We also collected several sets of data as independent external validation sets: 61 orally available drugs from the Physician's Desk References (PDR),[36] 166 drug-like compounds from the Comprehensive Medicinal Chemistry (CMC) database,[37] and 1404 compounds from the PHYSPROP database.[35] The 61 compound PDR data set was obtained by searching through the solubility data for all 438 orally available PDR drugs in archival data sources available to us (Merck Index,[38] AQUASOL,[34] and the PHYSPROP[35] database). We were able to find solubility data for 61 out of 438 compounds. Similarly, the 166 compound CMC data set (out of a total of 5836 drug-like CMC compounds) was obtained in a similar manner. The 5836 drug-like CMC compounds were extracted from 7577 entries in the CMC database as described in Egan et al.[11] These data sets were kept separate to allow us to monitor the performance of the model on different series of compounds, such as drug-like compounds (PDR and CMC series) and environmentally interesting compounds (PHYSPROP data set). There are many other ways to separate the validation data set based on one's particular goals and interests. It is important that the model perform well on the entire validation data set. We also predicted the solubility profile for 438 orally available drugs from the PDR and 5836 drug-like compounds from the CMC database. Since the solubility data are only available for a small fraction of these compounds, we compare the profile of the entire collection to that of the subset with known solubility data.

In summary, we used 775 compounds as a training set to develop models and 1665 external compounds for model validation. Huuskonen's model was based on 884 compounds and was validated on 442 external compounds.[28] It is impossible for us to compare our data sets with theirs since their paper did not disclose the compounds used in the model and it was not clear how the compounds were selected. It is likely that our compound list overlaps with theirs to some extent since both efforts shared common data sources (AQUASOL database and PHYSPROP database), but our dataset is significantly larger than the one reported by these authors.

Model Development. The Cerius^2 4.0 package[39] was used to build the QSPR models for the prediction of aqueous solubility. The structures were minimized using the universal force field implemented within the open force field (OFF) module of Cerius^2. All 136 descriptors available within Cerius^2 were computed, which included descriptors describing structural, topological,[40−42] electrotopological indices known as E-state keys,[43] and spatial parameters.[44,45] The structural descriptors include MW, number of rotatable bonds, number of hydrogen bond donors and acceptors. We also calculated the water−octanol partition coeffient, AlogP,[46] as implemented in Cerius^2 4.0. The spatial descriptors include Jurs's charged partial surface area parameters and shadow indices represented as the area of the projections of the molecular volume in three perpendicular planes.[44,45] The pairwise correlation coefficients reveal that many descriptors were highly correlated to each other. Among a group of highly correlated descriptors, only one was retained for regression. The cutoff for the correlation coefficient is 0.9 (i.e., if the correlation coefficient was greater than 0.9, they were considered highly correlated). Some descriptors have zero value for some compounds, and those descriptors that had very few nonzero values were eliminated since they only represented a small number of molecules. The remaining 30−40 descriptors were used for the regressions.

Several regression methods were used to build a model. G/PLS, which is a combination of a genetic algorithm and a partial least-squares method, was used to find the best combination of descriptors. Typically, G/PLS regression gives many models with similar accuracy. Each model was analyzed, and the most statistically significant group of descriptors was used for the final model. Stepwise regression was also used to find the most significant descriptors and to confirm the descriptors obtained from the G/PLS method. Multiple linear regressions were performed on the small set of descriptors selected by G/PLS and stepwise methods to build the final models. The regression results from different methods agree with each other well, so we only present the results from the multiple linear regressions because it gave us our final solubility model.

We developed models using two different training sets. The first model was based on the experimental solubility of 517 small molecules. We derive models using all 517 compounds in the training set instead of subdividing the training set into fairly homologous groups of compounds and then developing individual models on these subsets. Because it is generally easier to derive a model from chemically related compounds, the use of subsetting is quite common in models for various physical and biological properties. However, while this procedure can provide better fits of the training set data, it can dramatically decrease the accuracy of predictions for external compounds, especially those that do not fall clearly into any one particular subset.

In choosing the final regression model, we were concerned that the resulting equation was simple but accurate and retained terms that were as physically intuitive as possible. We also included electrotopological indices and 2-D topological connectivity indices to refine the model. Regressions were carried out using an increasing number of descriptors, and both coefficient of determination R^2 and leave-one-out (LOO) validated R^2 were monitored. The LOO validation was performed by holding one compound out and by developing a model based on the rest of the training set. The model was then used to predict the solubility of the compound being held out. This process was carried out for all of the compounds in the training set. The predicted values and the experimental values are used to calculate R^2. This parameter reflects the quality of the model on the validation data set. R^2 generally improves as the number of descriptors is increased. However, LOO validated R^2 increases first and then starts to decline as the number of descriptors increases further, which indicates overfitting. The final model is a linear model

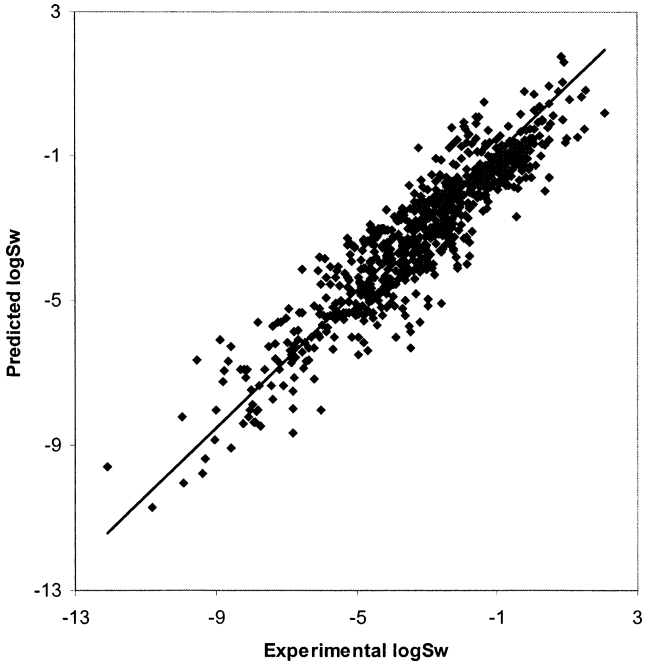

Figure 3. The predicted vs experimental solubility for the training set of 775 compounds.

using eight descriptors that do not require the 3-D structure of a molecule. The regression model based on eight descriptors has a coefficient of determination $R^2 = 0.892$ and root-mean-square error (RMSE) = 0.75 for the compounds in the training set.

This model was validated on several external data sets. It was noticed that the errors for high MW compounds were slightly greater than that observed for low MW molecules. It was determined that this was likely because the MW of the larger molecules was outside of the range of the training set. This led to the generation of a second training set of 775 compounds that had an increased number of higher MW drug and drug-like compounds with multiple functional groups.

The regression procedure described above was also used to develop the second solubility model. An eight-descriptor regression gave the optimal LOO validated R^2. This model has a coefficient of determination, $R^2 = 0.84$, which is slightly lower than the previous model, the unsigned error = 0.68, the RMS = 0.87, the LOO validated $R^2 = 0.833$ and $F_{8,766} = 502.4$. The regression equation is as follows:

$$\log(S_w) = -0.7325*\langle \text{AlogP98} \rangle - 0.4985*\langle \text{HBD} \rangle *\langle \text{HBA} \rangle - 0.5172*\langle \text{Zagreb} \rangle - 0.0780*\langle \text{S\_aaaC} \rangle + 0.1596*\langle \text{Rotlbonds} \rangle + 0.2057*\langle \text{HBD} \rangle + 0.1834*\langle \text{S\_sOH} \rangle + 0.2539*\langle \text{Wiener} \rangle$$

where <x> represents the mean-centered, unit-variance scaled value of each descriptor based on the training set values. The coefficients then illustrate the relative importance of each descriptor to show the significance of each descriptor; the variables are mean-centered, and then scaled accordingly. Figure 3 shows the predicted solubility versus experimental solubility. The slightly lower correlation compared to the previous model is probably related to the data quality of some of the higher MW compounds. In this equation, all the descriptors are very significant as indicated by the *F* ratio and the significance probability as shown in Table 2. The average of 28 pairwise correlation coefficients between eight descriptors for the entire training set is 0.37 and SD = 0.31. Three pairs have correlation above 0.8 (they are 0.822, 0.824, and 0.864), correlation coefficients for four pairs are between 0.6 and 0.8 and those for 21 pairs are less than 0.6. The quality of this simple linear model with eight descriptors is comparable with that of the recent neural network model,[30] which involves 19 adjustable variables and has slightly better $R^2 = 0.86$. With the neural network method, it is sometimes difficult to interpret the physical meaning of the model and relative importance of the descriptors.

Table 2. *F* Ratio and the Significance Probability for Each Descriptor in the Regression Equation

descriptors	*F* ratio	prob > *F*
AlogP98	1332.23	<0.0001
HBD*HBA	160.86	<0.0001
HBD	42.69	<0.0001
Rotlbonds	73.84	<0.0001
Wiener	88.21	<0.0001
Zagreb	370.40	<0.0001
S_aaaC	23.35	<0.0001
S_sOH	36.53	<0.0001

The AlogP98 descriptor was the most significant contributor to the solubility prediction. It appears in both the 517- and the 775-compound training set model. It is not surprising that this term carries a negative sign since solubility and water−octanol partition coefficient represent two opposite properties of a compound: hydrophilicity and hydrophobicity. The greater the AlogP98, the more hydrophobic the compound, and the lower the solubility. In fact, the logP parameter has been used to model the solubility of homologous and congeneric series of compounds.[47,48]

The descriptor, (HBD*HBA), proportional to the product of the number of hydrogen bond donors (HBD) and acceptors (HBA), is also very significant for the 517- and 775-compound training set. We interpret this parameter as reflecting intermolecular hydrogen bonding between two solute molecules. Thus, this variable can be interpreted as representing the solid-state cohesive energy or crystal packing forces to some extent. The more intermolecular hydrogen bonding, the stronger the molecules are bound to one another in the crystalline phase, and, hence, the lower the solubility. To test this hypothesis, we selected two subsets of compounds from the training set. One group consisted of solids and the other group contained liquids. Two separate models were developed based on each subset. It was found that the HBD*HBA term was present in the model based on solids, but was absent in the model based on liquids. This indicates that the HBD*HBA term is likely correlated with the crystal packing energy, at least from the statistical point of view. To solely attribute this term as representing crystal packing is an oversimplification of the complex interactions present in solids.

The number of hydrogen bond donors (HBD), present in both models, is interpreted to reflect the hydrogen bonding capacity of a solute with water. The electronic state key, S_sOH, appears in both models and is closely related to the hydrogen bonding ability of a molecule. The hydroxyl group OH is both a hydrogen bond donor and acceptor. The more hydrogen bond donors and

acceptors the more soluble a compound is likely to be. Attempts to interpret all parameters in a QSAR equation may lead to an oversimplification of the physical processes being represented. Thus, it is not possible to unambiguously determine if a descriptor is clearly responsible for enhancing or decreasing the solubility of a molecule. For example, hydrogen bond donors and acceptors are important for both molecule−molecule interactions in the crystalline phase and solute−solvent interactions, but we have interpreted the HBD*HBA term as modeling solute−solute interactions, while the HBD term models solute−solvent. This may not be entirely correct, but we believe it is the most plausible interpretation. Most importantly, the descriptors, HBD*HBA, HBD, and S_sOH, representing somewhat different aspects of hydrogen bonding interactions, work synergistically with the other parameters to explain the variation of the solubility of the training set.

The number of rotatable bonds, Rotlbonds, did not appear in the 517-compound model. In light of this, we believe that it is related to the large MW compounds added to this bigger data set. Hence, a parameter closely related to the size of a compound, such as number of rotatable bonds, was necessary to account for the wide MW range.

The other two geometrical indices, Wiener and Zagreb also correlate with the size of the molecule. These two parameters, along with the number of rotatable bonds, account for the molecular size dependency of solubility. Here, the opposite contribution of the Zagreb term might be necessary to balance out the overcompensation of Rotlbonds and Wiener.

E-state key, S_aaaC, is related to the aromatic nature of a compound. The two geometrical indices, Wiener and Zagreb, and E-state key, S_aaaC and S_sOH were used to further refine the model. The interpretation of the physical significance of each term is not straightforward. As mentioned above, mechanistic interpretation of individual terms in a regression equation may not always result in an entirely meaningful conclusion. However, the descriptors as a group gave the best explanation of the variation of the dependent variable. The weight associated with each descriptor depends on the composition of the training set of compounds. If a training set covers the chemical space uniformly, the resulting equation is likely to reflect the mechanism of solvation.

Interestingly, regressions including descriptors based on the 3-D structural information did not lead to significant improvement. Considering that these descriptors are more time-consuming to compute, we think that a model using descriptors based only on molecular connectivity (or 2-D structure) is a better choice, especially since we were interested in developing a model that could be applied to large (multi-million molecules) libraries and virtual libraries. Regressions using nonlinear terms other than HBD*HBA did not improve the model.

Model Validation. The final model was used to predict the solubility of several sets of external compounds (i.e., these compounds are not included in the training set). These data sets cover very diverse compounds: some small organic molecules, 61 orally available compounds from the PDR, 166 drug-like compounds from the CMC database, and a diverse collection of 1404 compounds from the PHYSPROP database. The PHYSPROP data set has many environmentally relevant chemicals, such as pesticides, herbicides, insecticides, and other industrial chemicals. The validation results on these data sets are shown in Table 3. The quality of the data varies from one data set to another. The solubility for CMC compounds shows the greatest variability and this is reflected by RMSE = 1.15. However, there are not a lot of data available for drug and drug-like compounds, and we kept most available data for use in the validation. Since these data are not used in the training set, they do not affect the quality of the model. The overall predictability of the model is very satisfactory with a combined RMS error of 1 log unit and unsigned error 0.77 (1665 compounds). Figure 4 plots the number of compounds as a function of the absolute deviation of the predicted solubility from the experimental value. The plot shows that if one can tolerate an error of 1 log unit, the model can correctly predict the solubility for 60−80% of the compounds. If one can tolerate 1.5 log unit of error, the model can correctly predict the solubility for 80−95% compounds. Only 5% of the predictions have errors greater than 2 log units.

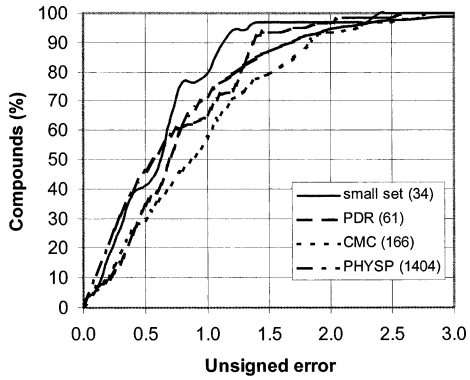

Figure 4. The number of compounds within a given deviation from the experimental value. This shows the fraction of compounds this model is able to correctly predict for a given error.

Table 3. Predicted Error for the 775 Molecule Training Set Model Relative to the Validation Data Sets

data set	no. of cmpds	signed error	unsigned error	RMS error
training set	775	0.00	0.68	0.87
small molecules	34	0.17	0.64	0.62
PDR	61	−0.28	0.80	0.95
CMC	166	−0.02	0.95	1.15
PHYSPROP	1404	−0.01	0.75	1.01
all validation compounds	1665	−0.02	0.77	1.01

Predicting solubility for drug-like compounds is of great importance for drug discovery. However, the validation of the predictive model is limited by the availability of experimental measurements for drug-like compounds. We predicted the solubility for 438 orally available drugs from the PDR and 5836 drug-like compounds from the CMC database. Since the experimental values are only available for a limited number of drugs, we chose to use the distribution of the experimental solubility for 93 orally available PDR compounds as a reference. Here, 61 of 93 compounds were the same as those in the previously discussed external validation set, 15 of the 93 compounds were in the training set, and 17 were obtained from searches of

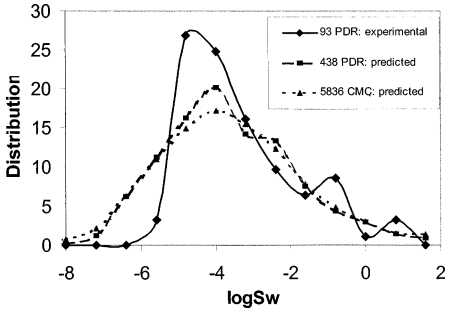

Figure 5. The distribution of the predicted solubility for 438 orally available PDR compounds and for 5836 drug-like compounds from the CMC database. This is compared to the experimental solubility distribution for 93 orally available PDR compounds.

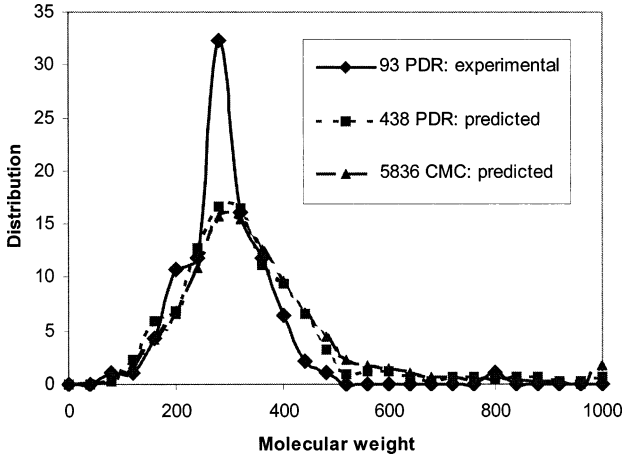

Figure 6. The distribution of MWs for 438 orally available PDR compounds and for 5836 drug-like compounds from the CMC database compared to that of 93 orally available PDR compounds.

other sources. The profile for these compounds indicates that the solubility for 95% of orally available compounds falls between −6 and 0 on the logarithmic scale. The distribution of the predicted solubility for the 438 orally available PDR compounds and 5836 drug-like CMC compounds overlaps well with that of the experimental profile for the reference compounds (see Figure 5). At the lower end of the solubility scale, the predicted solubility distributions drop less abruptly than that of the experimental profile. Comparison of the distributions of the MW for the two data sets with that of the reference set shows that high MW compounds are slightly under-represented in the reference set (see Figure 6). This might explain the discrepancy between the solubility distributions at the low solubility region. Solubility values of 11% of the PDR compounds are outside the 95% range (−6 to 0) of the experimental profile. The solubility distribution of the CMC compounds is slightly wider. 15% of the compounds are outside the 95% drug range. This is not surprising given that the CMC database includes many drugs administrated via forms other than oral; therefore, compounds with lower solubility than orally administered drugs are possible.

Application to Data Mining and Library Design. Here, we demonstrate how this type of model can be used in the compound and library design. These models were used to predict solubility of several Pharmacopeia

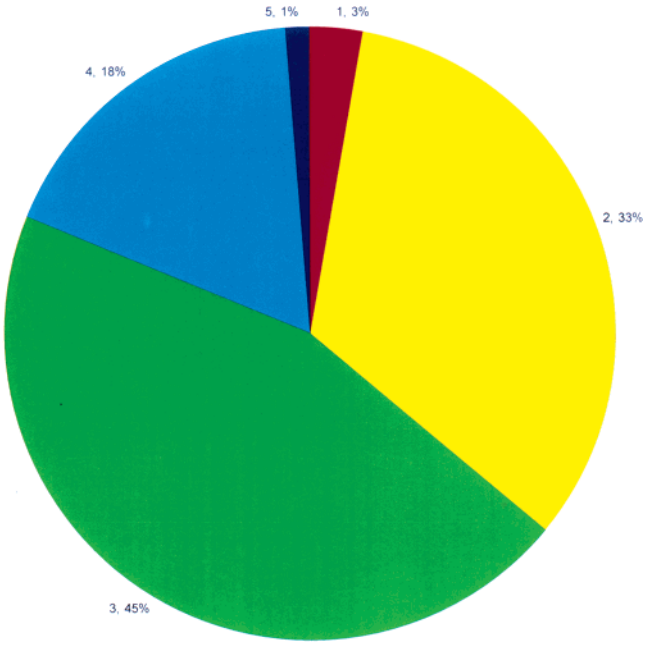

Figure 7. The solubility profile of a library. Black, red, yellow, green, cyan, and blue are for ranking 0, 1, 2, 3, 4, 5, respectively. The first number next to each slice is the ranking and the second is the percent of compounds in this ranking. For this library, there are no compounds with ranking 0.

Table 4. Definition of Solubility Ranking

solubility ranking	solubility value	drug-likeness
0	$\log(S_w) < -8$	no, impossible
1	$-8 < \log(S_w) < -6$	no, very low, but possible
2	$-6 < \log(S_w) < -4$	yes, low
3	$-4 < \log(S_w) < -2$	yes, good
4	$-2 < \log(S_w) < 0$	yes, optimal
5	$0 < \log(S_w)$	no, too soluble

libraries. These models can compute the solubility for 25−200 compounds/s depending on the composition of the library. In another words, it can compute solubility for 1 million compounds in 1.4−11 h on a SGI workstation. The computation time is quite acceptable and can be used for data mining and library design purposes.

This quantitative structure−property relationship model is able to predict the solubility for each compound with an average error of 0.77 log unit. This value can be used to prioritize and optimize the synthesis in the rational design. For a design of a large combinatorial library, the predicted solubility value of all compounds is too overwhelming to be analyzed. Therefore, we constructed a ranking scheme based on the solubility distribution of available drug-like compounds and the standard deviation of the predictive model. The ranking and the definition are listed in Table 4. We tried to use several rankings instead of just two (drug-like vs non-drug-like) because we think the model should be used to provide a solubility profile for chemists to select compounds for further pursuing. The higher the ranking, the more soluble a compound. The compound in the ranking zero is clearly too insoluble to be considered further in the discovery effort. The compound with solubility ranking 1 is only slightly soluble; however, there are some marketed drugs in this range. Ranking 2, 3, and 4 are all in the drug-like category. Though ranking 5 is clearly extremely soluble, not many marketed drugs are found in this category. As mentioned

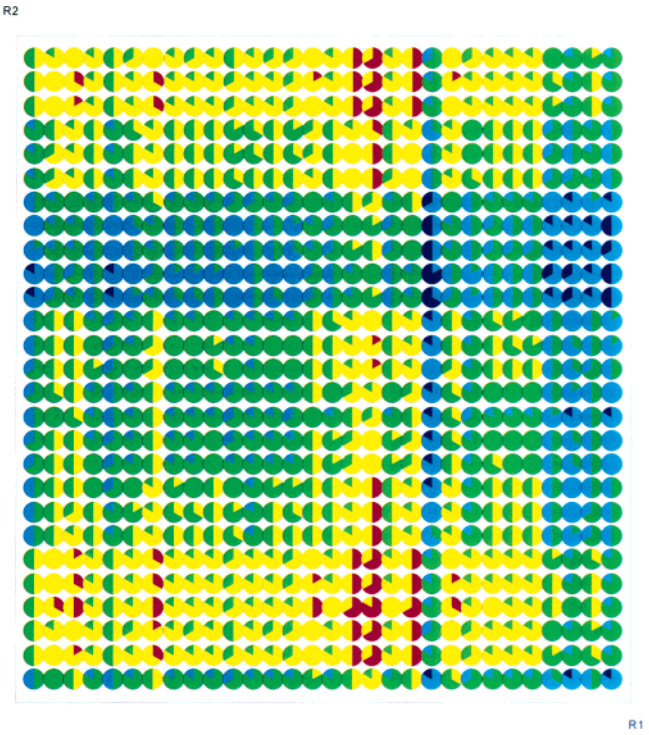

Figure 8. The solubility profile for sublibraries as a function of R1 and R2 substituents. The color scheme is the same as in Figure 7. This plot shows how R1 and R2 building blocks influence the solubility profile. For example, sublibraries with R1 = 17, 18 or R2 = 21 are very soluble and those with R1 = 4 or R2 = 18 show poor solubility. Sublibraries with the combination of R1 = 18 and R2 = 18 are reasonably soluble.

earlier, a good drug has to balance many physicochemical and pharmacokinetic properties. The extremely soluble compounds can be too hydrophilic to penetrate the biological membrane.

With this kind of ranking scheme, one can obtain a profile for any synthesized or virtual library. Figure 7 shows a pie chart of solubility distribution of a library. To aid library design effort, it is important to understand the relationship between structural characteristics and the solubility if possible. Figure 8 shows the solubility profile of sublibraries as a function of R1 and R2 substituents. From the plot, one can see how each substituent contributes to the solubility. R1 and R2 substituents also interact with each other. In another words, a compound with a particular R1 or R2 may not be very soluble, but a compound with a combination of specific R1 and R2 can be very soluble or vice versa. This information can help chemists to select the specific building blocks in the library design, thus, optimizing the solubility profile.

Conclusions

In conclusion, we were able to develop a fast, yet reasonably accurate, model for the prediction of aqueous solubility based on a diverse set of literature data. The RMS error for 1665 test compounds is 1 log unit and unsigned error is 0.77. The quality of the model was limited by the accuracy of the experimental measurements on the training set compounds. These models do not require the 3-D structure of a compound and are, therefore, suitable for use in data mining and library design applications. This model is based on a diverse set of compounds, and, therefore, can be used to predict the solubility for pharmaceuticals and compounds of general interest.

Acknowledgment. We would like to thank I-Ping Cheng for her assistance in collecting experimental solubility data. We thank Jack Baldwin for his generous support and many insightful discussions. A. Cheng would like to thank Steve Dixon for his support and many valuable suggestions.

References

(1) Yalkowsky, S. H. *Solubility and Solubilization in Aqueous Media*; American Chemical Society and Oxford University Press: Washington, DC, New York, Oxford, 1999.
(2) PhRMA. Annual Report 2001−2002: New Medicines New Hope. http://www.phrma.org/publications/publications/annual2001/innovation.phtml, 2001.
(3) DiMasi, J. A. Success Rates for New Drugs Entering Clinical Testing in the United States. *Clin. Pharmacol. Ther.* **1995**, *58*, 1−14.
(4) Prentis, R. A.; Lis, Y.; Walker, S. R. Pharmaceutical Innovation by the Seven UK-owned Pharmaceutical Companies (1964−1985). *Br. J. Clin. Pharmacol.* **1988**, *25*, 387−396.
(5) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2001**, *44*, 235−249.
(6) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(7) Walters, W. P.; Murcko, M. A. Library Filtering Systems and Prediction of Drug-like Properties. *Methods Principles Med. Chem.* **2000**, *10*, 15−32.
(8) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-Based Design: Optimization of Drug Absorption and Pharmacokinetics. *J. Med. Chem.* **2001**, *44*, 1313−1333.
(9) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.
(10) Dixon, S. L.; Merz, K. M., Jr. One-Dimensional Molecular Representations and Similarity Calculations: Methodology and Validation. *J. Med. Chem.* **2001**, *44*, 3795−3809.
(11) Egan, W. J.; Merz, K. M., Jr.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.
(12) Yalkowsky, S. H.; Banerjee, S. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Dekker: New York, 1992.
(13) Mitchel, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.
(14) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.
(15) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds using a Quantitative Structure−property Relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.
(16) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.
(17) Bodor, N.; Huang, M.-J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954−960.
(18) Huibers, P. D.; Katritzky, A. R. Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.
(19) Yalkowsky, S. H.; Valvani, S. C.; Roseman, T. J. Solubility and Partitioning VI: Octanol Solubility and Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1983**, *72*, 866−870.
(20) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Octanol/water Partition Coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.
(21) Wakita, K.; Yoshimoto, M.; Miyamoto, S.; Watanabe, H. A Method for Calculation of the Aquous Solubility of Organic Compounds by Using New Fragment Solubility Constants. *Chem. Pharm. Bull.* **1986**, *34*, 4663−4681.
(22) Suzuki, T. Develpment of an Automatic Estimation System for Both the Partition Coefficient and Aqueous Solubility. *J. Comput.-Aided Mol. Design* **1991**, *5*, 149−166.
(23) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.

(24) Kuhne, R.; Ebert, R.-U.; Kleint, F., Schmidt, G.; Schuurmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.
(25) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(26) Ruelle, P.; Kesselring, U. W. Aqueous Solubility Prediction of Environmentally Important Chemicals from the Mobil Order Thermodynamics. *Chemosphere* **1997**, *34*, 275−298.
(27) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(28) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(29) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(30) Liu, R.; So, S.-S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation in Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.
(31) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.
(32) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.
(33) Shargel, L.; Yu, A. B. C. *Applied Biopharmaceutics and Pharmacokinetics*; Appleton & Lange: Stamford, Connecticut, 1993.
(34) Adb, the AROZONA dATAbASE of Aqueous Solubility, Yalkowsky, S. H., College of Pharmacy, University of Arizona, Tucson, AZ, 1997.
(35) Physical/Chemical Property Database (PHYSPROP), Syracuse Research Corporation, SRC Environmental Research Center, Syrucuse, NY, 1999.
(36) Physicians' Desk Reference: Electronic Library; Medical Economics Co., Inc.: Montvale, NJ, 1999.
(37) Comprehensive Medicinal Chemistry; MDL Information Systems, Inc.: San Leandro, CA, 1999.
(38) The Merck Index, Version 12.2 CD-ROM, Chapman & Hall, 1998.
(39) Cerius2., Accelrys, San Diego, CA, 2000.
(40) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: New York, 1986.
(41) Katritzky, A. R.; Gordeeva, E. V. Traditional Topological Indices vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(42) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(43) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(44) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/activity and Structure/property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.
(45) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(46) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP adn CLOGP methods. *J. Phys. Chem. A* **1998**, *102*, 3762−3772.
(47) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The Linear Free-Energy Relationship between Partition Coefficients and the Aqueous Solubility of Organic Liquids. *J. Org. Chem.* **1968**, *33*, 347−350.
(48) Valvani, S. C.; Yalkowsky, S. H.; Roseman, T. J. Solubility and Partitioning IV: aqueous Solubility and Octanol−Water Partition Coefficients of Liquid Nonelectrolytes. *J. Pharm. Sci.* **1981**, *70*, 502−507.

JM020266B