Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis

Junmei Wang,*,† George Krudy,† Tingjun Hou,‡ Wei Zhang,§ George Holland,† and Xiaojie Xu*,§

Encysive Pharmaceuticals Inc., 7000 Fannin Street, Houston, Texas 77030, 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 March 14, 2007

In this work, two reliable aqueous solubility models, ASMS (aqueous solubility based on molecular surface) and ASMS-LOGP (aqueous solubility based on molecular surface using ClogP as a descriptor), were constructed by using atom type classified solvent accessible surface areas and several molecular descriptors for a diverse data set of 1708 molecules. For ASMS (without using ClogP as a descriptor), the leave-one-out q^2 and root-mean-square error (RMSE) were 0.872 and 0.748 log unit, respectively. ASMS-LOGP was slightly better than ASMS ($q^2 = 0.886$, RMSE = 0.705). Both models were extensively validated by three cross-validation tests and encouraging predictability was achieved. High throughput aqueous solubility prediction was conducted for a number of data sets extracted from several widely used databases. We found that real drugs are about 20-fold more soluble than the so-called druglike molecules in the ZINC database, which have no violation of Lipinski's "Rule of 5" at all. Specifically, oral drugs are about 16-fold more soluble, while injection drugs are 50−60-fold more soluble. If the criterion of a molecule to be soluble is set to −5 log unit, about 85% of real drugs are predicted as soluble; in contrast only 50% of druglike molecules in ZINC are soluble. We concluded that the two models could be served as a rule in druglike analysis and an efficient filter in prioritizing compound libraries prior to high throughput screenings (HTS).

INTRODUCTION

According to the Tufts Center for the Study of Drug Development (CSDD), the cost of bringing a drug to market has risen from an average of $231 million in 1991 to roughly $900 million in 2003.[1] Usually, the later the development phase, the more costly a drug candidate is. For example, in a study conducted by the Tufts CSDD, the out-of-pocket cost in the clinical period phase for 27 approved drugs are 15.2, 41.7, and 115.2 million for phase I, phase II, and phase III, respectively. For the traditional drug discovery philosophy, one first identifies an active inhibitor as a lead and then optimizes its selectivity and other physicochemical, physiological, and pharmacokinetic properties (such as absorption, distribution, metabolism, excretion, and toxicity, in short ADMET) sequentially. It is gradually realized that this serial protocol is inferior to a parallel protocol for which the drug lead's activity and selectivity as well as ADMET/pharmacokinetic properties are optimized simultaneously, simply because the parallel protocol can effectively eliminate bad drug candidates from the earlier stages and then substantially reduce the development costs. What makes a compound a good drug candidate? A good drug candidate should be potent against the drug target to induce some biological response, highly selective, has good ADMET/pharmacokinetic properties, and so on. Druglike analysis is useful to discriminate good drug candidates from the screening compounds. The most famous druglike filter is the "Rule of 5" suggested by Lipinski,[2] which states that a good drug candidate should have molecular weight smaller than 500, the calculated log*P* (ClogP) smaller than 5.0, and the numbers of hydrogen bond donors and acceptors less than 5 and 10, respectively. Although most drugs obey the four rules of the filter, "Rule of 5" is only the minimum criterion of a molecule to be druglike. It is very easy for a compound to fall within the "Rule of five" but have no potential to become a drug. As a matter of fact, there are about 2 million vendor compounds in the ZINC database having no violation of "Rule of 5" at all.[3] Evidently, not all of the 2 million compounds have the potential to become drugs. Thus, more stringent criteria should be built up to efficiently enrich druglike compounds from the others. Reliable in silico models that predict ADMET/pharmacokinetic properties (aqueous solubility, membrane permeability, intestinal absorption, metabolism, toxicity, oral bioavailability, plasma protein binding, urinary excretion, area under the plasma concentration-time curve (AUC), total body clearance (Cl), volume of distribution, elimination half time ($t_{1/2}$), etc.) can be applied for this purpose.

Adequate aqueous solubility is important for a drug to be administrated orally or by injection. Aqueous solubility and membrane permeability are the two key factors that affect a drug's oral bioavailability. 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.

* Corresponding author e-mails: jwang@encysive.com and xiaojxu@pku.edu.cn.
† Encysive Pharmaceuticals Inc.
‡ University of California at San Diego.
§ Peking University.

10.1021/ci700096r CCC: $37.00
Published on Web 06/15/2007

In concept, aqueous solubility S of a nonelectrolyte is the concentration (mol/L) of its saturated aqueous solution. Usually, the logarithm of solubility, log$S$, is used for convenience. 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 descriptors to build up models and predict this property.

In the following, we will give a brief review of solubility predication. A more detailed summary on solubility prediction was presented by Lipinski et al.[2] and Jorgensen et al.[4] Those methods can be categorized into two types: those that correlate with experimentally determined properties and those that do not. The performance of a QSPR model is described by a set of 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.

The first type of models was exemplified by Jain and Yalkowsky's recent work in which log$S$ was correlated with an experimentally determined melting point (MP) and the logarithm of octanol/water partition coefficient (log$P$): n = 580 and AUE = 0.45.[5] The first type of models has little use in high throughput prediction of aqueous solubility since experimental melting points are typically not available. Quite a few second types of models are calculated by eq 1 in accordance with the additive characteristics of aqueous solubility, where c_i is the number of occurrence of a molecular fragment i or an atom type i, and w_i is its weight, which is determined by regression analysis, and c_0 is a constant. Certainly, other descriptors, such as molecular weight, molecular polarizability, and calculated logP, can be incorporated into eq 1.

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

Klopman and Zhu reported a set of models with the counts of fragments as descriptors and the best model utilized 171 fragments: n = 1168, m = 171, r^2 = 0.95, AUE = 0.49.[6] In a more recent report by Hou et al. the counts of atom types in addition to two correction factors (hydrophobic carbon and square of molecular weight) were used to build up a model for 1290 organic molecules that covered a large variety of chemical classes (n = 1290, m = 78, r^2 = 0.92, AUE = 0.48, RMSE = 0.61 and n_{test} = 120, AUE_{test} = 0.57, $RMSE_{test}$ = 0.79).[7] The fragment/atom type contribution method does not need any descriptors based on other theoretical models, and they only need to count the occurrence of functional groups or atom types in a molecule, so they are extremely time-saving. One potential disadvantage of this kind of method is that new fragments or atom types not defined in the training sets may cause substantial errors.

Besides the counts of fragments or atom types, a lot of theoretical descriptors have been successfully applied in predicting aqueous solubility. Mitchell and Jurs utilized topological, geometric, and electronic descriptors to predict aqueous solubility for a set of diverse molecules (n = 295, m = 9, r^2 = 0.93, RMSE = 0.64).[8] Huuskonen developed a set of solubility models for a data set of 1297 diverse compounds that were described by 24 atom-type E-state indices and 6 topological indices (n = 884, m = 30, r^2 = 0.89, RMSE = 0.67 and n_{test} = 413, q^2 = 0.88, $RMSE_{test}$ = 0.71).[9] The Huuskonen data set was also studied by Tetko et al. with purely 38 atom-type E-state indices as descriptors. The best model was generated by an artificial neural network (ANN): m = 33, r^2 = 0.91, and RMSE = 0.62.[10] Again the same data set was used by Liu and So to develop a simple ANN model using 19 descriptors (hydrophobicity, hydrophilicity, molecular weight, and 2D-topolgical indices): n = 1033, m = 19, r^2 = 0.86, RMSE = 0.70 and q^2 = 0.85, RMSE = 0.72 for leave-one-out cross-validation.[11] In a recent report, Yan and Gasteiger developed two QSPR models for the Huuskonen data set by a multilinear regression and an ANN.[12] The descriptors comprised a set of 32 values of a radial distribution function (RDF) code representing the 3D compounds and 8 additional correction factors that characterize molecular polarizability, relative aromatic and aliphatic degree, and the ability of atoms to participate in hydrogen bonding. A good predictive power was achieved: n = 797, r^2 = 0.79, AUE = 0.70, RMSE = 0.93 and n_{test} = 496, q^2 = 0.82, AUE_{test} = 0.68, $RMSE_{test}$ = 0.79 for the regression model. Recently, Delaney studied a much larger data set of 2874 compounds by using 9 simple descriptors that included calculated log$P$, molecular weight, aromatic proportion, non-carbon proportion, polar surface area, etc.[13] The performance of the model was listed as follows: n = 2874, m = 9, R^2 = 0.69, UAE = 0.75, RMSE = 1.01. In another report, Votano and Parham constructed a set of models with topological structure indices as descriptors using a variety of data analysis methods.[14] For the data set of 4115 aromatic compounds, the RMSE of the 772 test set molecules were 0.91, 1.01, and 1.04, for models developed with an ANN, partial least-square, and multiple linear regression analysis, respectively. Regarding the 1874 nonaromatic compounds, the RMSE of the 166 test set molecules were 0.75, 0.87, and 0.88, for the three aforementioned analysis methods, respectively. The disadvantage of this kind of methods is that they are dependent on the descriptors calculated from other theoretical models, and this kind of dependence produces additional difficulties to estimate the solubility of a molecule automatically.

In summary, most aqueous solubility models have a RMSE between 0.7 and 1.0 log units, and good performance is relatively easy to achieve for a smaller data set; QSPR models calculated with eq 1 usually require more descriptors; and most solubility models are constructed with linear regression and artificial neutral network. An ANN usually outperforms linear regression in generating QSPR models, but the difference is much smaller in terms of q^2, AUE, and RMSE for the external test sets. In addition, ANN models may not be interpretable and have a higher chance to be overfitted than regression models.

In this work, we introduced a new type of descriptor, solvent accessible surface area (SASA) classified by atom

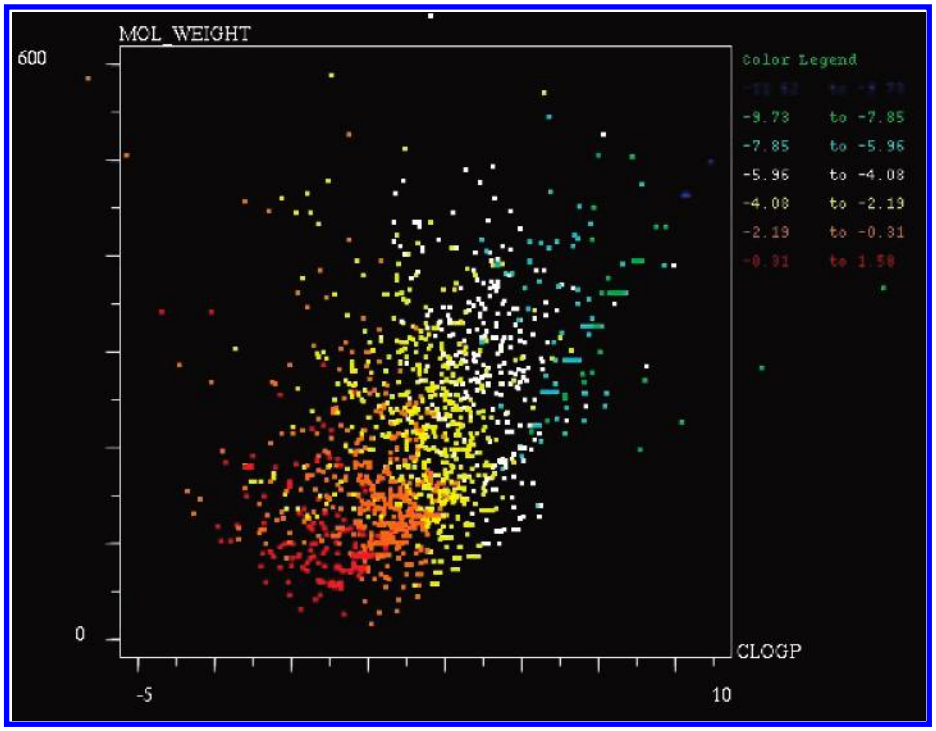

Figure 1. Distribution of 1708 aqueous solubility data in a 2D-chemical space defined by molecular weight and ClogP. The dots are colored by experimental solubility − log*S*.

types, to construct aqueous solubility models for a relatively large data set that consisted 1708 diverse molecules. In eq 1, c_i is now the SASA of atom type *i*, rather than the count of fragment *i* or atom type *i*. We believe that SASA is superior to the number of occurrences of a fragment or an atom type to correlate with aqueous solubility, simply because the solubility of a molecule depends on the relative interactions between solute−solute, solute−water, and water−water, and more exposed atoms have bigger contributions than inside atoms. In contrast, the application of counts of fragments or atom types as descriptors totally neglects the difference of exposed and inside atoms.

The aqueous models constructed 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 a database for high throughput screenings.

METHODS

1. Experimental Data Source. There were two major sources of experimental solubility data applied in this study, namely, the low molecular weight subset (1144) of the Delaney data set and the Huuskonen data set, which consisted of 1297 diverse compounds taken from the AQUSOL database of the University of Arizona and the PHYSPROP database. In addition, 28 novel compounds in Hou's data set and 9 more compounds through personal correspondence were included. The canonical SLNs (Sybyl Line Notation) of all 2478 molecules were generated and compared to each other. After removing duplicated entries, 1708 molecules were left. The experimental data of the Huuskonen set had higher priority to be adopted when duplication occurred. The compound names and experimental aqueous solubility values as well as SMILES are listed in Table S1 of the Supporting Information. The distribution of 1708 aqueous solubility data in a 2D-chemical space defined by molecular weight (MW) and ClogP is shown in Figure 1. The dots are colored by the experimental log*S*. It is shown that the 1708 molecules are almost evenly distributed in a chemical space defined by druglike molecules (MW from 50 to 550 and ClogP from −4 to 8).

2. Descriptors. 2.1. Atom Type Classified SASA. In the first place, for each molecule, a 3D-structure was generated by Concord module in Sybyl7.0.[15] Then its solvent accessible surface area was computed with a program developed by ourselves. 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. Similar to our previous work in developing a solvation free energy model based on SASA, a probe radius of 0.6 Å, rather than the water probe (1.4 Å), was applied to penetrate the molecular surface deeper to explore more details.[16] An atom type assignment was conducted with the Antechamber module in the Amber package.[17] At the beginning, each element had only one atom type, and additional atom types were introduced only if they were able to significantly improve the fitting. Besides SASA, a set of molecular properties that frequently appears in QSPR modeling was tested whether it could improve the model performance. It is notable that c_i in eq 1 are not only the SASA of atom type *i* but also the molecular properties *i* that enters the regressions.

2.2. Molecular Polarizability − Pol. Static molecular polarizability expresses how a molecule responds to an external electric field. Molecular polarizability is a measure of inductive and coefficient dispersion interactions within a molecule or a molecular system. The correlation square between polarizability and aqueous solubility was found to be 0.44. Molecular polarizability was calculated with an

Table 1. Coefficients of Each Descriptor of Two Aqueous Solubility Models Based on Solvent Accessible Surface Areas and Several Molecular Properties

descriptor	description	ASMS	ASMS-LOGP
CONSTANT	constant	0.973488	0.731810
CLOGP	calculated logP		−0.418476
POL	polarizability	−0.029167	−0.025182
MW2	square of molecular weight	0.000014	0.000009
HB	number of intramolecular hydrogen bonds	−0.168311	−0.146716
HYDRO-PHOB_C3	number of sp3 carbon in hydrophobic cores	−0.298797	−0.227676
HYDRO-PHOB_C2	number of sp2 carbon in hydrophobic cores	−0.078249	−0.076251
ho	H−O	−0.015301	−0.013147
hn	H−N	0.020421	0.009108
h4	H on sp2 carbon with one electron-withdrawal group	0.029957	0.018237
h5	H on sp2 carbon with two electron-withdrawal group	0.020605	0.004941
ha	H on sp1 and other sp2 carbons	0.007661	0.006955
h1	H on aliphatic sp3 carbon with one electron-withdrawal group	0.002733	0.002641
h23	H on aliphatic sp3 carbon with two or three electron-withdrawal group	0.016835	0.010372
hc	all other H	−0.007467	−0.000465
c1	sp1 C	−0.015212	−0.011148
c	C═O, C═S, or C═N	−0.009333	−0.009773
ca3	aromatic C with three other aromatic atoms, such as central atom of 1H-phenalene	−0.078561	−0.060715
ca2	aromatic C without hydrogen	−0.030905	−0.013338
ca	aromatic C attached to one hydrogen	−0.009187	0.001597
c2	all other sp2 carbon	−0.012682	−0.003022
c3a	sp3 carbon connected to a 11 or longer sp3 carbon-chain	0.086805	0.079860
c3	all other sp3 carbon	0.031380	0.028129
n1	nitrogens in cyano, N═N═R or N[XXX]N−R	0.008449	0.002777
nb	aromatic nitrogen, two substituents	0.011172	−0.000774
n2	other two-substitutent sp2 nitrogen	0.004006	−0.007129
n	N in amide group	−0.043182	−0.022602
na	sp2 N in planar ring with three substituents	−0.080639	−0.045499
nh	other sp2 N with three substituents	−0.019853	−0.013968
no	N in nitro group	0.115853	0.097132
n3_2	N in amine group with two hydrogens	−0.019914	−0.036936
n3_1	N in amine group with one hydrogen	0.063802	0.051510
n3	all other nitrogen	0.177633	0.149629
o21	O in aldehydates	0.007037	0.007653
o22	O in ketones	0.016423	0.002194
o23	sp2 O in carboxyl group, O═C−SH, COO−, or COS-	0.016058	−0.000931
o24	sp2 O in amide group	0.010781	0.001608
o26	sp2 O in ester	0.006691	−0.000406
o2n	O in nitro group	−0.017969	−0.014710
o2s	O═S	−0.003398	−0.006002
o2p	O═P	0.077398	0.045194
oh′	hydroxyl O in HO−C═O, HO−C═S, HO−C═NR, or HO−C═PR	0.040671	0.040837
oh	all other hydroxyl O	0.059812	0.040755
os′	sp3 O in RO−C═O, RO−C═S, RO−C═NR, or RO−C═PR	0.003099	0.011531
osp	sp3 O in RO−S═O, RO−S═S, RO−P═O, or RO−P═S	−0.030296	0.016977
os	all other sp3 O	−0.018591	−0.017665
sh	sp3 sulfur in thiol groups	−0.018385	−0.013581
ss	sp3 sulfur in -SR or S−S	−0.009582	0.000273
s	sp2 sulfur in S═P, S═C, etc.	−0.011547	−0.012442
s4	hypervalent sulfur, four substituents	0.056366	0.040530
p	any P	0.017886	−0.076058
f	any F	−0.012698	−0.005571
cl	any Cl	−0.012931	−0.004618
br	any Br	−0.015655	−0.006918
i	any I	−0.015565	−0.004802

empirical additive formula: $\alpha = -1.529 + 10.152\#C_{sp1} + 8.765\#C_{sp2} + 5.702\#C_{sp1} + 3.391\#H + 3.833\#F + 16.557\#Cl + 24.123\#Br + 38.506\#I + 10.488\#N_{nitro} + 6.335\#N_{othes} + 4.307\#O + 15.726\#S_{sulfone} + 22.366\#S_{others} + 11.173\#P$, where $\#C_{sp1}$ denotes the numbers of sp^1 carbon atoms in the molecules, etc. This polarizability model was developed for a data set of 420 molecules ($n = 420$, $m = 14$, $r^2 = 0.998$, RMSE = 1.490 and leave-one-out $q^2 = 0.998$, RMSE = 1.566).[18]

2.3. Calculated Logarithm of Water−Octanol Partition Coefficient − ClogP. log*P* is an index of molecular polarity, and it is highly correlated with aqueous solubility. The log*P* data in this study were estimated with the ClogP module implemented in Sybyl7.0.[15] The correlation coefficient square between ClogP and aqueous solubility was found to be 0.69. However the use of ClogP as a descriptor ruins our effort of developing a totally independent software package of predicting drugs ADMET properties. Therefore, two models, one using and the other not using ClogP, were constructed.

2.4. Molecular Weight − MW. Molecular weight is correlated with the size of the molecule. The correlation coefficient squares between aqueous solubility and molecular weight and square of molecular weight (MW2) were 0.39 and 0.29, respectively. In combination with SASA, we found that MW2, rather than MW, gave a better fitting. Hou et al. also found that MW2 was a better descriptor when combined with their atom type count descriptors.[7]

2.5. Intramolecular Hydrogen Bonding − HB. Hydrogen bond formation within solute itself or between solute and water can affect its solubility. We believe that the ability of a molecule to participate in hydrogen bonding with water may be well represented by some SASA descriptors.

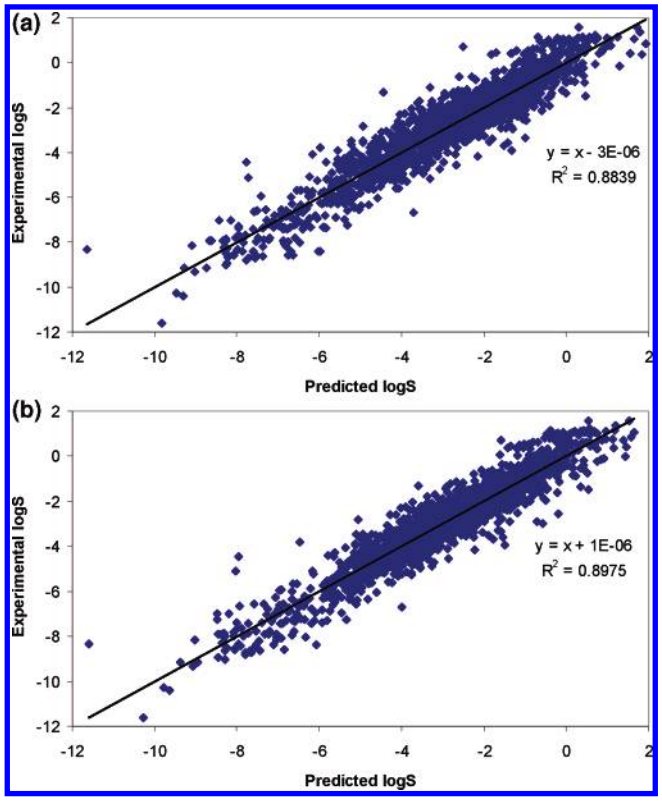

Figure 2. Plots of calculated versus experimental log*S* of 1708 molecules using ASMS and ASMS-LOGP with the corresponding regression equations: (a) model ASMS and (b) model ASMS-LOGP.

However, the ability of a molecule to form intramolecule hydrogen-bonding needs to be quantified. In this work, hydrogen donor D is either N or O with an attached hydrogen H, and hydrogen acceptor A is either N or O in functional groups except nitro and cyano. An intramolecule hydrogen bond was considered to be formed when the distance of AD was smaller than 2.5 Å and the angle of AHD was larger than 100°. A simple conformational search was performed when the above criteria were not met, as long as AD was smaller than 4.5 Å and D was not in a ring. A bond formed with D except H was rotated six times at a step of 60°, and a hydrogen bond was considered to be formed if at least one conformation met the criteria. HB is the total number of all possible hydrogen bonds. The correlation coefficient square was calculated as 0.013 between HB and experimental aqueous solubility.

2.6. Hydrophobicity − HB_C2 and HB_C3. Hydrophobicity is another factor that makes substantial contribution to aqueous solubility. Although it is partially accounted by ClogP and some of the SASA descriptors, adding hydrophobicity terms explicitly can improve the QSPR models. In this work, a hydrophobic cluster was a collection of sp2 and sp3 carbons, and from any atom in the cluster there was no other kind of heavy atoms within any atomic path that had less than or equal to seven atoms. HB_C2 and HB_C3 are the numbers of sp^2 carbon and sp^3 carbons in hydrophobic clusters of a molecule, respectively. The r^2 of aqueous solubility to HB_C2 and HB_C3 are 0.10 and 0.06, respectively.

3. Model Validation. The two aqueous solubility models were extensively validated by three types of tests. First of all, a leave-one-out analysis was carried out for both models.

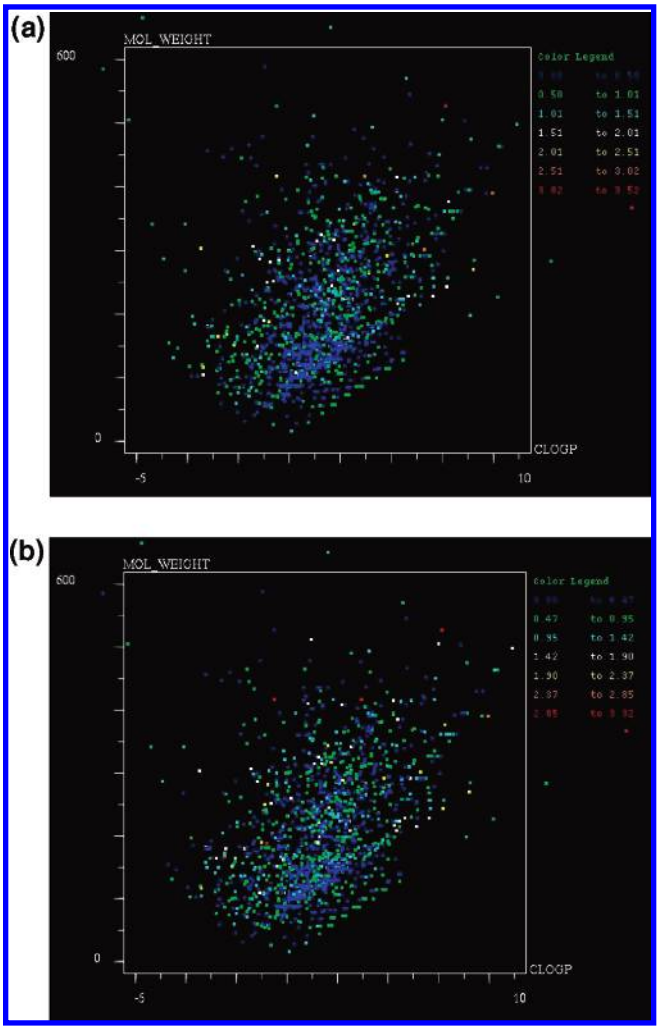

Figure 3. Distribution of 1708 aqueous solubility data in a 2D-chemical space defined by molecular weight and ClogP. The dots are colored by residues of prediction with (a) model ASMS and (b) model ASMS-LOGP.

Table 2. Performance of 10,000 Times 90/10 Cross-Validation for Both Models

	minimum	maximum	mean	RMSD
		ASMS		
AUE	0.41	0.68	0.527	0.034
RMSE	0.54	0.97	0.699	0.054
q^2	0.77	0.94	0.884	0.021
		ASMS-LOGP		
AUE	0.46	0.70	0.570	0.034
RMSE	0.58	1.00	0.742	0.053
q^2	0.75	0.93	0.869	0.022

Second, the whole molecular set was classified into 240 groups by a cluster analysis based on 2D-similarity.[15] One molecule from each group was randomly picked up to compose a test set (15% of the whole data set). The other molecules entered the training set. Two corresponding QSPR models, which were constructed by using the 1468 molecules in the training set, were used to make predictions for the other 240 molecules in the test set. Third, a 90/10 (10% randomly selected data in the test set and 90% data in the training set) cross-validation was run for 10,000 times. For each run, the aqueous solubility of the test set molecules was predicted with the models based on the training set.

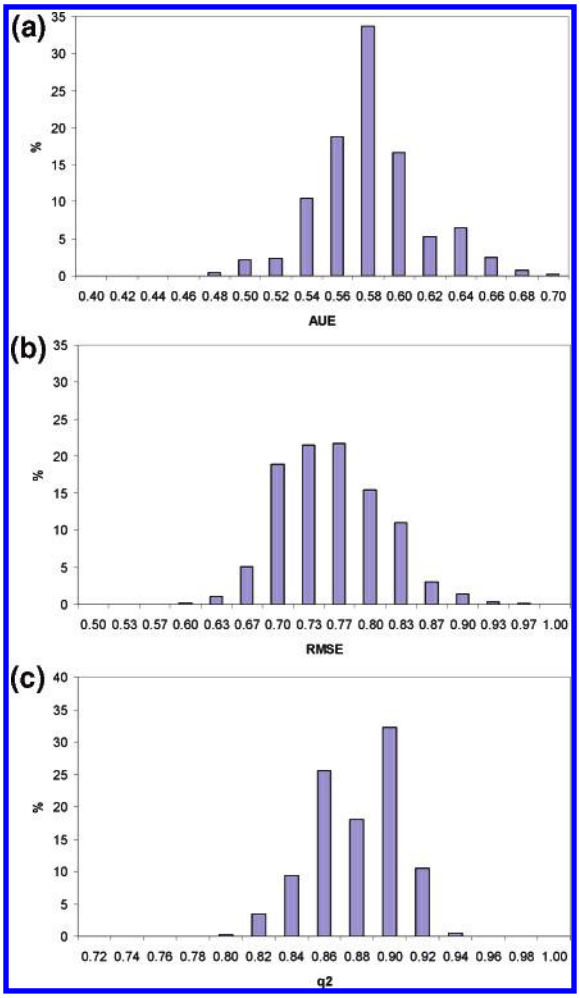

Figure 4. Distributions of AUD (a), RMSE (b), and q^2 (c) of 10,000 times 90/10 cross-validations by clustered column charts for model ASMS.

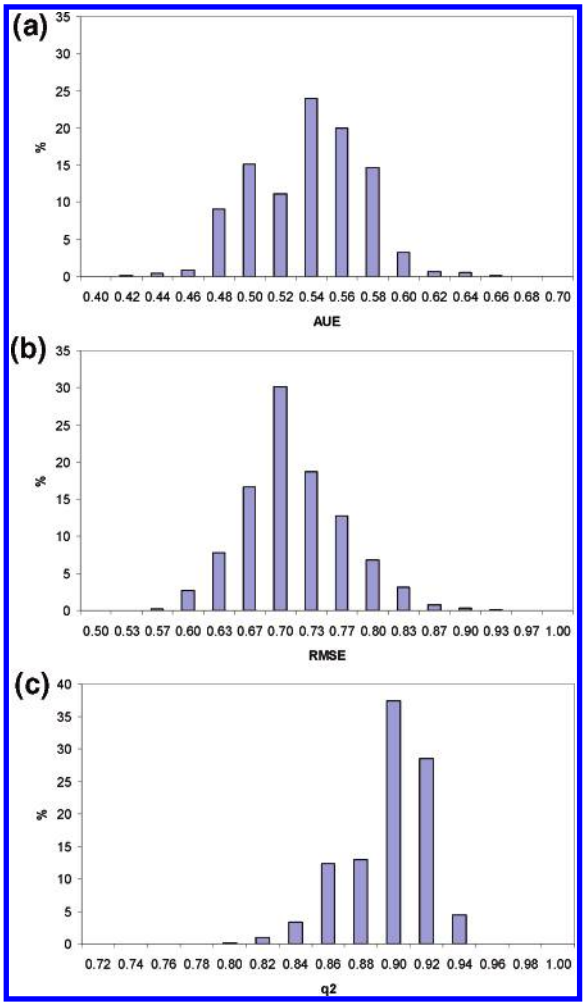

Figure 5. Distributions of AUD (a), RMSE (b), and q^2 (c) of 10,000 times 90/10 cross-validations by clustered column charts for model ASMS-LOGP.

4. Aqueous Solubility Prediction for a Number of Conformation Sets. To investigate how sensitive the aqueous solubility models to different conformers, 100 molecules that have a number of rotatable bonds ranging from 0 to 10 were randomly selected and submitted for conformational searches with the Omega program (Openeye Inc.).[19] Then aqueous solubility prediction was conducted with both models.

5. Application of Aqueous Models in Druglike Analysis. Both ASMS and ASMS-LOGP were applied to predict aqueous solubility for molecules in seven prestigious databases: (1) ZINC druglike subset[3] − 104,712 molecules, ~5% of the whole data set; (2) Available Chemical Directory (ACD)[20] − 61,258 molecules, ~30% of the whole database; (3) Comprehensive Medicinal Chemistry (CMC)[20] − 4134 molecules, ~50% of the whole database; (4) World Drug Index (WDI)[21] − 25,783, ~40% of the whole database; (5) Drug − 1536 molecules; (6) Injection Drug − 314 molecules; and (7) Oral Drug − 643 molecules. For (1)−(4), certain percentages of molecules in the whole databases were selected randomly to form the data sets for studies. Molecules in (5)−(7) are actual drugs approved by the U.S. Food and Drug Administration since 1970.

RESULTS AND DISCUSSION

1. Two Aqueous Solubility Models Based on SAS. In total, we have collected 1708 diverse molecules to develop aqueous solubility models. This data set is significantly larger than most data sets used by other researchers. Two aqueous solubility models, either without (ASMS) or with (ASMS-LOGP) ClogP as a descriptor, have been developed using eq 1. The application of SASA as descriptors enables us to use only a very limited number of atom types to achieve good fitting performance. The definition of 50 atom types defined in this work is listed in Table 1. In comparison, Klopman and Zhu[6] applied 171 fragments and Hou et al.[7] utilized 76 atom types to study much smaller data sets (1168 and 1290, respectively). Since most atom types in this work were defined according to atomic number, hybridization, or a very localized chemical environment (mostly in a single functional group), it is unlikely that our models suffer from the problem of missing atom types and make biased prediction for novel molecules.

The performance of the two models is very encouraging. For ASMS, leave-one-out $q^2 = 0.872$ and RMSE = 0.748 log unit using the first 18 components suggested by a partial least-square (PLS) analysis. For full component analysis, $r^2 = 0.884$, AUE = 0.547, RMSE = 0.707, $F = 812.5$. ASMS-LOGP is slightly better than ASMS: $q^2 = 0.886$ and RMSE

Table 3. Statistical Results of Aqueous Solubility Prediction Using Model ASMS and Model ASMS-LOGP for 55 Molecules That Have Multiple Conformations

	comp ID	expt	rotatable bonds	confrmtn number	ASMS MEAN	ASMS RMSE	ASMS-LOGP MEAN	ASMS-LOGP RMSE
1	28	−2.52	1	4	−3.37	0.04	−3.36	0.04
2	30	−2.65	4	18	−3.35	0.13	−3.58	0.08
3	33	−2.28	5	9	−1.54	0.24	−2.25	0.03
4	131	−3.01	4	4	−2.22	0.11	−2.4	0.06
5	313	−3.26	3	2	−3.12	0.00	−3.14	0.00
6	361	−5.27	1	2	−4.47	0.00	−4.61	0.00
7	387	−4.16	7	105	−4.14	0.30	−4.05	0.26
8	435	−5.74	8	62	−5.72	0.16	−5.92	0.10
9	447	−4.82	4	14	−5.22	0.13	−4.94	0.13
10	457	−3.66	7	108	−4.35	0.14	−3.94	0.11
11	464	−4.88	2	23	−5.15	0.11	−4.97	0.06
12	465	−4.23	1	3	−4.63	0.05	−4.56	0.02
13	484	−4.19	4	9	−4.94	0.06	−4.92	0.03
14	493	−3.12	4	9	−2.17	0.06	−2.4	0.03
15	502	−1.5	4	10	−1.71	0.13	−1.11	0.09
16	526	−2.68	3	4	−3.03	0.08	−3.36	0.08
17	533	−1.74	1	2	−2.69	0.01	−2.64	0.01
18	622	−1.32	5	168	−4.15	0.26	−3.55	0.19
19	645	−3.68	0	2	−4.05	0.02	−3.83	0.01
20	668	−2.86	1	2	−3.35	0.01	−2.95	0.01
21	714	0.79	1	4	−0.8	0.07	−0.74	0.04
22	759	−5.07	6	64	−3.98	0.13	−4.26	0.10
23	774	−1.5	2	3	−1.33	0.15	−1.48	0.09
24	776	−0.45	2	8	−0.96	0.17	−0.87	0.10
25	788	−0.53	5	12	−0.76	0.09	0.06	0.07
26	856	−5.68	8	3	−5.06	0.01	−5.39	0.00
27	914	−1.37	3	3	0.09	0.11	0.37	0.08
28	916	−2.22	8	41	−1.84	0.15	−2.52	0.13
29	1003	−2.1	10	163	−2.76	0.25	−3.05	0.18
30	1025	−1.76	1	2	−0.92	0.00	−1.37	0.00
31	1033	−0.35	2	3	−1.39	0.07	−1.5	0.05
32	1077	−3.02	6	17	−2.79	0.09	−3.14	0.05
33	1079	-1.51	5	7	0.57	0.20	0.52	0.16
34	1111	−0.85	0	2	0.26	0.06	−0.23	0.04
35	1185	−2.18	1	4	−2.31	0.06	−2.25	0.05
36	1262	−2.35	4	2	−1.47	0.06	−1.83	0.06
37	1304	−1.2	4	7	−1.45	0.06	−1.6	0.07
38	1338	−0.43	4	4	−0.97	0.04	−0.69	0.05
39	1390	−4.07	6	2	−4.07	0.02	−4.52	0.01
40	1463	−0.72	4	6	−0.65	0.03	−0.78	0.02
41	1502	−0.54	7	17	−1.05	0.18	−0.91	0.14
42	1581	−1.88	2	4	−1.76	0.09	−2.09	0.05
43	1598	−3.93	6	400	−3.33	0.11	−3.32	0.06
44	1605	−1.34	2	3	−1.09	0.03	−1.07	0.01
45	1608	−1.62	4	21	−2.13	0.15	−2.43	0.09
46	1616	−2.93	8	213	−3.03	0.14	−3.28	0.09
47	1624	−4.6	1	2	−4.68	0.00	−5.44	0.00
48	1642	−2.3	7	66	−2.55	0.15	−2.55	0.12
49	1653	−3.81	4	4	−2.72	0.08	−3.15	0.03
50	1655	−1.83	2	4	−1.08	0.01	−1.7	0.01
51	1677	−1.23	2	5	−1.74	0.09	−1.22	0.05
52	1689	−4.19	9	113	−4.84	0.17	−4.72	0.14
53	1691	0.54	2	24	0.41	0.20	0.27	0.14
54	1705	−2.42	3	7	−1.91	0.07	−1.93	0.03
55	1707	−4.86	9	67	−5.07	0.12	−5.27	0.12

= 0.705 (leave-one-out cross-validation using the first 18 components according to PLS analysis), r^2 = 0.897, RMSE = 0.664, AUE = 0.505, F = 706.0 (full component analysis).

The plots of experimental versus predicted aqueous solubility are shown in Figure 2. The distribution of the prediction errors in a two-dimensional space defined by molecular weight and ClogP is shown in Figure 3. It is clear that prediction errors are evenly distributed in the space for both models.

The performance of ASMS-LOGP is only slightly better than that of ASMS, indicating log*P* is implicitly taken into account in the ASMS model. The advantage of ASMS over ASMS-LOGP is that the ASMS model is not affected by the potential error introduced by the log*P* calculations. More importantly, ASMS can be applied directly to conduct high throughput aqueous solubility prediction with our ADMET software package.

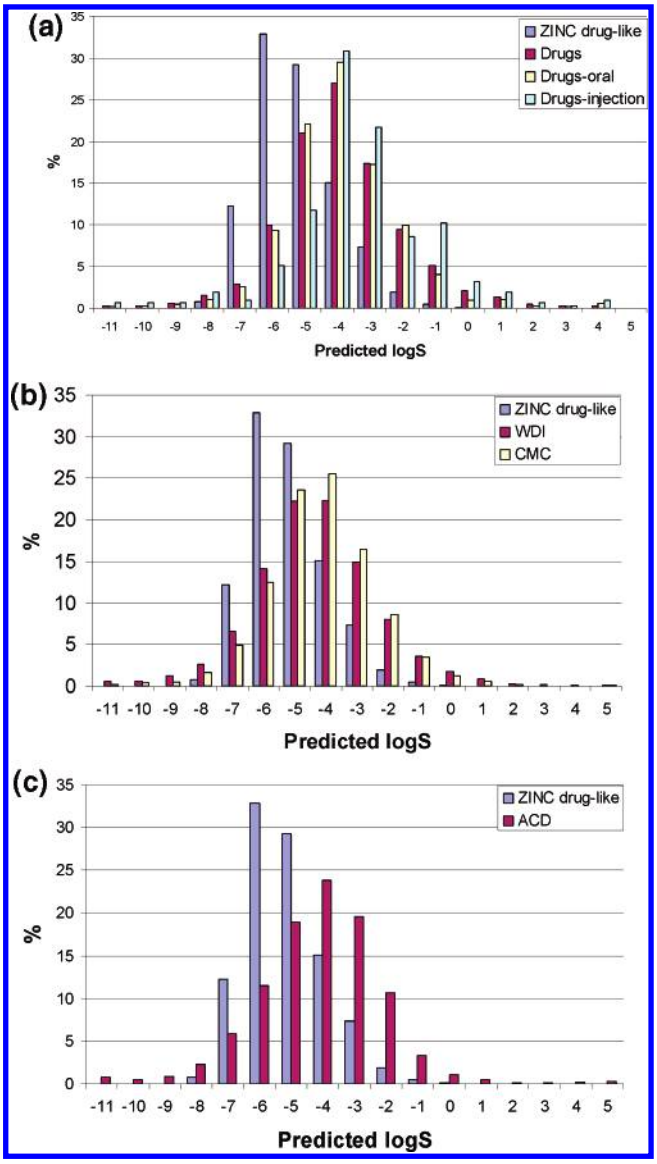

Figure 6. Distribution of predicted aqueous solubility − log*S* by model ASMS for seven data sets: (a) ZINC (in blue) versus drug data sets, (b) ZINC versus WDI and CMC, and (c) ZINC versus ACD.

Table 4. Statistical Results of Predicting Aqueous Solubility with Both Model ASMS and Model ASMS-LOGP for a Set of Databases

model	database	no. of compds	min.	max.	mean	RMSD
ASMS	ZINC	10412	−8.45	3.28	−4.58	1.22
	ACD	61258	−18.35	18.12	−3.68	2.04
	CMC	4134	−12.19	6.11	−3.73	1.73
	WDI	25783	−16.38	8.90	−3.88	2.01
	Drug	1536	−11.91	4.47	−3.38	1.92
	Drug-Injection	314	−10.83	4.47	−2.90	2.12
	Drug-Oral	643	−10.71	4.47	−3.41	1.79
ASMS-LOGP	ZINC	10412	−9.14	4.77	−4.72	1.23
	ACD	61258	−17.982	7.10	−3.65	1.99
	CMC	4134	−12.00	4.91	−3.66	1.74
	WDI	25783	−16.63	4.91	−3.83	2.06
	Drug	1536	−11.78	3.11	−3.36	1.93
	Drug-Injection	314	−11.63	3.11	−2.93	2.03
	Drug-Oral	643	−10.55	2.84	−3.42	1.78

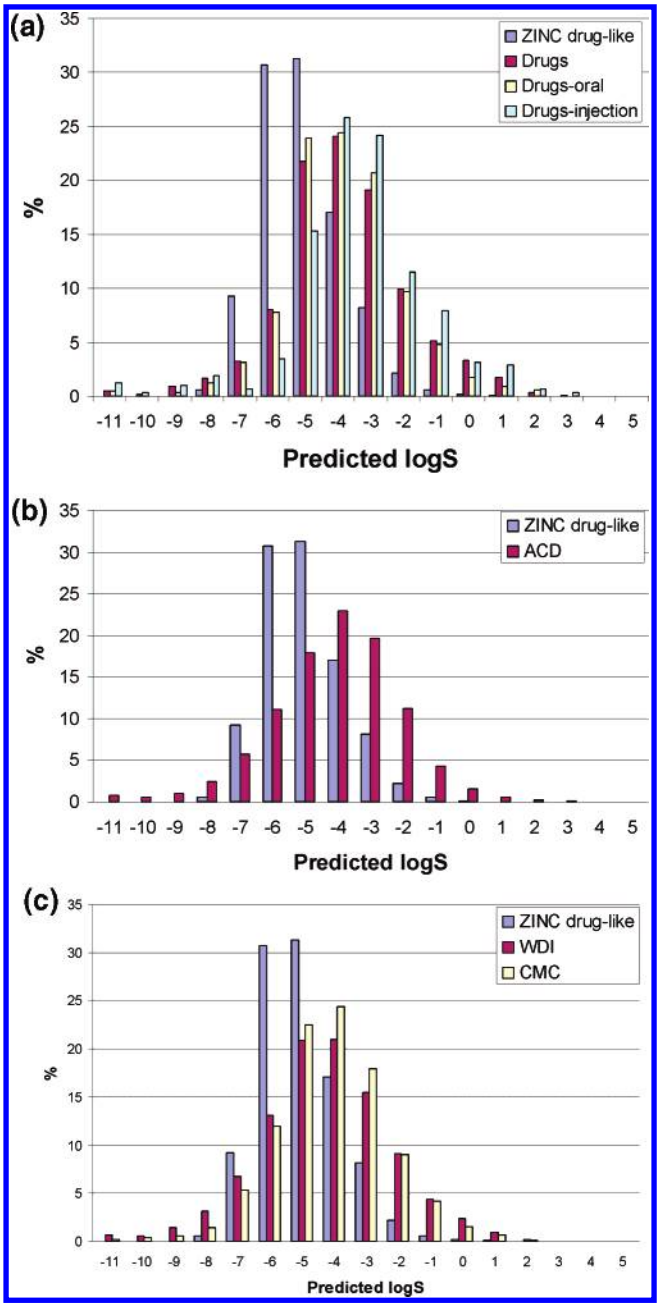

Figure 7. Distribution of predicted aqueous solubility − log*S* by model ASMS-LOGP for seven data sets: (a) ZINC (in blue) versus drug data sets, (b) ZINC versus WDI and CMC, and (c) ZINC versus ACD.

2. The Predictability and Applicability of Two Models. Both ASMS and ASMS-LOGP have been extensively validated in three tests. First of all, the leave-one-out q^2 is very close to r^2 of full component regression analysis: 0.872−0.884 for ASMS and 0.886−0.897 for ASMS-LOGP. Similarly, the RMSE values are also very close. In the second test, 240 structurally diverse molecules are selected to compose a test set. The models constructed on the other 1468 molecules are used to predict solubility of the molecules in the test set. The AUE and RMSE are 0.617 and 0.792 log units for ASMS, respectively, and 0.558 and 0.724 for ASMS-LOGP, respectively. In the third, 10,000 times 90/10 cross-validation are carried out for both models. The results are listed in Table 2, and the distributions of AUE, RMSE, and q^2 are represented by clustered column charts in Figures 4 and 5. Interestingly, for both models the mean RMSE and q^2 are very close to those of leave-one-out analysis: 0.748−0.742 (RMSE) and 0.872−0.869 (q^2) for model ASMS and 0.705−0.699 (RMSE) and 0.886−0.884 (q^2) for model ASMS-LOGP. It is concluded that leave-one-out analysis is a more reliable cross-validation approach than simply dividing a whole data set into training and test sets arbitrarily. And both models are reliable in predicting aqueous solubility of novel compounds.

The applicability of the two models is well characterized by Figure 1. In this figure, molecular weight, which describes the size of a molecule, and ClogP, which describes the polarity of a molecule, define a two-dimensional space. It is clear that experimental aqueous solubility is well distributed in the whole space, indicating the molecular set we studied is adequately diverse and the model can make a suitable prediction for a variety of molecules. Another reason that MW and ClogP were chosen to define the chemical space is because both parameters have high correlation to the experimental solubility.

Compared to other well-established solubility models including those implemented in commercial software packages, our models achieved better or comparable performance, although we studied a considerable larger data set. The high predictability is demonstrated by the very small differences between RMSE and $RMSE_{test}$ of both models.

3. How Do Conformations Affect the Performance of the Models? For the 100 molecules randomly selected to investigate how different conformers affect the solubility prediction, only 55 have more than one conformer. The number of rotatable bonds, the number of conformations, and the mean and root-mean-square error of predicted solubilities are listed in Table 3. The average AUE and RMSE to the means are 0.099 and 0.122 log unit, respectively. The largest RMSE among 55 molecules is 0.30 for compound 387. The very small average AUE and RMSE indicate that the prediction is not sensitive to conformations. However, it is recommended to use 2D to 3D conversion programs, such as Concord as we used,[15] to produce reasonable 3D structures for SASA calculations. It is notable that no further structural minimization is needed prior to solubility prediction with both ASMS and ASMS-LOGP. Furthermore, when a conformational ensemble is used to predict aqueous solubility, it is very unlikely to produce much different results for the conformational ensemble in water from that in vacuum.

4. Application of the Aqueous Models in Drug Design. A model is useful only when it can be applied to solve practical problems. Both models were utilized to predict aqueous solubility for the seven data sets described in the methodology section. The average, minimum, and maximum solubility of each data set are listed in Table 4. The distributions of predicted aqueous solubility are represented by a set of clustered column charts in Figures 6 and 7. It is concluded that real drugs are about 20-fold more soluble than the so-called druglike molecules in the ZINC database. More specifically, injection drugs are about 50−60-fold more soluble, while oral drugs are about 16-fold more soluble. The enrichment plots for both models are shown in Figures 8 and 9. If the cutoff of a molecule to be soluble is −5.0 log unit, about 85% of true drugs are soluble, whereas only about 50−55% of "druglike" molecules in the ZINC database are

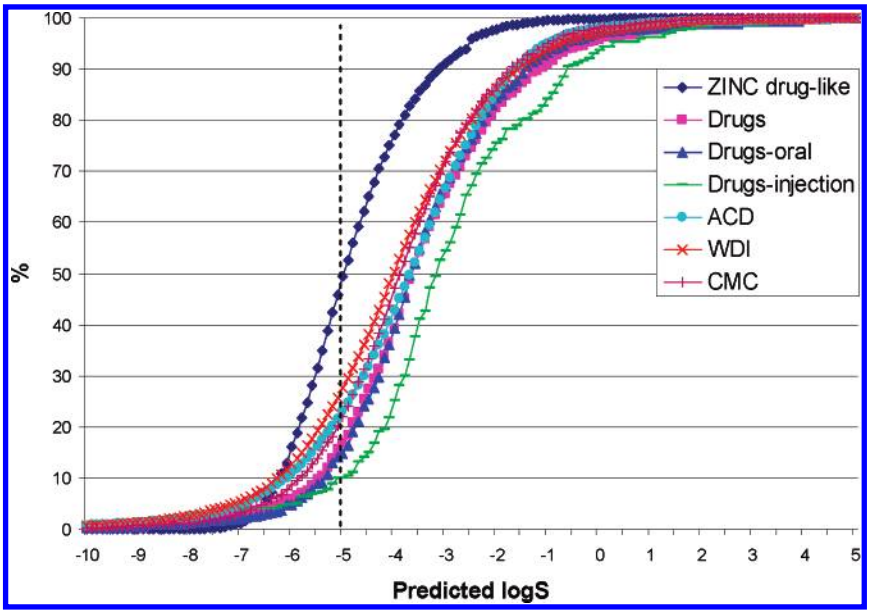

Figure 8. Enrichment plots of screening a number of data sets using model ASMS. Take the predicted log*S* of −5 as a threshold, about 15% of the molecules in the Drugs data set were screened out, while about 50% of the molecules in the ZINC druglike data set were removed.

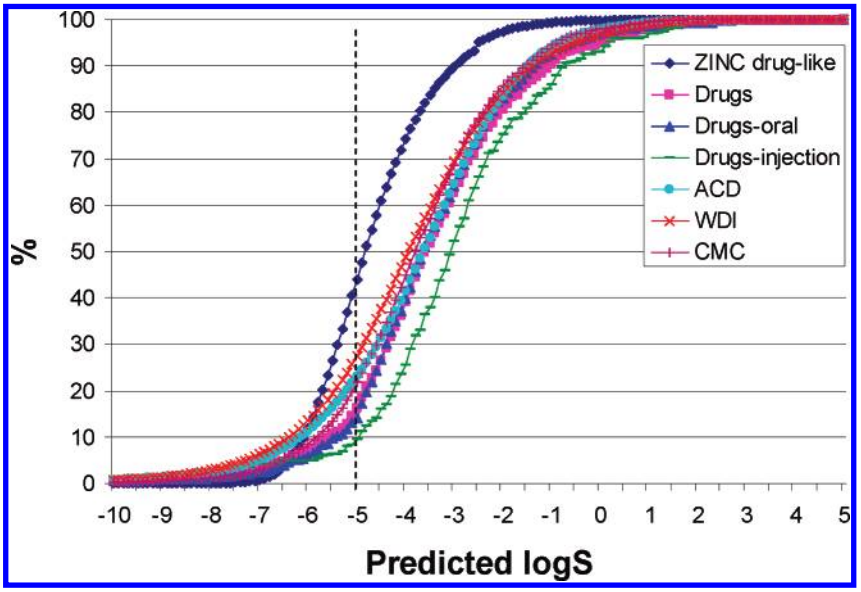

Figure 9. Enrichment plots of screening a number of data sets using model ASMS-LOGP. Take the predicted log*S* of −5 as a threshold, about 13% of the molecules in the Drugs data set were screened out, while about 45% of the molecules in the ZINC druglike data set were removed.

soluble. This is a remarkable result, and it might imply the necessity of further improving the quality of compound libraries for HTS. Based on Table 1, one might get some hints on how to modify the molecules to improve aqueous solubility. For example, the introduction of a hydroxyl group could improve solubility about 0.4−0.6 log unit if the effect on other molecular descriptors (polarizability, molecular weight, number of the intramolecular hydrogen bonds, etc.) is ignored. That is to say, the contribution of one hydroxyl functional group is simply estimated by summing up the products of solvent accessible surface areas of hydrogen and oxygen multiplying their corresponding coefficients (−0.0153 for H and 0.0598 for O in the ASMS model). In practice, one should also consider the possible side effects caused by the modification on other important properties, especially, bioactivity, selectivity, bioavailability, etc.

In conclusion, our solubility models can serve as a new rule to evaluate how a molecule is druglike. They can also serve as a filter to prioritize databases prior to HTS. Unlike other filters, such as similarity based on fingerprints or pharmacophore models, the solubility filter is research project independent and may serve as a general filter as the "Rule of 5" in drug discovery.

CONCLUSIONS

In this work, we successfully developed two aqueous solubility models using atom type classified solvent-accessible surface areas for a large diverse molecular data set. For model ASMS, $r^2 = 0.884$, RMSE = 0.707, leave-one-out $q^2 = 0.872$, and RMSE = 0.748. ASMS-LOGP, which included ClogP as a descriptor, achieved better performance: $r^2 = 0.897$, RMSE = 0.664, leave-one-out $q^2 =$

0.886, and RMSE = 0.705. Both models were thoroughly evaluated by three cross-validation tests and showed good predictability. Both models were applied to HTS prediction of aqueous solubility for a variety of data sets extracted from a set of widely used databases in pharmaceutical companies. It is found that the real drugs are about 20 times more soluble than those "druglike" molecules in the ZINC database. With a cutoff of −5 applied in log*S* prediction with both ASMS and ASMS-LOGP models, 50% of "druglike" molecules in ZINC were eliminated, while only less than 15% of real drugs were screened out. We believe our reliable aqueous solubility models will have a great use in drug discovery.

Abbreviations. QSPR, quantitative structure-property relationship; SASA, solvent accessible surface area; ADMET, absorption, distribution, metabolism, excretion, and toxicity; HTS, high throughput screening; CLogP, calculated logP; 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; PLS, partial least-square.

ACKNOWLEDGMENT

We are grateful to acknowledge the financial support from BASF, the Chemical Company, and the research support from NCSA (MCB000013N (J. Wang and P.I.)).

Supporting Information Available: Compound names and the SMILES as well as the experimental values of the 1708 molecules (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151−185.
(2) 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.
(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) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.
(5) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.
(6) 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.
(7) 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.
(8) 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.
(9) 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.
(10) 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.
(11) 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.
(12) 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.
(13) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(14) 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.
(15) *Sybyl user manual*; Tripos Inc.: St. Louis, MO, U.S.A., 1995.
(16) Wang, J.; Wang, W.; Hou, S.; Lee, M.; Kollman, P. A. Solvation model based on weighted solvent accessible surface area. *J. Phys. Chem. B* **2001**, *105*, 5055−5067.
(17) 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.
(18) 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.
(19) *Openeye Scientific Software*; Santa Fe, NM, U.S.A., 2006.
(20) MDL Information Systems, Inc., San Ramon, CA, U.S.A., 2004.
(21) Thomson Scientific, Inc. Philadelphia, PA, U.S.A., 2006.

CI700096R