Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas

Junmei Wang,*,† Tingjun Hou,‡ and Xiaojie Xu*,§

Department of Pharmacology, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, Texas 75390-9050, Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92093, and College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, P.R. China

Received November 6, 2008

In this work, four reliable aqueous solubility models, ASM-ATC (aqueous solubility model based on atom type counts), ASM-ATC-LOGP (aqueous solubility model based on atom type counts and *ClogP* as an additional descriptor), ASM-SAS (aqueous solubility model based on solvent accessible surface areas), and ASM-SAS-LOGP (aqueous solubility model based on solvent accessible surface areas and *ClogP* as an additional descriptor), have been developed for a diverse data set of 3664 compounds. All four models were extensively validated by various cross-validation tests, and encouraging predictability was achieved. ASM-ATC-LOGP, the best model, achieves leave-one-out correlation coefficient square (q^2) and root-mean-square error (*RMSE*) of 0.832 and 0.840 logarithm unit, respectively. In a 10,000 times 85/15 cross-validation test, this model achieves the mean of q^2 and *RMSE* being 0.832 and 0.841 logarithm unit, respectively. We believe that those robust models can serve as an important rule in druglikeness analysis and an efficient filter in prioritizing compound libraries prior to high throughput screenings (HTS).

INTRODUCTION

Aqueous solubility is one of the major drug properties to be optimized in drug discovery. Aqueous solubility and membrane permeability are the two key factors that affect a drug's oral bioavailability.[1,2] Generally, a drug with high solubility and membrane permeability is considered exempt from bioavailability problems. Otherwise, it is a problematic candidate or needs careful formulation work. Recently, we have found that the real drugs are about 20 times more soluble than those "druglike" molecules in the ZINC database.[3] The "druglike" molecules are those passing the widely known filter of Lipinski's "Rule of five", which states that a good drug candidate should have a molecular weight smaller than 500, a calculated logP (*ClogP*) smaller than 5.0, and the number of hydrogen bond donors and acceptors less than 5 and 10, respectively.[4] The aqueous solubilities were predicted using a reliable model called ASMS-LOGP.[5] In a virtual screening study using logS, the logarithm of aqueous solubility in mol/L, as the filter, less than 50% of "druglike" molecules in the ZINC database passed the threshold of logS being larger than −5.0, while 85% real drugs passed. This result indicates that an aqueous solubility model can serve as a screening filter to prioritize compound libraries in drug lead optimization.

Because of the importance of aqueous solubility, a lot of effort has been put into developing reliable models to predict this physiochemical property. Although some progress has been made and a lot of models have been constructed, very few models, no matter implemented in commercial packages or in public domain, have satisfactory predictability. Recently, Goodman et al. sponsored a solubility prediction competition in conjunction with the *Journal of Chemical Information and Modeling* to predict the solubilities of 32 molecules.[6] Another data set of 100 druglike molecules with experimental solubilities was provided as the training set. The goal of the competition was to find out how well solubility could be predicted with all kinds of means. Unfortunately, the summary of this contest had not been released at the time of this writing. However, the performance of recently published models, the algorithms beyond those models, were reviewed by Lipinski et al.,[4] Jorgensen et al.,[7] Wang et al.,[5] and Hou et al.[2]

What are the challenges in aqueous solubility prediction? We believe the absence of high-quality data is one of the major reasons causing the current models lack of high predictability. Most models were developed with training sets fewer than 2000 molecules, such as the models reported by Jain and Yalkowsky,[8] Klopman and Zhu,[9] Hou et al.,[10] Mitchell and Jurs,[11] Huuskonen,[12] Tetko et al.,[13] Liu and So,[14] Yan and Gasteiger,[15] Wang, Hou, and Xu et al.,[5] and so on.[2,5] Although the performance of the above models is encouraging and the *RMSE* are typically smaller than 0.80 logarithm unit for their own data sets, those models may experience a significant performance drop when being utilized for external data sets. For example, we developed a similar model to that of Jain and Yalkowsky using the same data set; that model, ASM-TM, gave a poor prediction for a set of 82 external molecules (Data Set 5), and the *RMSE* increased from 0.720 of the training set to 1.374 logarithm unit of the test set. More details on this model are presented in the following sections.

* Corresponding author e-mail: junwang@yahoo.com (J.W.) and xiaojxu@pku.edu.cn (X.X.).
† University of Texas Southwestern Medical Center at Dallas.
‡ University of California at San Diego.
§ Peking University.

10.1021/ci800406y CCC: $40.75
Published on Web 02/19/2009

Table 1. Correlation Coefficient Squares between Predicted Solubility by Different Models for the 1210-Molecule Beilstein Data Set

model	ACDLAB	Hou et al.	ASMS	ASMS-LOGP	VOLSURF
ACDLAB	1.000	0.702	0.674	0.748	0.479
Hou et al.		1.000	0.767	0.774	0.524
ASMS			1.000	0.915	0.588
ASMS-LOGP				1.000	0.558
VOLSURF					1.000

Recently, prediction models based on large data sets are beginning to emerge. For example, Delaney developed an aqueous solubility model for a 2874-compound data set using nine simple descriptors which included the calculated logP, molecular weight, aromatic proportion, noncarbon proportion, polar surface area, and so on.[16] Votano and Parham constructed a set of models using topological structure indices as descriptors for a data set consisting of 4115 aromatic compounds.[17] Naturally, the larger a data set is, the more reliable the models based on it are. Unfortunately, the performance of models based on large data sets is typically not as good as that of models based on small data sets. The *RMSE* of the aforementioned two models are both larger than 1.0 logarithm unit.

To test the performance and reliability of some popular aqueous solubility models, we recently conducted intrinsic solubility prediction for a set of 1210 molecules extracted from the Beilstein database using five models/packages, namely, the ACD/Laboratories solubility module (Version 11),[18] the Volsurf solubility model implemented in Sybyl 7.0,[19] the Hou et al.'s model,[10] and our two models (ASMS and ASMS-LOGP) recently published.[5] Surprisingly, all five models gave poor predictions, and the *RMSE* of logS were almost doubled in comparison to the reported *RMSE* of individual models. The correlation coefficient squares between the predicted solubilities by different models are listed in Table 1. The highest correlation occurs between ASMS and ASMS-LOGP, and the r^2 for the other model pairs cover the range from 0.4 to 0.77. Considering no model can be efficiently replaced by the others, for each compound, a consensus score, which was defined as the mean of the predicted solubilities with the five models, was calculated. The clustered column representation of the prediction residuals is shown in Figure 1. Those having residuals larger than 3.2 or smaller than −3.2 logarithm units are classified as the most suspicious data (in red); those having residuals ranging from −3.2 to −1.6 and 1.6 to 3.2 are less suspicious data (in yellow); and the others are classified as "good" data (in green). Then the experimental data, especially those in the red and yellow zones, were verified using the original publications. Interestingly, more than 90% of data in the red zone and about 50% of data in the yellow zone were erroneous. After the correction, the *RMSE* values decrease significantly for the sanitized data set, which now are 1.164, 1.366, 1.303, 1.281, and 1.381 logarithm units for the ACD/Laboratories model, the Hou's model, ASMS, ASMS-LOGP, and the Volsurf solubility model, respectively. The *AUE* and *RMSE*, both before and after the experimental data correction, are listed in Table 2 for the five models. Evidently, more robust models are needed to make a satisfactory prediction since those corrected *RMSE* are still considerably larger than the reported ones, which are roughly 0.6−0.8 logarithm units.

The purpose of this work is to develop a set of robust models to predict aqueous solubilities not only for the molecules in the training set but also for those in the external data sets. The reliability and predictability are evaluated through a set of stringent cross-validation experiments. To achieve the goal, a set of high-quality solubility data from a variety of sources will be collected. The models will then be optimized to reduce the *RMSE* by seeking a different combination of descriptors. Finally, leave-one-out analysis and 85/15 cross-validation analysis as well as scrambling analysis will be carried out to thoroughly evaluate the developed models.

Another goal of this study is to develop some models, which may not necessarily have good performance, but only utilize a limited number of descriptors, such as *ClogP*, the temperature of melting point to enable the simplicity of the models. Those models can be used by organic chemists or medicinal chemists to fast estimate the solubility of a compound or the solubility change between the precursor compound and its derivatives.

The aqueous solubility models reported in this work should have great applications in drug design: they can be used to predict a molecule's aqueous solubility prior to its synthesis; they can be also applied as a rule to define druglikeness; and they can serve as a filter to prioritize the compounds in a database for high throughput screenings.

METHODS

1. Experimental Data Source. There are five sources of experimental solubility data applied in this study. Data Set 1, which is the training set used to develop ASMS and ASMS-LOGP models in the previous study,[5] has 1708 high quality data. Most data in this data set are from the low molecular weight subset (1144) of the Delaney data set[16] and the Huuskonen data set.[12] Data Set 2 is a clean version of the data set used by Jain and Yalkowsky.[8] After eliminating the duplicated and obviously wrong entries, 578 data are left. Data Set 3, which has 1210 data, is the sanitized version of the Beilstein data set. Data Set 4 is a subset of the Goodman's data set for solubility challenge.[6] All 90 molecules in this data set had experimental intrinsic logS. Data Set 5 is a collection of molecules that have experimental T_m. The 119 molecules in this data set are selected from some recent publications.[20−29]

The canonical SLNs (Sybyl Line Notation) of all the molecules in five data sets were generated and compared each other.[19] After eliminating the duplicates, 3664 molecules were left. The experimental data of the duplicated entries were adopted according to the following order of priority: Data Set1 > Data Set 2 > Data Set 4 > Data Set 5 > Data Set 3. For each data set, the compound names, experimental aqueous solubilities, *ClogP*, molecular polarizabilities, and the predicted solubilities as well as molecular structures in SLNs are listed in Tables S1−S5 in the Supporting Information. The distribution of the 3664 aqueous solubility data in a chemical space defined by molecular weight (*MW*) and *ClogP* is shown in Figure 2. As shown in the figure, the 3664 molecules are almost evenly distributed in the druglike chemical space, which has molecular weight from 50 to 550

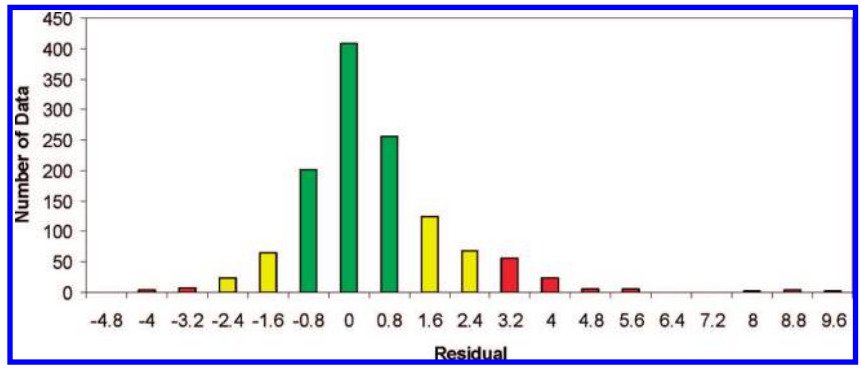

Figure 1. Clustered column representation of the residuals for the 1210 Beilstein data set. Average unsigned errors (*AUE*), root-mean-square errors (*RMSE*), and regression coefficient squares (q^2) of the 10,000 times 85/15 cross-validations for model ASM-ATC-LOGP: (a) *AUE*, (b) *RMSE*, and (c) q^2.

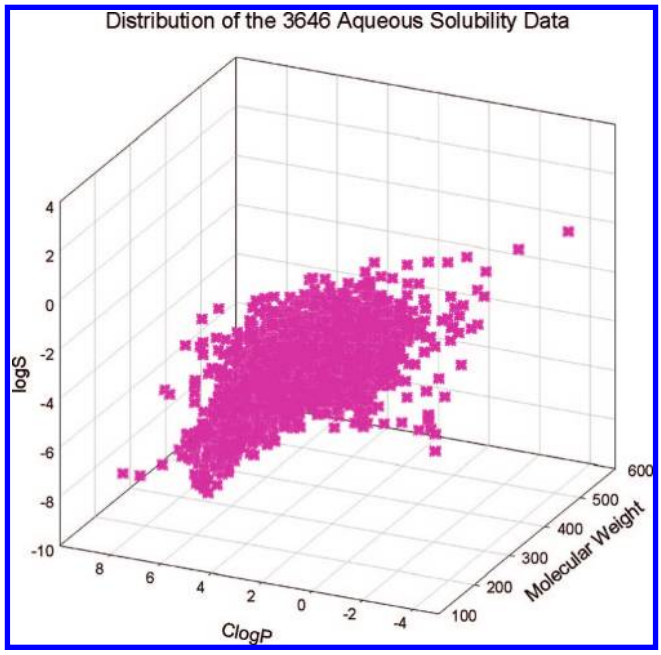

Figure 2. Distribution of aqueous solubility in a chemical space defined by molecular weight and *ClogP*.

Table 2. Performance of Prediction Using Five Published Models for the 1210 Beilstein Data Set

	after correction		before correction	
model	*AUE*	*rms*	*AUE*	*rms*
ACDLAB	0.837	1.164	1.012	1.514
Hou et al.	0.988	1.366	1.191	1.744
ASMS	0.982	1.303	1.170	1.657
ASMS-LOGP	0.937	1.281	1.129	1.642
VOLSURF	1.082	1.381	1.278	1.748
ASM-ATC-LOGP	0.745	0.978		
ASM-ATC	0.772	1.000		
ASM-SAS-LOGP	0.761	1.000		
ASM-SAS	0.791	1.027		

and *ClogP* from −4 to 8. The distributions of *MW* and *ClogP* by clustered columns are shown in Figures 3 and 4.

2. Descriptors. What are the proper descriptors to characterize aqueous solubility? To answer this question, we need to consider the major molecular interactions involved in the procedure of dissolving a compound in water. Aqueous solubility is almost exclusively dependent on the intermolecular adhesive interactions between solute−solute, solute−water, and water−water. The solubility of a compound is thus affected by many factors that include the size and shape of the molecule, the polarity and hydrophobicity of the molecule, and the ability of some groups to participate in intra- and intermolecular hydrogen bonding as well as the state of the molecule (for example, additional lattice energy is paid for a compound in the crystalline state to dissolve) etc. One may take those factors into consideration to select proper molecular descriptors to build up models to predict this important property.

Indeed, lots of successful models were developed using experimental or predicted physiochemical properties as descriptors, such as temperature of melting point, logP, polarizability, topological, and geometric indexes, etc. Considering one or several physiochemical properties may not well account for the interactions between solute−solute, solute−water, and water−water for different compounds, solubility models can be constructed based on the additive characteristics of aqueous solubility using eq 1. In this equation, c_i is the number of occurrence of atom type i or the total solvent accessible surface area of atom type i; w_i is its weight, which is determined by regression analysis; and c_0 is the constant. Practically, other molecular descriptors, such as molecular weight, molecular polarizability, and experimental or calculated logP, can also be incorporated into eq 1

$$\log S = \sum_{i=1}^{N} w_i c_i + c_0 \quad (1)$$

2.1. Atom Type Count (ATC). Atom type assignment was conducted with the Antechamber module of the Amber 8.0 software package.[30] Starting from ten basic atom types representing ten elements (C, N, O, S, P, H, F, Cl, Br, and I), prediction error sources were analyzed, and new atom types were introduced. Those new atom types were rejected if they could not or only marginally improve the fitting performance. This procedure was repeated time and again, and finally fifty atom types were defined. The fifty atom types together with their definitions are listed in Table 3.

2.2. Atom Type Classified Solvent Accessible Surface Area (SAS). To calculate the SAS of a molecule, a 3D-structure was first generated by the Concord module of Sybyl7.1.[19] Then an internal program developed by ourselves was applied to calculate the solvent accessible surface area of each atom. The element-based radius parameters (in Å) are listed as follows: H − 1.2, C − 1.74, N − 1.54, O − 1.40, S − 2.0, P − 2.0, F − 1.60, Cl − 1.79, Br − 2.04, I − 2.15. To penetrate the molecular surface deeper to explore

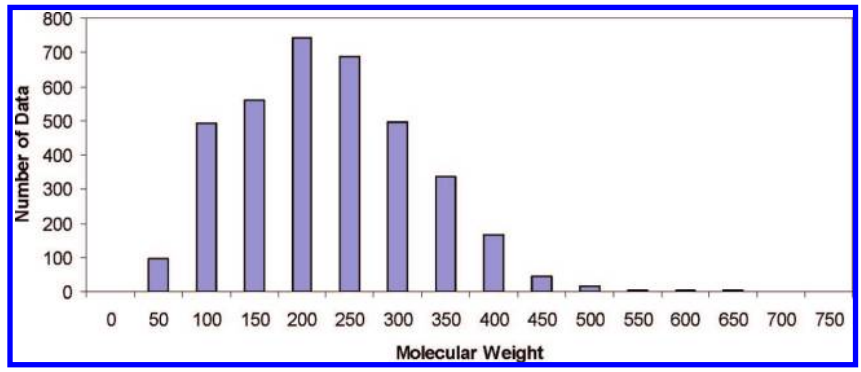

Figure 3. Clustered column representation of the distribution of molecular weight for the 3664-molecule set.

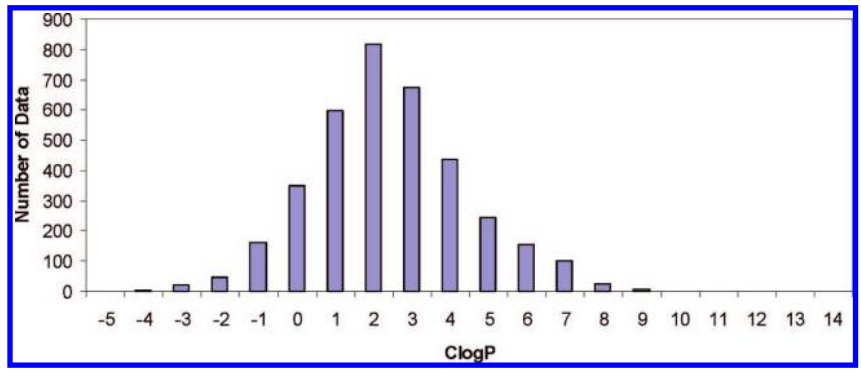

Figure 4. Clustered column representation of the distribution of *ClogP* for the 3664-molecule set.

more details, a probe radius of 0.6 Å, rather than the water probe of 1.4 Å, was used in this work. The SAS of atoms having the same atom type were summed up. It is notable that the same atom type definitions in Table 3 were used both for the ATC and SAS descriptors.

2.2. Molecular Descriptors. A set of molecular physiochemical properties, including *ClogP*, molecular polarizability, molecular weight, intrahydrogen bond number, and hydrophobicity parameters, was utilized to improve the aqueous solubility models.

Among the aforementioned molecular properties, the calculated logP, *ClogP*, has the highest correlation with aqueous solubility, and the correlation coefficient square is 0.625. It is not surprising at all since logP measures how strong the solvent molecule interacts with water in comparison to *n*-octanol. In this work, the logP of each molecule was calculated using the *ClogP* module of ACD/Laboratories. Molecular polarizability, *Pol*, which is a measure of inductive and coefficient dispersion interactions within a molecule or a molecular system, has the second largest correlation with aqueous solubility, and the r^2 was found to be 0.301 for the 3664-molecule set. In this work, an empirical model recently developed by ourselves was utilized to calculate the molecular polarizabilities.[31] Molecular weight is a good measure of molecular size, especially for druglike molecules. The r^2 of aqueous solubility versus molecular weight (*MW*) and aqueous solubility versus square of molecular weight (*MW2*) are 0.313 and 0.272, respectively. However, in combination with ATC and SAS, we found that *MW2*, rather than *MW*, gave better fitting, which is also consistent with the observation reported by Hou and co-workers.[10] Considering all three physicochemical properties, *ClogP*, *Pol*, and *MW* are additive, their contributions to an aqueous solubility model can be partially or even entirely compensated by the ATC or SAS descriptors.

The following molecular properties may not correlate well with the ATC or SAS descriptors. The first such kind of property is the number of intrinsic hydrogen bonds (*IHB*). It is believed that the ability of a molecule to participate in hydrogen bonds with water may be well represented by some ATC or SAS descriptors. However, its ability to form intramolecule hydrogen bonds needs to be quantified explicitly. The algorithm of searching intrinsic hydrogen bonds was described in our previous work.

Hydrophobicity is another factor that makes substantial contribution to aqueous solubility. In our previous work,[5] the numbers of sp^3 carbon (*HC_C3*) and sp^2 carbons (*HC_C2*) in hydrophobic clusters were used to describe hydrophobic cores in a molecule. A hydrophobic core is defined as a collection of sp2 and sp3 carbons linked by covalent bonds, and from any atom in the core there is no heavy atom other than carbon within any seven-atom or shorter paths. Besides *HC_C3* and *HC_C2*, two more hydrophobic parameters, hydrophobic score for sp^3 carbon (*HS_C3*) and sp^2 carbon (*HS_C2*), were introduced. For any sp^3 carbon *i*, hydrophobic chain length *l* loops from 3 to 9, if there is a hydrophobic chain having a chain length equal to *l* and *i* is a member of the chain, *HS_C3* is added by *l*. In other words, the longer the hydrophobic chain, the larger the contribution to the score is. *HS_C2* was calculated in a similar way. Interestingly, the correlations between the *HC_C3*, *HC_C2*, *HS_C3*, and *HS_C2* were found to be very small, indicating that all four parameters can be applied in regression analysis simultaneously.

Although both the intrinsic hydrogen bond and hydrophobicity parameters have much weaker correlations with aqueous solubility (r^2 are smaller than 0.1), these descriptors may not be substituted by the ATC and SAS descriptors. Including these descriptors may improve the performance of aqueous solubility models.

Table 3. Coefficients of Each Descriptor of Two Aqueous Solubility Models Based on Solvent Accessible Surface Areas and Several Molecular Properties

descriptor	description	ASM-ATC-LOGP	ASM-ATC	ASM-SAS-LOGP	ASM-SAS
constant	constant	0.630431	0.850743	0.573996	0.833213
Clogp	calculated logP	−0.341315	-	−0.320238	-
Pol	polarizability	−0.041121	−0.049542	−0.025454	−0.024524
MW2	square of molecular weight	0.000011	0.000013	0.000009	0.000011
IHB	number of intramolecular hydrogen bonds	-	-	−0.075903	−0.079692
HC_C3	number of sp^3 carbon in hydrophobic cores	−0.007790	−0.016665	−0.004233	−0.013631
HC_C2	number of sp^2 carbon in hydrophobic cores	−0.192247	−0.199750	−0.219447	−0.219280
HS_C3	hydrophobic score of sp^3 carbons	−0.010495	−0.010982	−0.011649	−0.012760
HS_C2	hydrophobic score of sp^2 carbons	−0.002622	−0.016382	0.001459	−0.001050
h1	H on aliphatic sp^3 carbon with one electron-withdrawal group	0.268291	0.250592	0.007315	0.006769
h23	H on aliphatic sp^3 carbon with two or three electron-withdrawal groups	0.268422	0.306753	0.006669	0.011226
h4	H on sp^2 carbon with one electron-withdrawal group	0.292334	0.289184	0.010652	0.010095
h5	H on sp^2 carbon with two electron-withdrawal group	0.371634	0.374227	0.018911	0.017672
ha	H on sp^1 and other sp^2 carbons	0.220868	0.140192	0.007622	0.002324
hn	H−N	0.149419	0.241277	0.001076	0.00237
ho	H−O	0.173435	0.287026	0.011549	0.014371
hc	all other H	0.198995	0.122077	0.003092	−0.00176
c	C═O, C═S, or C═N	0.259139	0.290402	−0.01222	−0.01481
c1	sp^1 C	−0.14394	−0.18729	−0.01102	−0.01663
ca	aromatic C attached to one hydrogen	0.13427	0.187427	0.00445	−0.00124
ca2	aromatic C without hydrogen	−0.04509	−0.01793	−0.02751	−0.04144
ca3	aromatic C with three other aromatic atoms, such as central atom of 1H-phenalene	−0.02884	−0.04576	−0.02113	−0.0378
c2	all other sp^2 carbon	0.104457	0.178686	0.002785	−0.00139
c3	all other sp^3 carbon	−0.04434	0.091534	0.020931	0.016362
n	N in amide group	−0.06712	0.125751	0.01084	0.017599
n1	nitrogens in cyano, N═N═R or N≡N-R	0.348715	0.576025	0.006789	0.010905
n2	other two-substituent sp^2 nitrogen	0.015109	0.084623	−0.01177	−0.00959
n3	all other nitrogen	0.602241	0.647543	0.182327	0.190327
n3_2	N in amine group with two H	0.793642	0.849957	0.120654	0.127373
n3_1	N in amine group with one H	0.010193	−0.01691	0.015576	0.018045
n_1	amide N in imides	−0.15778	−0.1473	−0.00236	−0.06236
n_2	sp^3 N in guanidines	−0.22828	−0.1718	−0.01561	−0.01685
na	sp^2 N in planar ring with three substituents	−0.15701	−0.15724	−0.04148	−0.03703
nb	aromatic nitrogen, two substituents	0.14704	0.271272	0.009332	0.020504
nh	other sp^2 N with three substituents	0.061282	0.081893	0.027903	0.038799
no	N in nitro group	0.00746	0.001611	0.054472	0.05089
o20	sp^2 O bonded to a planar or aromatic sp^2 carbon, such as O in 3H-pyrrol-3-one	−0.35104	−0.21787	−0.01321	−0.00475
o21	O in aldehydates	−0.04708	0.02669	0.012728	0.013106
o22	O in ketones	−0.08304	0.131527	0.004413	0.011793
o23	sp^2 O in carboxyl group, O═C-SH, COO- or COS-	0.21874	0.141418	0.009432	0.002861
o24	sp^2 O in amide group	0.09392	0.232048	0.009428	0.019948
o26	sp^2 O in ester	−0.25288	−0.05187	−0.00548	−0.00211
o2n	O in nitro group	0.014921	0.003221	−0.01008	−0.01357
o2p	O═P	0.977503	1.359498	0.024012	0.049886
o2s	O═S	0.568359	0.385238	0.008361	0.007518
oh	all other hydroxyl O	0.153793	0.14722	0.011746	0.01489
oh′	hydroxyl O in HO-C═O, HO-C═S, HO-C═NR, or HO-C═PR	0.019643	0.139806	0.013298	0.021975
os	all other sp^3 O	−0.12238	−0.19444	−0.02262	−0.02776
os′	sp^3 O in RO-C═O, RO-C═S, RO-C═NR, or RO-C═PR	0.002121	−0.16112	−0.00497	−0.01536
osp	sp^3 O in RO-S═O, RO-S═S, RO-P═O or RO-P═S	0.246556	0.127534	−0.03613	−0.04574
p	any P	−1.07776	−0.80764	0.023468	0.001787
s	sp^2 sulfur in S═P, S═C, etc.	0.219865	0.262606	−0.00373	−0.00631
s4	hypervalent sulfur, four to six substituents	−0.69599	0.063326	−0.0138	0.002359
sh	sp^3 sulfur in thiol groups	−0.54912	−0.39044	−0.01826	−0.02033
ss	sp^3 sulfur in -SR or S−S	0.194372	0.035753	−0.00398	−0.01419
f	any F	−0.11499	−0.28472	−0.00708	−0.01285
cl	any Cl	0.022437	−0.18123	−0.00678	−0.01463
br	any Br	−0.05124	−0.34974	−0.00709	−0.01563
i	any I	0.121981	−0.24096	−0.00624	−0.01604

3. Model Construction and Validation. Regression analysis was performed to construct a set of aqueous solubility prediction models with a different combination of descriptors. The performance of an aqueous solubility model is described by the following parameters: n - number of data points, m - number of descriptors, *AUE* -average unsigned error in log unit, *RMSE* - root-mean-square error in log unit, r^2 - square of correlation coefficient, and q^2 - square of correlation coefficient for the test set.

Each aqueous solubility model developed in this work was extensively validated by three types of experiments. First of all, a leave-one-out analysis was conducted. Second, a 85/15 (15% randomly selected data entered the test set and the remaining data in the training set) cross-validation test was

Table 4. Performance of Full Regression Analysis of Four Solubility Models for the Whole Data Set (3664 Molecules)

model	M^a	*AUE*	*RMSE*	r^2	*F*
ASM-ATC-LOGP	57	0.624	0.840	0.832	18279.5
ASM-ATC	56	0.654	0.866	0.822	16975.4
ASM-SAS-LOGP	58	0.631	0.850	0.828	17770.9
ASM-SAS	57	0.656	0.872	0.819	16692.3

[a] Number of descriptors.

performed for 10,000 times. For each run, the aqueous solubilities of the test molecules were predicted with the models based on the training set. In the third, the scrambling test, which continually swapped the experimental values for two randomly selected molecules until none molecule had its original experimental data, was performed for 1000 times. If the r^2 of the scrambled data sets are close to 0.0, the QSPR models developed using the original data set should be valid and not be obtained by chance.

RESULTS AND DISCUSSION

1. Performance of Four Aqueous Solubility Models. Four aqueous solubility models have been developed by using a different combination of descriptors for a large data set of 3664 molecules. Models ASM-ATC-LOGP and ASM-ATC mainly utilized atom type counts as descriptors, while models ASM-SAS-LOGP and ASM-SAS used atom type classified solvent accessible surface areas as the major descriptors. As indicated by their names, ASM-ATC-LOGP and ASM-SAS-LOGP employed *ClogP* as a descriptor, while ASM-ATC and ASM-SAS did not. Considering a 3D-structure is required to calculate the number of intrinsic hydrogen bonds, *IHB* did not show up in both ASM-ATC-LOGP and ASM-ATC models, since 3D-structures were not required for both models. Other molecular descriptors, including *Pol*, *MW2*, *HC_C3*, *HC_C2*, *HS_C3*, and *HS_C2*, entered all four models.

The performance of the four models is very encouraging. The regression coefficient squares r^2 are 0.839, 0.828, 0.835, and 0.826 for ASM-ATC-LOGP, ASM-ATC, ASM-SAS-LOGP, and ASM-SAS, respectively. The root-mean-square errors are also very close, which are 0.823, 0.850, 0.832, and 0.855 for the four models. Surprisingly, the SAS descriptor does not show a better performance than ATC. We think that the SAS descriptor is superior to ATC since the former can differentiate the contributions of the exposed and interior atoms to the solute−solute and solute−water interactions, while ATC totally neglects the difference of the two kinds of atoms. Unfortunately, the advantage of the SAS descriptors does not lead to better models than the ATC descriptors. Considering models ASM-ATC-LOGP and ASM-ATC do not need a 3D-structure to calculate the solvent accessible surface areas, they are more suitable to be applied in virtual high throughput screenings. The fact that the performance of ASM-ATC-LOGP and ASM-SAS-LOGP is only slightly better than that of corresponding ATC-

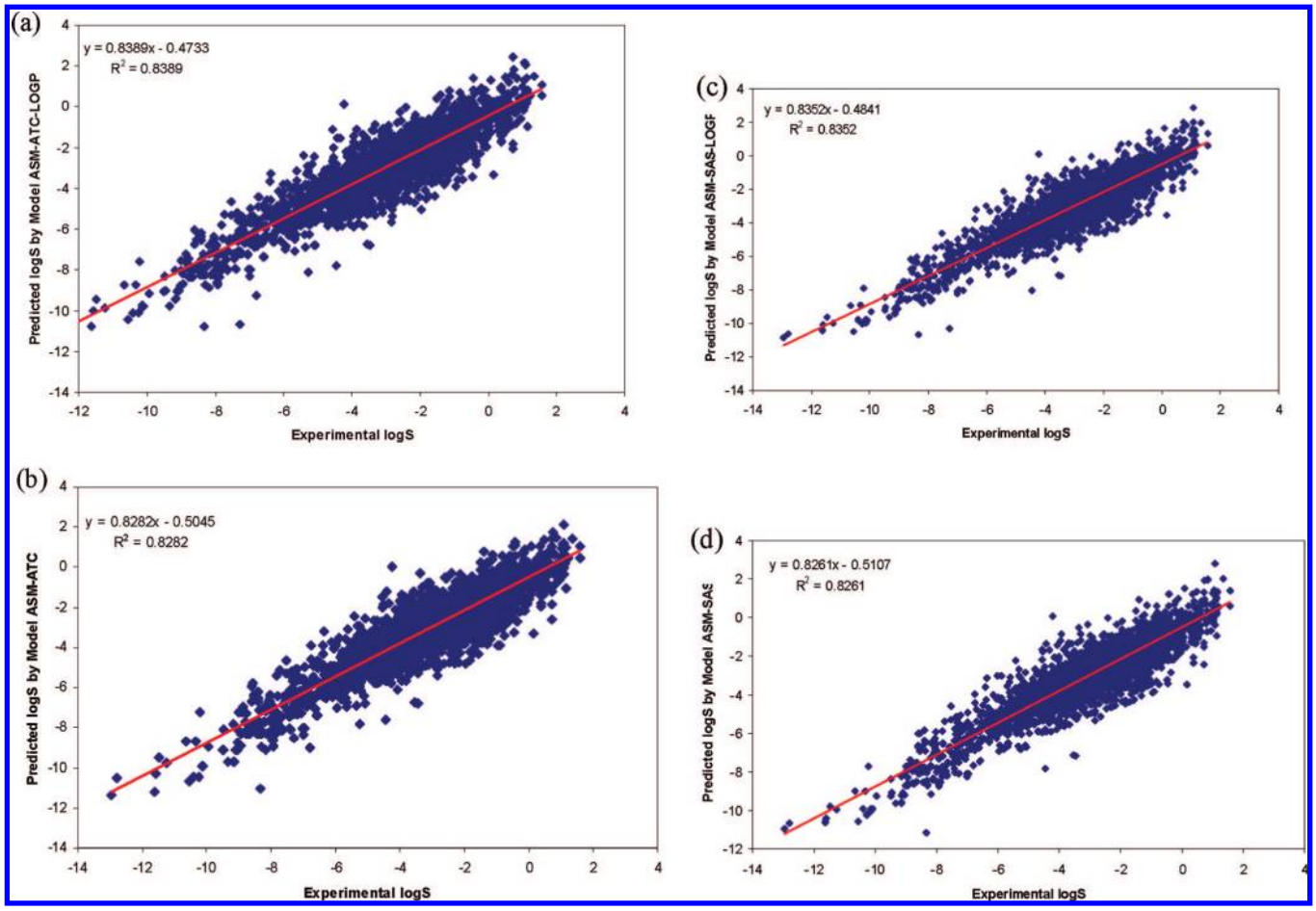

Figure 5. Plots of experimental versus calculated logS using four models for 3646 molecules: (a) ASM-ATC-LOGP, (b) ASMS-ATC, (c) ASM-SAS-LOGP, and (d) ASM-SAS.

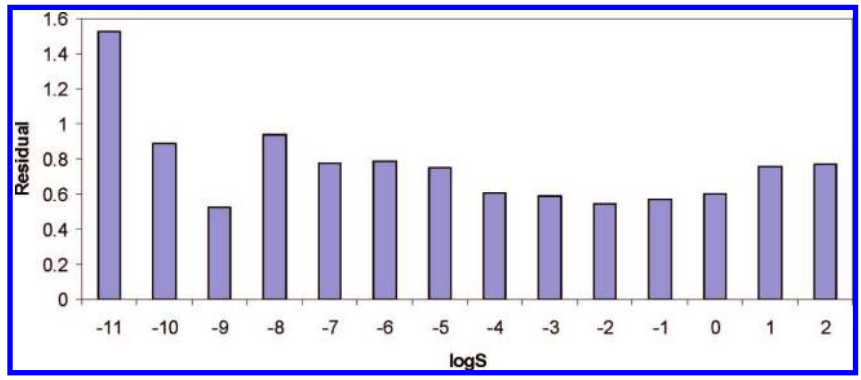

Figure 6. The distribution of the prediction errors by model ASM-ATC-LOGP.

Table 5. Performance of logS Prediction Using ASM-ATC-LOGP and ACD/Labs Solubility Module (Version 11) for Five Data Sets

		ASM-ATC-LOGP			ACD/Laboratories		
	N^a	*AUE*	*RMSE*	r^2	*AUE*	*RMSE*	r^2
Data Set 1	1708	0.545	0.717	0.881	–	–	–
Data Set 2	578	0.507	0.722	0.943	–	–	–
Data Set 3	1210	0.745	0.978	0.565	0.837	1.164	0.516
Data Set 4	90	0.672	0.909	0.527	0.746	0.940	0.494
Data Set 5	119	0.662	0.851	0.627	0.884	1.178	0.434

[a] Number of molecules in a data set.

Table 6. Performance of Leave-One-out Regression Analysis of Four Solubility Models for the Whole Data Set (3664 Molecules)

model	M^a	*AUE*	*RMSE*	q^2	*F*
ASM-ATC-LOGP	57	0.613	0.823	0.839	18969.4
ASM-ATC	56	0.642	0.850	0.828	17571.4
ASM-SAS-LOGP	58	0.619	0.832	0.835	18466.9
ASM-SAS	57	0.644	0.855	0.826	17315.6

[a] Number of descriptors.

Table 7. Minimum, the Maximum, the Mean, and Root-Mean-Square Deviation of Three Statistical Parameters of the 10,000 Times 15/85 Cross-Validation Analysis[a]

	minimum	maximum	mean	rmsd
		ASM-ATC-LOGP		
AUE	0.541	0.723	0.625	0.022
RMSE	0.732	0.959	0.841	0.031
q^2	0.762	0.884	0.832	0.016
		ASM-ATC		
AUE	0.574	0.739	0.655	0.022
RMSE	0.748	0.979	0.867	0.030
q^2	0.754	0.882	0.821	0.016
		ASM-SAS-LOGP		
AUE	0.550	0.714	0.632	0.022
RMSE	0.732	0.963	0.850	0.032
q^2	0.760	0.882	0.827	0.017
		ASM-SAS		
AUE	0.565	0.745	0.657	0.022
RMSE	0.737	0.993	0.873	0.031
q^2	0.734	0.875	0.818	0.017

[a] Each time 3100 molecules were randomly selected in the training set and the others in the test set. The statistical parameters, *AUE*, *RMSE*, and q^2 are the mean errors, the root-mean-square errors, and the correlation coefficient squares, respectively.

LOGP and SAS-LOGP models indicates *ClogP* is implicitly taken into account in the later two models. The advantage of ASM-ATC and ASM-SAS over ASM-ATC-LOGP and ASM-SAS-LOGP is that the former two models are not affected by the potential error introduced by the logP calculations and can be applied directly to make a prediction with our free software package.

The major parameters of regression analysis for the four models are listed in Table 4. The plots of experimental versus predicted aqueous solubilities are shown in Figure 5. The distribution of the prediction errors for different ranges of experimental aqueous solubility is shown in Figure 6. It is clear that the prediction errors are smaller for logS from −5.0 to 0.0. Too insoluble and too soluble molecules have larger prediction errors, for which the uncertainty of the experimental values are also larger.

How well does an aqueous solubility model perform for individual data sets? Table 5 lists the *AUE*, *rms*, and r^2 of the five data sets for model ASM-ATC-LOGP. Data Sets 1 and 2, which are widely used in other models, have a significantly better performance than the other three data sets. The *AUE*, *RMSE*, and r^2 of Data Set 1 are 0.545, 0.717, and 0.881 logarithm units, respectively. This performance is very close to that of ASMS-LOGP, the best model developed in our previous work only using Data Set 1 ($AUE = 0.505$, $RMSE = 0.664$, $r^2 = 0.897$). As to Data Set 3, the *AUE*, *RMSE*, and r^2 are 0.745, 0.978, and 0.565, respectively. Although this performance is not as good as that of Data Sets 1 and 2, it is much better than that of ASMS-LOGP ($AUE = 0.937$, $RMSE = 1.281$) and ACD/Laboratories ($AUE = 0.837$, $RMSE = 1.164$). It must be pointed out that this comparison may not fair for ASMS-LOGP and ACD/Laboratories, since Data Set 3 was not included or only partially included in the training sets of ASMS-LOGP and ACD/Laboratories. In summary, ASM-ATC-*CLOGP* is a very robust model, and it achieves an encouraging prediction performance for all five data sets, while ASMS-LOGP and ACD/Laboratories do not give a satisfactory prediction for Data Set 3.

2. The Predictability of the Four Aqueous Solubility Models. A QSPR model is useful only if it has a good predictability. All four models, ASM-ATC-LOGP, ASM-ATC, ASM-SAS-LOGP, and ASM-SAS, were extensively validated in three experiments. First of all, the leave-one-out analysis was performed, and the results are listed in Table 6. Clearly, q^2, the leave-one-out regression coefficient squares, are very close to r^2 of a full component regression analysis: 0.832−0.839 for ASM-ATC-LOGP, 0.822−0.828 for ASM-ATC, 0.828−0.835 for ASM-SAS-LOGP, and 0.819−0.826 for ASM-SAS. Similarly, the *RMSE* values are also very close.

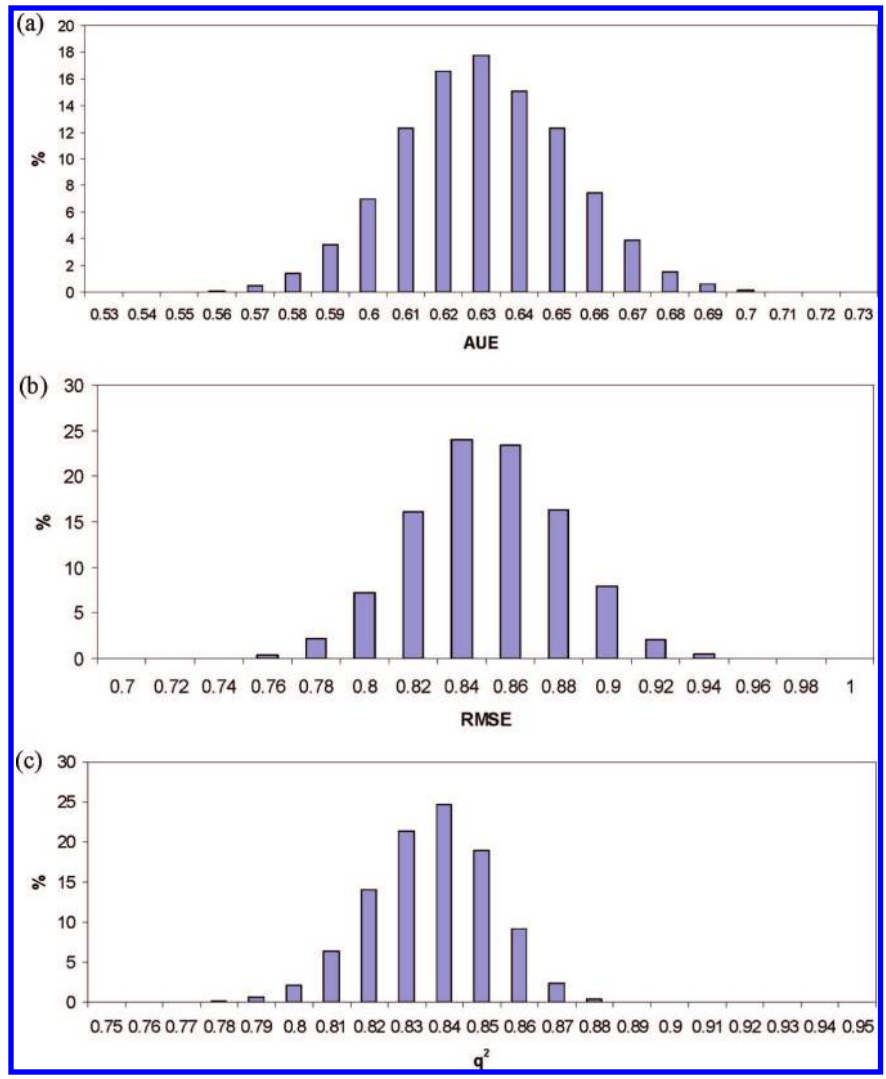

Figure 7. Clustered column charts of the average unsigned errors (*AUE*), root-mean-square errors (*RMSE*), and regression coefficient squares (q^2) of the 10,000 times 85/15 cross-validations for model ASM-ATC-LOGP: (a) *AUE*, (b) *RMSE*, and (c) q^2.

Table 8. Statistics on r^2 of 1000 Times Scrambling Tests

model	minimum	maximum	mean	rmsd
ASM-ATC-LOGP	0.014	0.046	0.025	0.004
ASM-ATC	0.011	0.040	0.024	0.004
ASM-SAS-LOGP	0.014	0.044	0.026	0.004
ASM-SAS	0.013	0.040	0.025	0.004

In the second experiment, a 15/85 cross-validation analysis was performed for 10,000 times for each model. In each cross-validation run, 362 molecules were randomly selected to enter the test set, and the other molecules entered the training set. The model constructed using the training set was applied to predict the solubilities for the 362 molecules in the test set. The average of the *AUE*, *RMSE*, and q^2 for 10,000 runs are summarized in Table 7. The distribution of *AUE*, *RMSE*, and q^2 represented by clustered column charts are shown in Figure 7. Interestingly, the average q^2 of 10,000 times 15/85 cross-validation are very close to those of leave-one-out analysis for all four models: 0.832−0.832 for ASM-ATC-LOGP, 0.821−0.822 for ASM-ATC, 0.827−0.828 for ASM-SAS-LOGP, and 0.818−0.819 for ASM-SAS. Similarly, the *RMSE* of both tests are also very close. This result indicates that the leave-one-out analysis is a very robust cross-validation approach. On the other hand, dividing a whole data set into the training and test sets randomly and then making a prediction for the test set for only once or several times may lead to a misleading conclusion. Even for a reliable model such as ASM-ATC-LOGP developed using a large data set, the *AUE*, *RMSE*, and q^2 still have broad distributions in 10,000 times 15/85 cross-validation experiment as shown in Figure 7.

In the third experiment, a scrambling test was conducted 1000 times for each model. The minimum, the maximum, and the average r^2 of each model were listed in Table 8. The mean r^2 of four models are all smaller than 0.026, indicating the performance of the aqueous solubility models is not achieved by chance.

A disadvantage of some atom type-based models lies in the fact that they cannot differentiate the solubility of isomers. An interesting example is anthracene carboxylic acids which have different aqueous solubilities when the carboxylic functional group are substituted at the three different sites (anthracene-9-carboxylic acid > anthracene-1-carboxylic acid > anthracene-2-carboxylic acid). We made a prediction for the three molecules with all four models developed in this study. For ASM-ATC-LOGP, the predicted logS are −4.37, −4.45, and −4.28 logarithm units for anthracene-1-carboxylic acid, anthracene-2-carboxylic acid, and anthracene-9-

Table 9. Performance of Predicting Aqueous Solubility for the 32 Test Set Compounds in the "Solubility Challenge" Campaign by Models ASM-ATC-LOGP and ACD/Labs

	experimental		predicted logS	
compound	solubility[a] (μg/mL)	logS[b]	ASM-ATC-LOGP	ACD/Labs
acebutolol	711	−2.675	−2.6362	−2.45
amoxicillin	3900	−1.9717	−3.0899	−2.16
bendroflumethiazide	21.2	−4.2984	−3.9542	−4.1
benzocaine	780	−2.3259	−2.6144	−2.04
benzthiazide	6.4	−4.8292	−4.893	−5.24
2-chloromandelic acid	TStM	TStM	−2.4423	−1.4
clozapine	188.9	−3.2381	−3.8106	−4.11
dibucaine	14	−4.3898	−3.2758	−3.35
diethylstilbestrol	10	−4.4287	−5.0145	−4.88
diflunisal	0.29	−5.9359	−4.3514	−4.23
dipyridamole	3.46	−5.1639	−1.5267	−2.5
1R-2S-ephedrine	TStM	TStM	−1.2416	−0.64
folic acid	2.5	−5.2469	−2.6132	−2.66
furosemide	19.6	−4.2272	−4.0347	−4.35
hydrochlorothiazide	625	−2.678	−2.422	−2.23
imipramine	22	−4.1054	−3.5886	−4.61
indomethacin	410	−2.9408	−4.4305	−4.1
ketoprofen	157	−3.2094	−3.6183	−3.37
lidocaine	3130	−1.8743	−2.5102	−2.19
marbofloxacin	TStM	TStM	−2.6827	−2.88
meclofenamic acid	0.16	−6.2674	−5.6626	−5.67
naphthoic acid	28.96	−3.7742	−3.226	−3.07
1R-2R-pseudoephedrine	TStM	TStM	−2.5569	−3.52
probenecid	3.9	−4.627	−1.1974	~
pyrimethamine	19.4	−4.1079	−4.3137	−3.63
salicylic_acid	1620	−1.9307	−2.0779	−1.35
sulfamerazine	200	−3.1211	−2.7729	−2.63
sulfamethizole	450	−2.7787	−3.1731	−2.7
terfenadine	0.01	−7.7412	−5.1415	−5.63
thiabendazole	66	−3.4842	−3.656	−3.53
tolbutamide	93	−3.4634	−3.528	−2.85
trazodone	127	−3.4666	−3.5931	−3.76

[a] Experimental solubility data were adopted from the result page of the "Solubility Challenge" campaign. TStM stands for "too soluble to measure". [b] logS were calculated using the solubility data in the second column of this table.

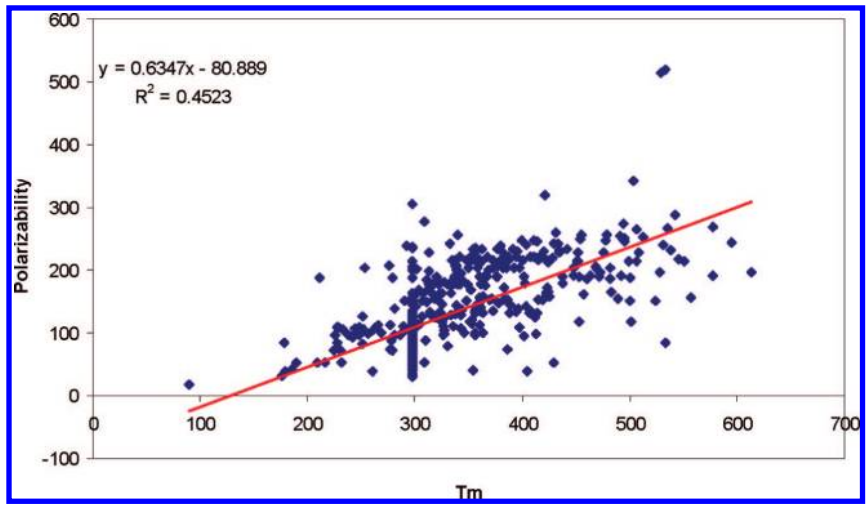

Figure 8. Correlation between the temperature of melting point (T_m) and molecular polarizability.

carboxylic acid, respectively. Similarly, the three other models, ASM-ATC (−4.25, −4.34, −4.15), ASM-SAS-LOGP (−4.24, −4.48, −4.02), and ASM-SAS (−4.02, −4.28, −3.78), also successfully predict the trend. Owing to the introduction of HC_C2 and HS_C2 descriptors, although the three molecules have the same atom type numbers and the same *ClogP* and polarizability values, both atom type-based models ASM-ATC and ASM-ATC-LOGP can successfully reproduce the experimentally determined trend. As to ASM-SAS and ASM-SAS-LOGP, not only the HC_C2 and HS_C2 parameters are different but also the atom type classified solvent-accessible surface areas are different for the three anthracene carboxylic acids.

During the time the manuscript was being reviewed, the "solubility challenge" campaign sponsored by Goodman et al.[6] released the results of predicting solubility for a set of 32 molecules on the home page of the *Journal of Chemical Information and Modeling*. One of the models developed in

this work, ASM-ATC-LOGP, has achieved a very encouraging performance: it was ranked as one of the top 10 models among a total of 99 models developed with all kinds of means. This performance is based upon the percents of correct prediction (with in ±0.5 logarithm unit of logS) for the Known28 Set and the Best24 Set. We listed the prediction results of both ASM-ATC-LOGP and the ACD/Laboratories models in Table 9 as suggested by one reviewer. Without considering the four too soluble to measure compounds and probenecid, for which ACD/Laboratories failed to make a prediction, the average unsigned errors of logS of the remaining 27 compounds are 0.766 and 0.716 for ASM-ATC-LOGP and ACD/Laboratories, respectively. The correlation coefficient between the predicted values by two models is 0.91, indicating the error sources of prediction are quite similar for the two models. It is pointed out that although such kind of a campaign is useful to evaluate how well aqueous solubility can be predicted with the current available models and software packages, it cannot be used to rank a specific model unbiasedly since the test set of 32 molecules is too small. Even with a much larger test set, as we did in the 10,000 times 15/85 cross-validation analysis for model ASM-ATC-LOGP, the unsigned average errors of the 362-test set molecules which were randomly selected for each run are ranged from 0.541 to 0.723 logarithm units (Table 7). Another missing part of the campaign is the evaluation of efficiency of the solubility models. Too expensive models, such as those based on molecular mechanical and quantum mechanical calculations and those using experimental determined descriptors, may not have a good applicability in drug design.

3. The Application of Aqueous Solubility. The applicability of the four solubility models is well characterized by Figure 2. The 3664 molecules are evenly distributed in a chemical space defined by molecular weight, which describes the size of a molecule, and *ClogP*, which describes the polarity of a molecule. The distribution of *ClogP* and molecular weights of 3664 molecules are shown in Figures 3 and 4. Evidently, most molecules fall into the ranges of druglike molecules for both molecular weight and *ClogP* according to Lipinski's "Rule of five".

In drug design, our solubility models, especially ASM-ATC-LOGP, can serve as a general filter to prioritize a compound library prior to doing virtual or real high throughput screenings. They can also serve as a new rule, in addition to the "Rule of five", to evaluate how a molecule is druglike, since drug molecules are typically more soluble than the screening molecules according to our previous findings.[5]

4. Simple Models To Estimate Aqueous Solubility. Although the four developed aqueous solubility models can be applied to make a reliable prediction for a given external molecule, they are too complicated to be used by organic chemists or medicinal chemists to fast estimate the solubility of a molecule or how solubility changes when altering the original compound. In this work, we redeveloped a simple model, ASM-TM, using *ClogP* and experimental temperature of melting points, T_m, as descriptors. The data set, Data Set 2, is the clean version of that used by Jain and Yalkowsky.[6] The regression equation is shown in eq 2.

$$\log S = 3.513 - 0.010 \times T_m - 1.112 \times ClogP \qquad (2)$$

The standard error and r^2 are 0.720 and 0.937, respectively. According to eq 2, if T_m increases 100 degrees or *ClogP* increases one logarithm unit, logS roughly drops 1.0 logarithm unit.

Since T_m is not always available, the above equation has a limited application. We tried to replace T_m with molecular polarizability, *Pol*, which can be easily calculated using our empirical models. The correlation coefficient square of T_m to *Pol* was 0.452 as shown in Figure 8. Encouragingly, the model constructed using *ClogP* and *Pol* (ASM-POL) is only marginally worse, and the standard error and r^2 are 0.887 and 0.905, respectively. The regression equation is shown in eq 3.

$$\log S = 1.095 - 0.008 \times Pol - 1.078 \times ClogP \qquad (3)$$

Again, logS drops one logarithm unit if *ClogP* increases one or *Pol* increases 125. It should be noted that the above empirical rules are very rough, and large errors may occur for some molecules.

CONCLUSIONS

In this work, we have successfully developed four aqueous solubility models mainly using atom type counts and atom type classified solvent accessible surface areas as descriptors. All four models have been extensively validated by three cross-validation experiments. Developed from a large data set of 3664 molecules, the four models are very robust and can make reliable predictions for external test molecules. In fact, model ASM-ATC-LOGP, the best among the four models, has achieved an encouraging performance in predicting the solubility for all five data sets, while our previous models, ASMS-LOGP and ACD/Laboratories, do not give a satisfactory prediction for the Belistein data set. We believe that our reliable aqueous solubility models will have a great use in drug discovery.

Abbreviations. QSPR - quantitative structure−property relationship; ATC - atom type count; SAS - solvent accessible surface area; ADMET - absorption, distribution, metabolism, excretion, and toxicity; HTS - high throughput screening; *ClogP* - calculated logP*; Pol* - molecular polarizability; T_m - temperature of melting point; *AUE* - average unsigned error; *RMSE* - root-mean-square error; r^2 - square of correlation coefficient for the training set; q^2 - square of correlation coefficient for the test set.

ACKNOWLEDGMENT

We are grateful to acknowledge the research support from NCSA (MCB000013N (J.W., P.I.)).

Supporting Information Available: Compound names, SLNs, and experimental and predicted values of the five data sets (Tables S1−S5). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. Recent advances in computational prediction of drug absorption and permeability in drug discovery. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.
(2) Hou, T. J.; Wang, J. M. Structure - ADME relationship: still a long way to go. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 759–771.

(3) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(4) 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.
(5) Wang, J.; Krudy, G.; Hou, T.; Holland, G.; Xu, X. Development of reliable aqueous solubility models and their application in drug-like analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395–1404.
(6) Llinàs, A.; Glen, R. C.; Goodman, J. M. Solubility challnge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements. *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.
(7) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355–366.
(8) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234–252.
(9) 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.
(10) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266–275.
(11) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489–496.
(12) 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.
(13) Tetko, I. V.; Tanchuk, Y. V.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemcial compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488–1493.
(14) 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.
(15) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
(16) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
(17) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodiversity* **2004**, *1*, 1829–1841.
(18) Toronto, Ontario, Canada, 2008. http://acdlabs.com (accessed Oct 2008).
(19) *Sybyl user manual*; Tripos Inc.: St. Louis, MO, U.S.A., 1995.
(20) Mithani, S. D.; Bakatselou, V.; TenHoor, C. N.; Dressman, J. B. Estimation of the increase in solubility of drugs as a function of Bile salt concentration. *Pharm. Res.* **1996**, *13*, 163–175.
(21) Nielsen, N. M.; Bungaard, H. Glycolamide esters as biolabile prodrugs of carboxylic acid agents: synthesis, stability, bioconversion, and physicochemical properties. *J. Pharm. Sci.* **1988**, *77*, 285–298.
(22) Stella, V. J.; Martodihardjo, S.; Terada, K.; Rao, V. M. Some relationships between the physical properties of various 3-acyloxymethyl prodrugs of phenytoin to structure: Potential in vivo performance implications. *J. Pharm. Sci.* **1998**, *87*, 1235–1241.
(23) Lapanje, S.; GrEar, B.; Trtnik, G. Thermodynamic studies of the interactions of guanidinium chloride with some oligopeptides containing l-valine, l-leucine, l-tryptophan, and l-tyrosine. *J. Chem. Eng. Data* **1980**, *25*, 320–323.
(24) Casini, A.; Scozzafava, A.; Mincione, F.; Menabuoni, L.; Ilies, M. A.; Supuran, C. T. Carbonic Anhydrase Inhibitors: Water-Soluble 4-Sulfamoylphenylthioureas as Topical Intraocular Pressure-Lowering Agents with Long-Lasting Effects. *J. Med. Chem.* **2000**, *43*, 4884–4892.
(25) Paulsson, M.; Edsman, K. Controlled drug release from gels using surfactant aggregates: I. Effect of lipophilic interactions for a series of uncharged substances. *J. Pharm. Sci.* **2001**, *90*, 1216–1225.
(26) Aldini, R.; Roda, A.; Montagnani, M.; Cerre, C.; Pellicciari, R.; Roda, E. Relationship between structure and intestinal absorption of bile acids with a steroid or side-chain modification. *Steroids* **1996**, *61*, 590–597.
(27) Thomas, E.; Rubino, J. Solubility, melting point and salting-out relationships in a group of secondary amine hydrochlorides. *Int. J. Pharm. Sci.* **1996**, *130*, 179–183.
(28) Schoenmakers, R. G.; Stehouwer, M. C.; Tukker, J. J. Structure-Transport relationship for the intestinal small-peptide carrier: Is the carbonyl group of the peptide bond relevant for transport. *Pharm. Sci.* **1999**, *16*, 62–68.
(29) Hartman, G. D.; Halczenko, W.; Smith, R. L.; Sugrue, M. F.; Mallorga, P. J.; Michelson, S. R.; Randall, W. C.; Schwam, H.; Sondey, J. M. 4-substituted thiophene- and furan-2-sulfonamides as topical carbonic anhydrase inhibitors? *J. Med. Chem.* **1992**, *35*, 3822–3831.
(30) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.
(31) Wang, J.; Xie, X.-Q.; Hou, T. J.; Xu, X. J. Fast approaches for molecular polarizability calculations. *J. Phys. Chem. A* **2007**, *111*, 4443–4448.

CI800406Y