Global and Local Computational Models for Aqueous Solubility Prediction of Drug-Like Molecules

Christel A. S. Bergström,[†] Carola M. Wassvik,[†] Ulf Norinder,[†,‡] Kristina Luthman,[§,*] and Per Artursson[†]

Center for Pharmaceutical Informatics, Department of Pharmacy, Uppsala University, Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden, Department of Medicinal Chemistry, AstraZeneca R&D, SE−151 85 Södertälje, Sweden, and Department of Chemistry, Medicinal Chemistry, Göteborg University, SE−412 96 Göteborg, Sweden

Received March 11, 2004

The aim of this study was to develop in silico protocols for the prediction of aqueous drug solubility. For this purpose, high quality solubility data of 85 drug-like compounds covering the total drug-like space as identified with the ChemGPS methodology were used. Two-dimensional molecular descriptors describing electron distribution, lipophilicity, flexibility, and size were calculated by Molconn-Z and Selma. Global minimum energy conformers were obtained by Monte Carlo simulations in MacroModel and three-dimensional descriptors of molecular surface area properties were calculated by Marea. PLS models were obtained by use of training and test sets. Both a global drug solubility model ($R^2 = 0.80$, $RMSE_{te} = 0.83$) and subset specific models (after dividing the 85 compounds into acids, bases, ampholytes, and nonproteolytes) were generated. Furthermore, the final models were successful in predicting the solubility values of external test sets taken from the literature. The results showed that homologous series and subsets can be predicted with high accuracy from easily comprehensible models, whereas consensus modeling might be needed to predict the aqueous drug solubility of datasets with large structural diversity.

INTRODUCTION

Virtual screening filters for pharmacokinetic properties such as absorption, distribution, metabolism, and elimination (ADME) could allow a cost-effective lead optimization process by replacing labor-consuming experimental techniques with computation. Ideally, such high-input screening would result in a higher output of developable compounds, which would imply a faster and less expensive drug discovery and development process. In this study, we have developed computational models based on molecular descriptors of drug-like molecules for prediction of aqueous drug solubility, one of the major factors influencing drug absorption and hence, the pharmacokinetic profile of a compound.

Computational models for the prediction of aqueous solubility from electrotopology, molecular surface areas, lipophilicity, and hydrophilic measures have been devised, and several of these show impressive statistics.[1−10] However, the majority of the solubility models are trained on non drug-like molecules containing structural features not commonly seen in drugs. A drawback of this approach is that functional groups frequently included in drug-like molecules, such as amines, amides, and carboxylic acids, are generally underrepresented in the non drug-like training sets. Hence, the descriptor space defined by such training sets does not cover the entire drug-like space and the applicability of the models in the prediction of aqueous drug solubility still remains to be shown.

In the effort to develop computational models, the quality of the experimental input data is often discussed. Several authors claim that the different techniques and solvents used in the determination of aqueous drug solubility result in inconsistent solubility data, as solubility data for one compound may differ 2−3-fold[5,11] and up to as much as 20-fold for some compounds[12] when determined in different laboratories. Thus, to be confident about the quality of the experimental data used for the computational analysis, it is preferable to perform solubility measurements in-house, using a standardized method. We have previously determined the solubility of drug-like compounds using two highly reproducible techniques; the small-scale shake flask method[13] and a potentiometric technique.[14] Due to limitations in experimental capacity, we have so far only worked with smaller datasets based on highly accurate solubility data. Computational models generated from these data showed that molecular surface areas obtained from the three-dimensional (3D) representation of the molecule are potential descriptors for aqueous drug solubility. Our, as well as others', computational models of aqueous solubility have relied on the assumption that the solvation process is more important for solubility than is the influence of the solid state. This assumption is supported by our recent multivariate data analysis of the melting point, a commonly used solid-state characteristic, which showed that the surface area descriptors selected in the solubility models were related to the solvation process rather than the solid-state properties.[15] This result

* Corresponding author phone: 46 31 772 2894; Fax: 46 31 772 3840; e-mail luthman@mc.gu.se. Department of Chemistry, Medicinal Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden.
[†] Uppsala University.
[‡] AstraZeneca R&D Södertälje.
[§] Göteborg University.

10.1021/ci049909h CCC: $27.50
Published on Web 06/23/2004

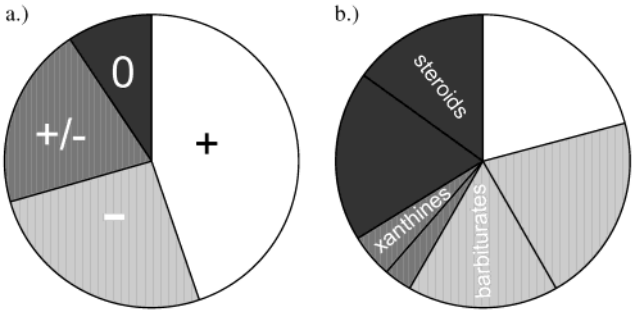

Figure 1. Schematic figure of the drug-like space used for building global and local models. (a) The drug-like compounds, which were used for the generation of global models, were partitioned into four subsets in the development of local models; acids (gray), bases (white), ampholytes (dark gray), and nonproteolytes (black). No large homologous series were found within the 85 compounds studied. (b) The distribution of the external test set within the four subsets colored as in figure a. The size of the largest homologous series found within each subset is shown. The external test set was skewed with regard to proteolytic function, since a majority of the compounds were acids or nonproteolytes.

made us interested in investigating to what extent solvation descriptors alone can predict aqueous solubility of larger datasets of structurally diverse drug-like molecules.

In this work we therefore studied the general applicability of the surface area descriptors in drug solubility prediction. To obtain a dataset that was large in number, broad in structural diversity, drug-like, and showed good experimental quality, we compiled our in-house drug solubility data[13,14] with high quality solubility data obtained from the drug development settings in the pharmaceutical industry.[16] Moreover, to analyze whether the surface area descriptors were better at estimating the solubility of compounds containing a particular proteolytic group, subset specific models were developed (Figure 1a). Furthermore, to study to what extent a further decrease in structural diversity results in more quantitative models, models for the prediction of compound homologues were devised (Figure 1b). The predictions obtained from the surface area descriptors (3D) were compared with predictions based on descriptors calculated from the 2D representation of the molecule, which previously have been shown to be successful in solubility predictions.[3,4,6,9] Finally, we challenged the computational models obtained with an external drug-like dataset ($n = 207$) taken from the literature.[4,7]

METHODS

Datasets. Values for the intrinsic solubility of 85 drug-like molecules were determined in-house[13,14] or obtained from reliable sources within the pharmaceutical industry.[16−19] The ChemGPS methodology[20] (Figure 2) and PCA (Figure 3) were applied to evaluate the diversity in physicochemical space for the selected molecules. ChemGPS is a navigation tool that defines the drug-like space using PCA prediction based on core and satellite structures. The 85 compounds included in our dataset were analyzed for structural diversity by ChemGPS predictions from descriptors generated by the program Selma (see the section 'Calculation of 2D Descriptors' below). This analysis, and the PCA analysis of physicochemical properties, showed that the compounds displayed large variations in the calculated descriptor space. To investigate the use of local models (Figure 1), the 85 compounds were divided into their chemical subgroups (acids, bases, ampholytes, and nonproteolytes). Compounds not displaying a proteolytic function with a pK_a between 2 and 12 were considered to be nonproteolytes. The dataset was comprised of 22 acids, 38 bases, 17 ampholytes, and 8 nonproteolytes.

An external test set compiled from data available in the literature commonly used for model development[4,7] was used to challenge the models obtained. Compounds that overlapped between the three datasets (i.e., compounds that were included in the 85 compounds in this work and in the Huuskonen dataset or the Jorgensen dataset) were excluded from the external test set, resulting in 207 compounds that were included in the final external test set used to challenge the global model. The compounds of the external test set were also divided into their chemical subgroups (78 acids, 43 bases, 16 ampholytes and 70 nonproteolytes) and used as validation for the local models.

Calculation of 2D Descriptors. The 2D descriptors were calculated with Molconn-Z[21] and Selma.[22] The Molconn-Z was used to calculate the atom-type electrotopological state indices. Briefly, the electrotopological state indices for a particular atom are values resulting from its topological and electronic environment. The indices encode the electronegativity as well as the local topology of each atom by considering perturbation effects from the neighboring atoms. Selma is a software package that generates descriptors related to size (e.g., molecular weight), ring structure (e.g., number of rings), flexibility (e.g., how many flexible side chains), number of hydrogen bonds, polarity (e.g., 2D PSA), Kier connectivity indices,[23] BCUT parameters related to connectivity and atom weights,[24] electronic environment, charge (e.g., partial charges) and lipophilicity. Molconn-Z and Selma generated 566 different descriptors.

Lipophilicity. The lipophilicity (ClogP) of the drugs was calculated with the BioByte software MacLogP version 2.0 (BioByte Corp., Claremont, CA).

Conformational Analysis. To speed up the process, a rapid conformational analysis was performed to obtain a reasonable conformation for the calculations of the 3D descriptors. A 500-step Monte Carlo conformational search was performed in MacroModel version 6.5 on a Silicon Graphics Octane workstation. The energy minimization was performed in a simulated water environment on the compounds in their unionized state using MMFF.

Calculation of 3D Descriptors. The in-house computer program Marea[25] was used to calculate the free surface area of each atom as well as the molecular volume of the conformation with the lowest energy. The atomic van der Waals radii used were as follows: sp and sp^2 carbons 1.94 Å; sp^3 carbons 1.90 Å; oxygen 1.74 Å; nitrogen 1.82 Å; sulfur 2.11 Å; chlorine 2.03 Å; fluorine 1.65 Å; electroneutral hydrogen 1.50 Å; hydrogen bond to oxygen 1.10 Å; and hydrogen bond to nitrogen, 1.125 Å (obtained from PC-MODEL version 4.0). Only static surface areas obtained from the global minimum conformer of each search were calculated and used in the model development.

The polar surface area (PSA) was defined as the area occupied by oxygen and nitrogen atoms and hydrogen atoms attached to these heteroatoms. The nonpolar surface area (NPSA) was defined as the total surface area (SA) minus the PSA. The SA was divided into the partitioned total surface area (PTSA). Each PTSA corresponds to the surface

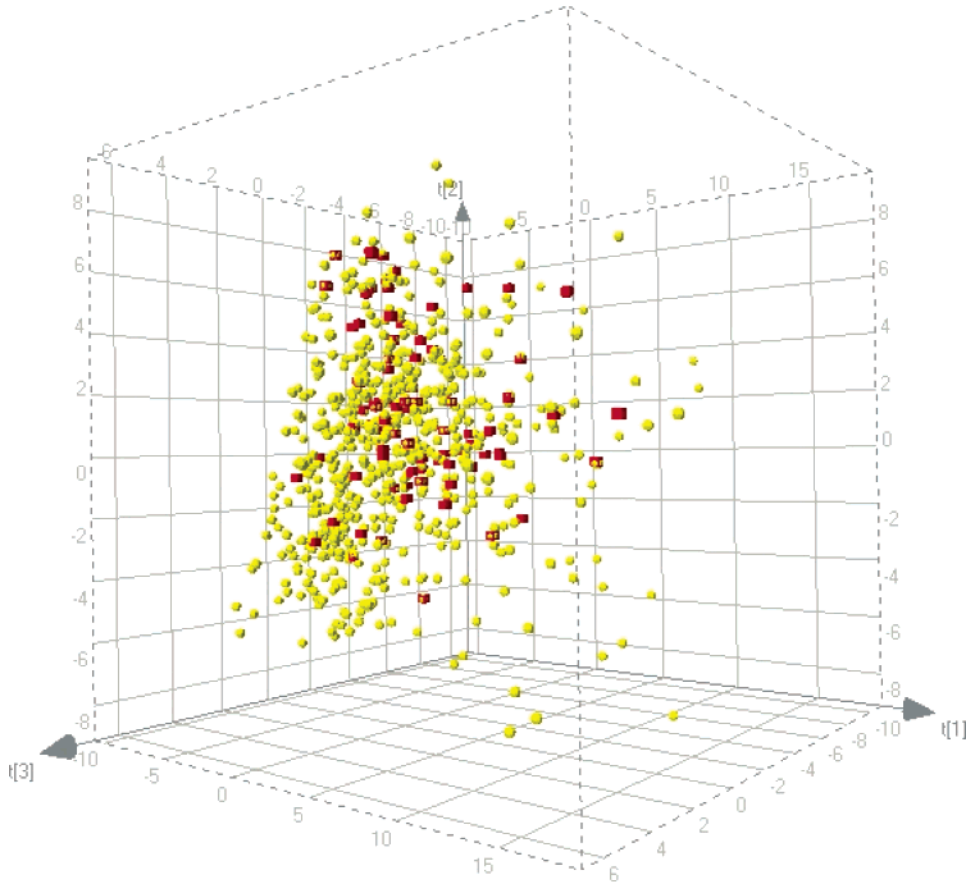

Figure 2. The dataset of 85 compounds, shown as red boxes projected on the three most important dimensions (t1, t2, and t3) of the ChemGPS training set. The ChemGPS satellites and core structures are shown as yellow spheres.

area of a certain type of atom.[13,14] For example, the NPSA originating from carbon atoms can be partitioned into the surface areas of sp-, sp^2-, and sp^3-hybridized carbon atoms and the hydrogen atoms bound to these carbon atoms. Both the absolute surface areas and the surface areas relative to the total surface area were calculated.

Data Analysis. The solubility values of the different datasets studied were predicted by partial least-squares projection to latent structures (PLS)[26] using Simca-P version 8.0.[27] The global dataset was divided into a training set and a test set of 56 and 29 compounds, respectively, according to the results from the PCA. The training set was selected to cover a maximum range in variable space described by the calculated descriptors. To study if local models were of better accuracy than the global model, the 85 drugs were divided into acids, bases, ampholytes, and nonproteolytes.[28] Since these datasets were rather small, we chose to include our compounds in the training sets used to develop the subset specific models and validated the models with external test sets taken from the literature.[4,7,29]

The response and the descriptors were preprocessed before the multivariate data analysis. The solubility values were log-transformed prior to the model development. All the calculated descriptors were mean centered and scaled to unit variance. The cube root of highly skewed descriptors was taken before they were subjected to further analysis. Variables that did not obtain a skewness within ± 1.5 after transformation were excluded from further data analysis, to avoid obtaining too heavy weighting in the models. The numbers of PLS components computed were assessed by Q^2, the "leave-many-out" cross-validated R^2, using four cross-validation rounds. Only PLS components resulting in a positive Q^2 were computed. The models were refined through stepwise selection of the variables. If the exclusion of the least important variable resulted in a more predictive model (higher Q^2), then that descriptor was permanently left out of the model. The variable selection procedure was repeated until no further improvement of the model was achieved. This selection resulted in ≤9 descriptors included in the final global models. To not overfit the models, the number of principal components was kept low (≤ 3) for the final models. Moreover, the models were validated by permutation test using scrambled solubility values to ensure that the models were not obtained by chance. The accuracy of the obtained global model was evaluated by the root-mean-square error of the test set ($RMSE_{te}$) comprised of 29 of the 85 compounds. Moreover, the global and local models were evaluated by the root-mean-square error of the external test set ($RMSE_{ext\ te}$) compiled from the publications by Huuskonen et al.[4] and Jorgensen and Duffy.[7]

Three models were generated for each dataset treated: one model based on 2D descriptors, one model based on 3D descriptors, and one model based on a combination of both types of descriptors. These three models were subjected to consensus modeling, where the average value of the three predicted values was used for solubility prediction.

There are several reasons why outliers of the models are obtained. First, the diversity of structural fragments included in the model development is important. If we use the obtained model to predict a structural fragment that has not been

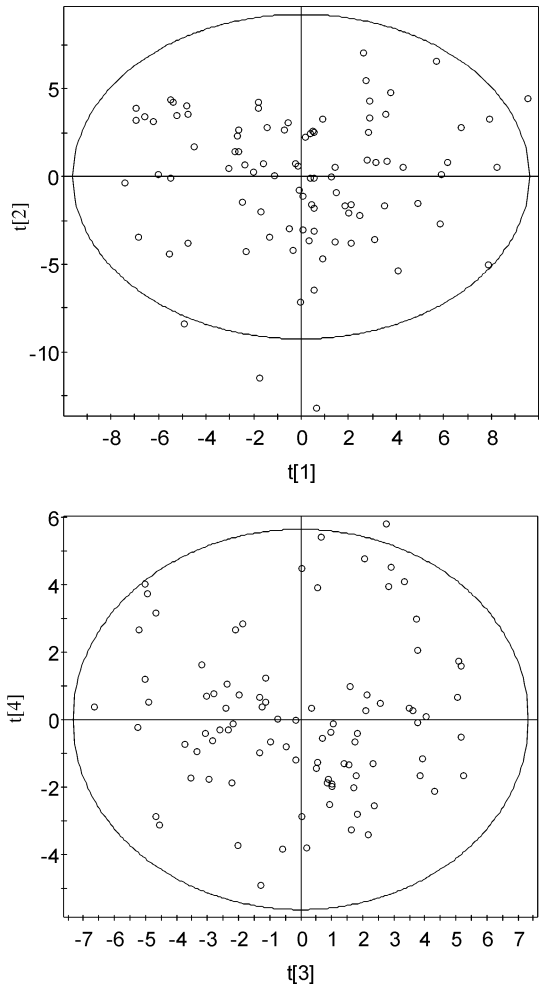

Figure 3. Score plot of the PCA performed on all calculated, nonskewed descriptors obtained for the 85 compounds included in the training and test set. The first four principal components (t1−t4) describe 59% of the x-space.

included in the training of the model, it is unlikely that the prediction will be successful. Second, the experimental setting is strongly influencing the accuracy of the prediction. It is, for instance, common to determine poorly soluble compounds with cosolvent methods and extrapolate to the aqueous solubilities. This approach may result in uncertain solubility values. Third, the validity of a model is within the range of responses used in the model development. Thus, predictions outside the solubility range of the training set, which for the global model was −1.2−−8.8 log units, should be performed with caution. To better predict compounds that are either extremely soluble or insoluble, a training set with a larger fraction of compounds with these solubilities should be used to reveal specific characteristics important for such molecules. The predictions obtained in this work are therefore presented after exclusion of compounds displaying solubility values outside the solubility range set by the training set. However, the data for all predicted compounds are given within parentheses in Tables 1−4, since these give a measure of the accuracy of the models if applied in early drug discovery.

RESULTS AND DISCUSSION

Datasets. The quality of the solubility data used for modeling was a high priority since large variability in solubility values has been shown.[5,11,12] The experimental variability can be dependent on the different end-points used, for example, the use of the solubility equilibrium to determine the thermodynamic solubility,[13,30] or the use of a value obtained at precipitation from a solution as in the case of

Table 1. Statistics of the Global Models

model	R^2	Q^2	$RMSE_{tr}$	R^2_{te}	$RMSE_{te}$[a]	$R^2_{ext\ te}$	$RMSE_{ext\ te}$[b]
global 2D[c]	0.75	0.68	0.92	0.62	0.86 (1.00)	0.56	0.80 (1.01)***
global 3D[d]	0.57	0.53	1.20 (1.03)[e]	0.67	0.94 (1.04)	0.52	0.89 (1.06)**
global comb[f]	0.78	0.71	0.86	0.54	1.04 (1.10)	0.54	0.93 (1.07)*
global cons[g]	0.80		0.90	0.71	0.83 (0.93)	0.59	0.82 (0.95)*

[a] RMSE-values for the test set. RMSE-values including the values for compounds with solubility data outside the solubility range covered by the training set (−1.2 to −8.8 log units) are given in parentheses. [b] RMSE-values for the external test set. The RMSE-values including compounds with solubility data outside the solubility range covered by the training set (−1.2 to −8.8 log units) are given within the parentheses. [c] Solubility model generated from descriptors obtained from the 2D representation of the drug-like compounds. [d] Solubility model generated from descriptors obtained from the 3D representation of the drug-like compounds. [e] The $RMSE_{tr}$ within parentheses is the RMSE-value calculated after the exclusion of SKF105657, which was identified as a large outlier falsely predicted by four log units from the surface area descriptors. [f] Solubility model generated from descriptors obtained from both the 2D and the 3D representation of the drug-like compounds. [g] Average consensus model of models c, d, and f. Statistically significant differences between the RMSE-values before and after exclusion of solubility data outside the solubility range covered by the training set are denoted with asterisks; $p < 0.05$ = *, $p < 0.01$ = **, and $p < 0.001$ = ***.

Table 2. Statistics of Subset Specific Models Obtained from 2D Descriptors[a]

model	R^2	Q^2	$RMSE_{tr}$	$R^2_{ext\ te}$	$RMSE_{ext\ te}$
acids	0.59	0.55	1.14	0.43	1.11 (1.27)
ampholytes	0.80	0.77	0.60	0.28	1.24 (1.14)
bases	0.80	0.78	0.82	0.55	1.06 (1.21)*

[a] Multivariate data analysis of the subsets. The RMSE-values for the external test set including compounds with solubility data outside the solubility range set by the training sets are given in parentheses. A statistically significant difference between RMSE values was obtained for the model of bases (* = $p < 0.05$).

Table 3. Statistics of the Subset Specific Models Generated from 3D Descriptors[a]

model	R^2	Q^2	$RMSE_{tr}$	$R^2_{ext\ te}$	$RMSE_{ext\ te}$
acids	0.55	0.42	1.20	0.29	1.25 (1.39)
ampholytes	0.53	0.37	0.91	0.33	0.86 (1.17)
bases	0.83	0.79	0.75	0.73	0.89 (0.98)

[a] Multivariate data analysis of the subsets. The RMSE-values for the external test including compounds with solubility data outside the solubility range set by the training sets are given in parentheses.

Table 4. Statistics of Models Obtained from Combinations of 2D and 3D Descriptors[a]

model	R^2	Q^2	$RMSE_{tr}$	$R^2_{ext\ te}$	$RMSE_{ext\ te}$
acids	0.62	0.58	1.11	0.43	1.32 (1.47)
ampholytes	0.83	0.74	0.55	0.27	1.28 (1.20)
bases	0.85	0.79	0.71	0.67	0.77 (1.03)*

[a] Multivariate data analysis of the subsets. The RMSE-values for the external test set including compounds with solubility data outside the solubility range set by the training sets are given in parentheses. A statistically significant difference between these RMSE values was obtained for the model of bases (* = $p < 0.05$).

turbidimetric solubility measurements.[31] Moreover, the use of different solvents (e.g., water), buffer systems, or the use of cosolvents, results in variability in the solubility values obtained, as do variations in the pH used for the solubility determination.[32] Finally, temperature variations can result in different solubility values, even though smaller variations in temperature (± 2 °C) generally do not affect the solubility largely.[30] Thus, even though we collected solubility data for a larger number of compounds, 85 compounds remained when the following criteria were applied: (i) the solubility data should be experimentally determined at a pH resulting in the measurement of intrinsic aqueous solubility values, (ii) the solubility data should be obtained at equilibrium, (iii) the solubility values should be determined at room temperature, and (iv) the compounds should be drug-like and structurally diverse. Despite these precautions, there is probably still some experimental variability in our solubility matrix. The reasons for variability in the solubility data can be related to the fact that the solubility has been determined in different laboratories by two techniques (i.e., shake-flask and potentiometric titration) and that the experimental settings have been different with regard to, for example, exact temperature (22 °C versus 25 °C) and analysis equipment (potentiometry, HPLC−UV, HPLC−FL). Our in-house correlations between solubility values obtained from the small-scale shake flask method and the potentiometric method resulted in an R^2 of 0.95 (data not shown). Whereas, correlations between the small-scale shake flask values and the shake flask data from the literature resulted in an R^2 of 0.98.[13] Therefore, our dataset will probably show a relation of R^2 0.9 rather than 1.0, which should be related to the R^2 and Q^2-values obtained from the multivariate data analysis. Thus, if our generated models were to obtain higher R^2 and Q^2 than 0.9, there is an obvious risk that the models have been over-fitted and that experimental variability has been modeled.

The ChemGPS analysis revealed that we had selected a diverse dataset, with the 85 drugs being scattered in the total drug-like space defined by the ChemGPS core and satellite structures (Figure 2).[20] In addition, principal component analysis (PCA) also showed that the compounds displayed a wide range in physicochemical properties (Figure 3). To our knowledge, this is the first time aqueous drug solubility models are derived from a dataset that covers such a large volume of the drug-like space. A distance to model analysis identified amoxicillin as an intermediate outlier in terms of its calculated 2D and 3D descriptors. However, in the partial least-squares projection to latent structures (PLS) analysis, amoxicillin was included in the training set, since it was not regarded as such a strong outlier that it should be excluded from the multivariate data analysis.

In this study both global and subset specific models of aqueous drug solubility were generated. This was done for two reasons. First, we wanted to investigate if the surface area descriptors that we have forwarded as alternative descriptors of drug solubility[13,14] and descriptors of electrotopology, which previously have been reported as solubility predictors,[1−4,6,9,10] are generally applicable or would form better or worse models for datasets with an overrepresentation of certain functional groups. Second, we wanted to study to what extent local models are of higher accuracy than global models. The 85 drugs were therefore divided into acids, bases, ampholytes, and nonproteolytes.[28]

The dataset used as an ad hoc test to challenge the developed models has been repeatedly used in modeling of aqueous drug solubility,[4,7,33−35] illustrating the lack of published solubility datasets suitable for modeling drug solubility. The clear drawback of this dataset is that it contains large homologous series of compounds. For instance, barbiturates and steroids comprised 17% and 14%, respectively, of this validation set. Moreover, smaller homologous series of reversed transcriptase inhibitors and xanthine analogues were also included in the external test set. Hence, the ad hoc test was biased to certain therapeutic areas and functional groups. Further analysis identified that this literature dataset was clustered in one-quarter of the PCA plot defined by the 2D and 3D descriptors obtained for the 85 compounds in the training and test sets, and covered only a limited volume of the drug-like space (Figure 4a). Moreover, the validation set was not well described physicochemically by the training set (Figure 4b). Despite these drawbacks, we used this dataset to allow a realistic comparison between our models and those published in the literature.

Global Models. Global models were developed from three different matrices: a matrix with 2D descriptors, a matrix with 3D descriptors and a matrix containing both 2D and 3D descriptors. The 2D descriptors used are related to electrotopology, flexibility, hydrophobicity, hydrophilicity, and charge, whereas the 3D descriptors are measures of nonpolarity, polarity, and size. The results from the three matrices were also used for average consensus modeling (see the Supporting Information).

2D Descriptors as Solubility Predictors of a Heterologous Dataset. The nonskewed calculated 2D descriptors obtained from Molconn-Z and Selma ($n = 33$) were included in the PLS analysis and a variable selection was performed. The resulting model for the 85 compounds had a good predictive power, with values of R^2 and Q^2 of 0.75 and 0.68, respectively (Figure 5a and Table 1). The variable influence on the projection (VIP) plot (Figure 6a) showed that ClogP was the most important descriptor followed by a BCUT eigenvalue[24] related to atomic weights and connectivity, and ranking numbers based on the Lipinski rule-of-five.[31] Other descriptors included in this model were (in order of importance): the flexibility of the second longest aliphatic chain, the Kier connectivity indices of three unbranched bonds, the energy of electrons of the π-shell, and the number of negatively ionizable groups.[36] Thus, the selected descriptors describe general properties such as hydrophobicity, hydrophilicity, flexibility, electron distribution and charge. The RMSEs obtained for the training set and the two test sets were of the same accuracy (0.92 to 1.01 log units for the training and external test sets, respectively).

3D Descriptors as Solubility Predictors of a Heterologous Dataset. The obtained PLS model for the 85 compounds based on surface area descriptors also included the lipophilicity descriptor ClogP, since this descriptor had proven important in earlier studies.[13] The obtained model had a lower predictive power than the model obtained from 2D descriptors (R^2 and Q^2 of 0.57 and 0.53, respectively; see also Figure 5b and Table 1). The VIP plot showed that ClogP was the most important descriptor, followed by hydrophobic surface

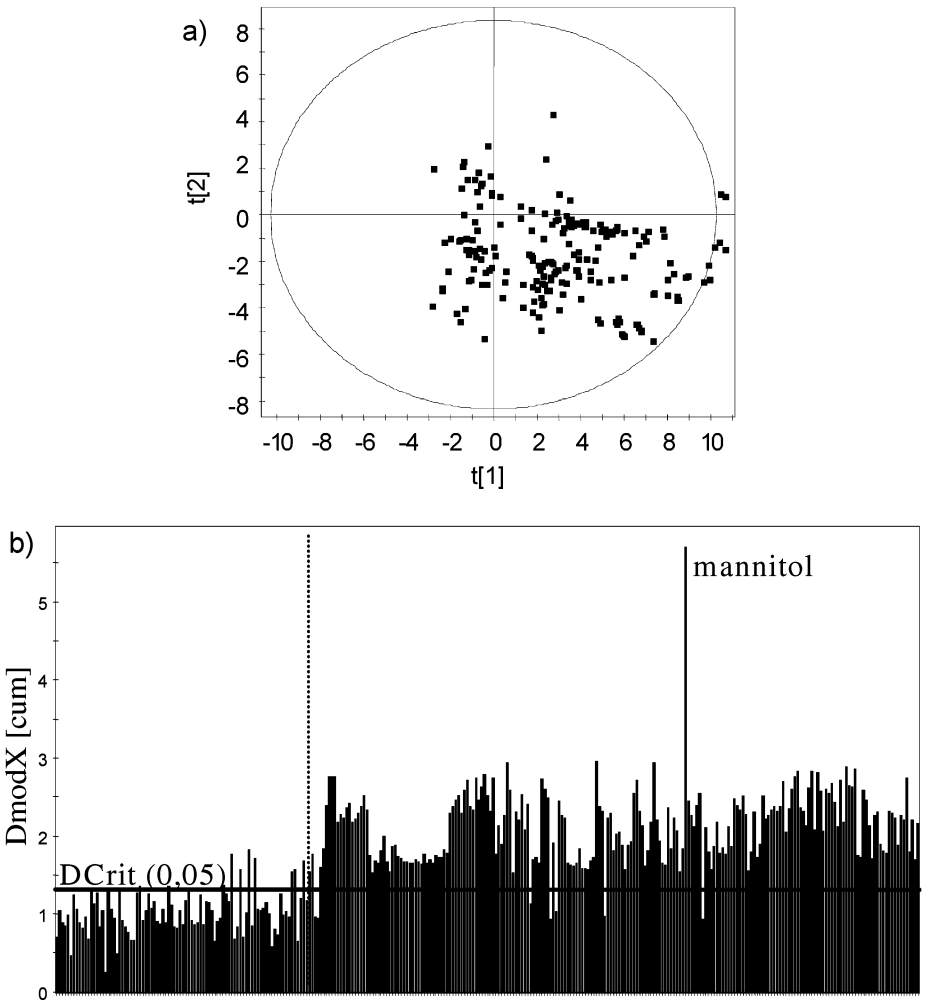

Figure 4. (a) Distribution of the external test set within the score plot of the PCA performed on all calculated, nonskewed descriptors obtained from the 85 compounds. The ellipse shows the 95% confidence interval (CI) for the first two principal components (t1−t2) extracted for the training set, which describe 30% of the x-space. (b) The distance to the model in x-space obtained from the training set used to build the global models. D Crit 0.05 marks 95% CI of the confidence interval of the obtained PCA model. Compounds on the left-hand side of the dashed line are the 85 compounds of the training and test sets, and the compounds on the right-hand side are the compounds that comprise the external test set. Mannitol was identified as a large outlier of the descriptor space set by the training set.

area descriptors (non polar surface areas (NPSA) of unsaturated atoms, the surface area of neutral hydrogen atoms, and the fraction of the surface area occupied by saturated nonpolar atoms) and a size descriptor (representing the total surface area) (Figure 6b). Least important in our model was the hydrophilic surface area descriptor for double-bonded oxygen. One large outlier was identified in the prediction: SKF105657 was predicted to be as much as four log units more soluble than the experimentally determined solubility value. As this compound was included in the PLS training set it might have influenced the model negatively; exclusion of this compound improved the R^2 and Q^2 to 0.64 and 0.58, respectively. However, neither the PCA nor the distance to model in x-space for the final model identified this compound as an outlier, therefore it was kept in the dataset. The *RMSE*s for the predictions using surface area descriptors were fairly good, despite the rather low R^2 value. Calculating the *RMSE*s after exclusion of the falsely predicted SKF105657 resulted in a $RMSE_{tr}$ and $RMSE_{te}$ of 1.03 and 1.04, respectively (Table 1). The external test set was predicted with an accuracy equal to that of our test set ($RMSE_{ext\ te} = 1.06$).

A Combination of 2D and 3D Descriptors for Prediction of a Heterologous Dataset. The solubility model based on both 2D and 3D descriptors had the highest predictive power of the global models generated, with values for R^2 and Q^2 of 0.78 and 0.71, respectively (Figure 5c and Table 1). The VIP plot showed that ClogP remained the most important descriptor for solubility, but eight more descriptors were selected, one of which was obtained from Molconn-Z calculations, four of which were calculated by Selma, and three were 3D generated descriptors (Figure 6c). BCUT eigenvalues of atomic weights together with nonpolar and polar surface area descriptors proved to be important, as did the Kier connectivity indices of five unbranched bonds. The least important descriptors in the model were the number of negatively ionizable groups, the surface area of single bonded oxygen, and the bond energy between single bonded carbon atoms. Thus, the selected descriptors were related to hydrophobicity, hydrophilicity, electrotopology, size, and charge. Even though the statistics for the training set were the best out of the three models developed, the prediction of the two test sets performed slightly worse than the predictions based

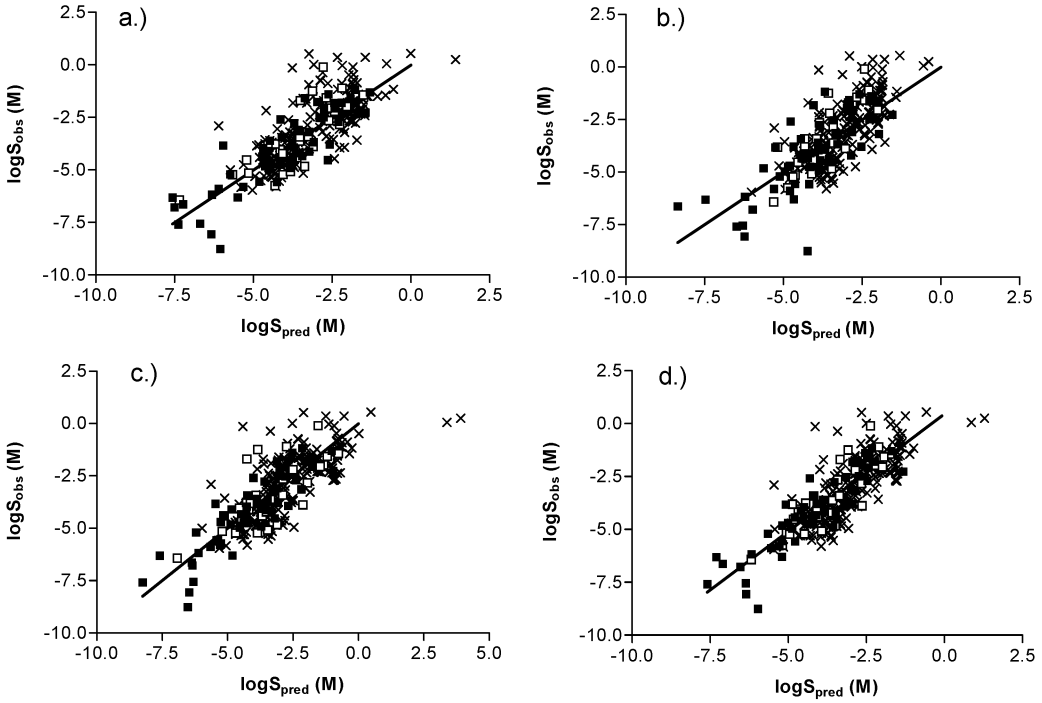

Figure 5. Global models generated on the basis of (a) 2D descriptors; (b) 3D descriptors; (c) combinations of 2D and 3D descriptors; and (d) an averaged consensus modeling approach based on models a−c. The following symbols are used: ■ = training set, □ = test set, and × = external test set.

on the models generated on either 2D or 3D descriptors (Table 1).

Consensus Modeling of a Heterologous Dataset. An average consensus approach was taken to investigate whether the global model improved when several different models were allowed to influence the final prediction. Indeed, this approach has previously been shown to improve calculations, since the combination of different techniques can give a mean value closer to the true value.[37] Thus, the predicted solubility values obtained from the three global models were averaged. This resulted in a slightly higher R^2-value and lower RMSE-values than the ones obtained with the three separate models. The accuracy of the prediction of the external test set from the consensus model is better or equal to previously published models based on structures included in the external test set. (see compilation in Huuskonen et al.,[4] Jorgensen and Duffy,[7] Wanchana et al.,[33] Stahura et al.,[34] and Wegner and Zell[35]).

In conclusion, the descriptors generated from the 2D representation of the molecules gave better solubility predictions of the global dataset than the descriptors calculated from the 3D structures. During the course of this work, a study of 809 compounds published by Cheng and Merz came to the same conclusion. Their work resulted in an aqueous solubility model with R^2 and *RMSE* values of 0.84 and 0.87, respectively, mainly based on electrotopological indices, lipophilicity, and size descriptors.[10] Only 7% of the compounds were drugs or drug-like molecules, and hence, the validity of their model for drug-like compounds remains to be shown. However, our results indicate that information gained from both types of descriptors is required to obtain the best possible prediction of this structurally diverse dataset.

Local Models. Subset-specific models were generated on the basis of the proteolytic group. Thus, both the dataset and the external test set were divided into subsets based on their proteolytic function (Figure 1).[28,29] A majority of the 85 compounds of the training and test sets was bases and a minority was nonproteolytes, a distribution in accordance to that of registered drugs.[38,39] In contrast, the largest subsets of the external test set were acids and nonproteolytes. We also observed that large homologous series were found in the external test set and, for instance, 10 of the 16 ampholytes were xanthine derivatives, 35 of the 78 acids were barbituric acids, and 31 of the 70 nonproteolytes were steroids.

PLS models were generated for bases (Figure 7), acids (Figure 8), and ampholytes (Figure 9). To perform a PLS analysis based on the eight nonproteolytes of the training set was regarded as a too complex treatment of this small dataset. Surprisingly, the nonproteolytes were predicted from the ClogP descriptor alone (Figure 10; R^2 of 0.85, $RMSE_{tr}$ and $RMSE_{ext\ te}$ of 0.75 and 0.89, respectively).[40] The subset of bases was the least diverse with regard to proteolytic groups, as the amino group (primary, secondary, and tertiary) was the proteolytic functionality for all compounds. However, with respect to lipophilicity, hydrophilicity, and size the dataset was diverse. For the subset comprised of acids, the acidic proteolytic function was represented by several different functional groups, that is, carboxylic acids, phenols, and sulfonamides. Last, the most diverse dataset was the subset with ampholytes; these compounds contained at least two ionizable groups, which were constituted of both acidic and basic functional groups.

2D Descriptors as Solubility Predictors of Bases, Acids, and Ampholytes. The 2D descriptors predicted the subsets in the following order of predictive power; bases (R^2 of 0.80, Figure 7a) = ampholytes (R^2 of 0.80, Figure 9a) > acids (R^2 of 0.59, Figure 8a and Table 2). The model for ampholytes was the only one obtained without the inclusion of ClogP as a descriptor (for VIP plots; see Supporting Information).

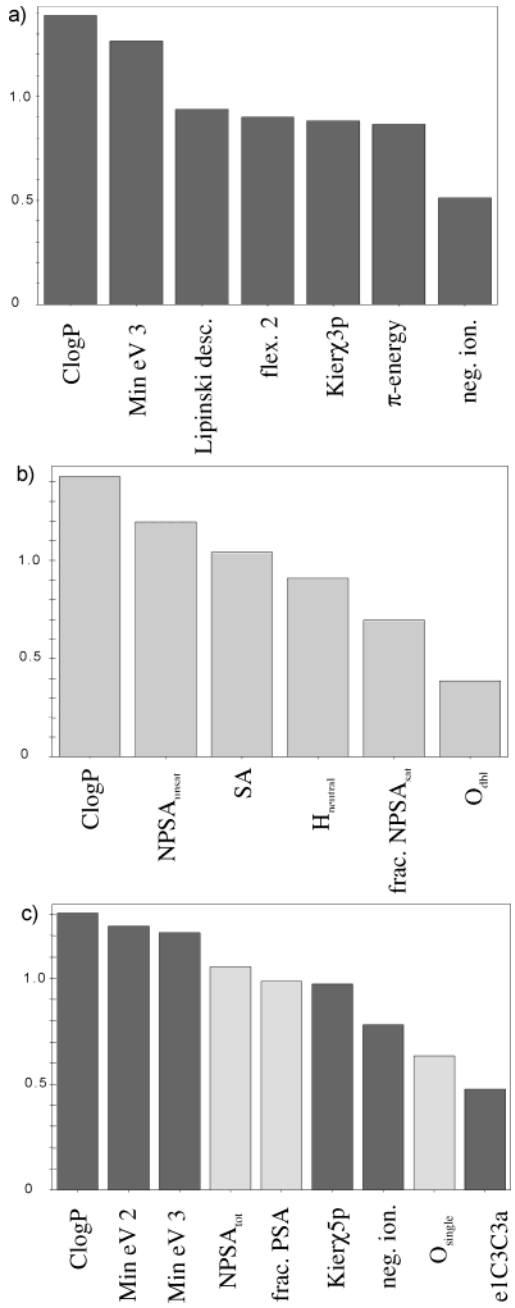

Figure 6. Variable Influence on Projection (VIP) plots of global models obtained from (a) descriptors generated from a 2D representation of the structure (dark gray bars); (b) descriptors generated from a 3D representation of the structure (light gray bars); and (c) descriptors generated from both 2D and 3D representations of the structure. The following abbreviations, written in parentheses, are used: the calculated octanol−water partition coefficient (ClogP); BCUT parameter for the second and third smallest eigenvalues of the graph adjacency matrix (Min eV 2 and Min eV 3, respectively); Lipinski rule-of-five indicator (Lipinski desc.); length of the second-longest chain containing only rotatable bonds (flex. 2); Kier connectivity indices for 3 and 5 unbranched bond lengths (Kierχ3p and Kierχ5p, respectively); the energy of electrons in the π-shell (π-energy); number of negatively ionizable groups (neg. ion.); surface area occupied by unsaturated nonpolar atoms ($NPSA_{unsat}$); the total surface area (*SA*); surface area occupied by hydrogen atoms bound to carbon atoms ($H_{neutral}$); fraction of the surface area occupied by saturated nonpolar atoms (frac. $NPSA_{sat}$); surface area occupied by double bonded oxygen atoms (O_{dbl}); surface area occupied by all nonpolar atoms ($NPSA_{tot}$); fraction of the surface area occupied by polar atoms (frac. *PSA*); surface area occupied by single bonded oxygen atoms (O_{single}); and e-state indices for single bonded carbon atoms bound to >CH− (e1C3C3a).

3D Descriptors as Solubility Predictors of Bases, Acids, and Ampholytes. The 3D descriptors were only successful in the prediction of bases (R^2 of 0.83, Figure 7b and Table 3). This model was obtained from surface area descriptors related to size and hydrophobic measures (for VIP plots; see the Supporting Information).

A Combination of 2D and 3D Descriptors for Prediction of Bases, Acids, and Ampholytes. The use of a combined matrix comprised of both 2D and 3D descriptors predicted the subsets in the following order of predictive power; bases (R^2 of 0.85, Figure 7c) > ampholytes (R^2 of 0.83, Figure 9c) > acids (R^2 of 0.62, Figure 8c and Table 4). These models did not result in large improvements in comparison to the best local model obtained from either the 2D or the 3D descriptors. Thus, combinations of descriptors of increasing complexity seem to be more useful in prediction of datasets of larger structural diversity, as identified in the modeling of the global dataset discussed above. It was found that acids and ampholytes were predominantly predicted from 2D descriptors, whereas equal numbers of 2D and 3D descriptors were used in the solubility prediction of bases. Consensus modeling did not result in large improvements of the predictive power of the subset-specific models as compared to the best local model obtained by either 2D or 3D descriptors (Figures 7d; 8d; 9d).

Comparison of 2D and 3D Models. The present study clearly showed that the investigated 2D- and 3D-generated descriptors had preferences for different subsets. Only the 2D descriptors were successful in the prediction of ampholytes. As could be expected for this group of compounds, two of the selected variables in this model represented the partial positive charge and the number of negatively ionizable groups (see the VIP plot in the Supporting Information). Such structural features are not correctly described by the 3D descriptors used in this study, which in the current setting are limited to the description of the total surface area occupied by heteroatoms rather than the total surface area of the proteolytic functions. Unfortunately, the model based on 2D descriptors obtained for ampholytes failed to predict the external test set used as an ad hoc test of the models. We noted that the subset used as the external test set was highly biased toward xanthine analogues, which comprised as much as 63% of the ad hoc test. These compounds were predicted to show approximately equal solubility. This was probably a result of the lack of variables describing the electronic environment of such compounds in the training set, which did not contain substituted xanthine derivatives (see Supporting Information).

Both 2D and 3D descriptors resulted in good prediction of the solubility of bases. A major reason for obtaining models with high predictivity of bases in comparison to the other subsets studied is probably the lower variability in the type of proteolytic group combined with the high diversity in other physicochemical properties. Moreover, the external test set used to challenge the base model did not include any large homologous series, and can thus be considered to be a fairly good external test for the generated models (Tables 2−4).

Neither of the descriptor matrices succeeded in the prediction of acids, resulting in low R^2 and high *RMSE*-values for all three developed models. This discouraging result may be an effect of the definition of acids applied in the study:

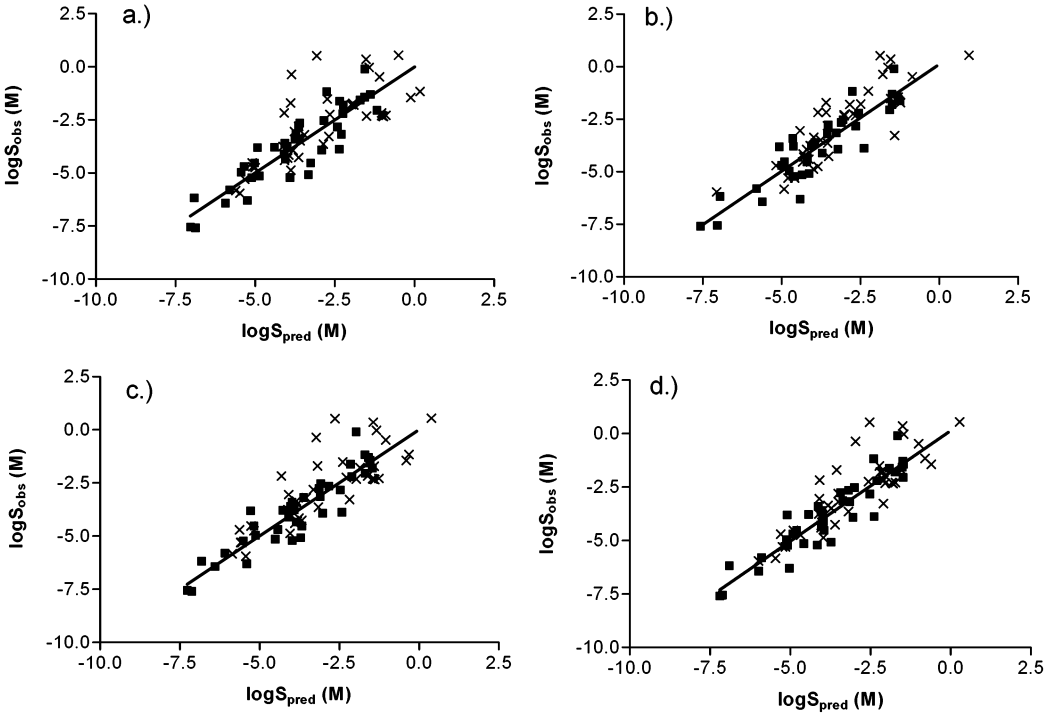

Figure 7. Models generated for bases. The models were obtained from multivariate data analysis of (a) 2D descriptors; (b) 3D descriptors; and (c) combinations of 2D and 3D descriptors. Graph (d) presents an averaged consensus modeling approach based on models a−c. The following symbols are used: ■ = training set (n = 38) and × = external test set (n = 43).

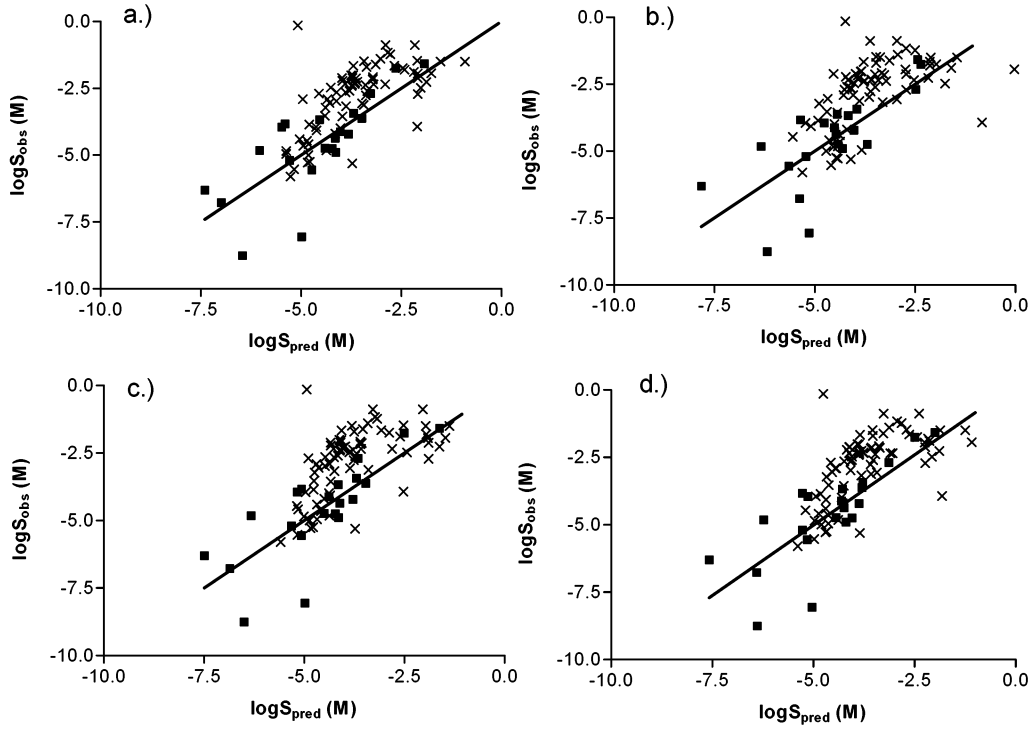

Figure 8. Models generated for acids. The models were obtained from multivariate data analysis of (a) 2D descriptors; (b) 3D descriptors; and (c) combinations of 2D and 3D descriptors. Graph (d) presents an averaged consensus modeling approach based on models a−c. The following symbols are used: ■ = training set (n = 22) and × = external test set (n = 78).

functional groups that have the ability to lose a proton within the pH-interval of 2−12 were regarded as acidic functions. In addition, barbituric acid derivatives constituted 43% of the external test set, which probably affected the result of the external validation (Tables 2−4). However, there is no reason to believe that less diverse datasets of, for example, carboxylic acids, are more difficult to predict than the bases in this study. Indeed, highly accurate models have been generated in-house for such datasets (data not shown).

Since the external test set was comprised of several homologous series, these compounds may have been better predicted by models generated on compounds with similar structural fragments. For instance, models devised for the two largest series of homologues found, that is, barbituric

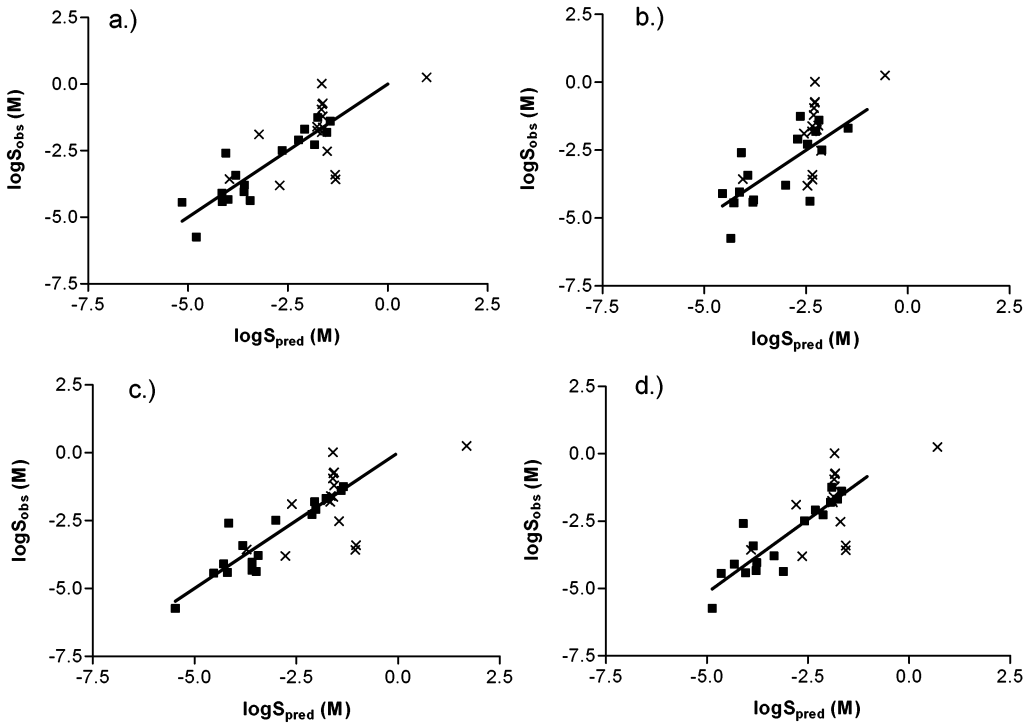

Figure 9. Models generated for ampholytes. The models were obtained from multivariate data analysis of (a) 2D descriptors; (b) 3D descriptors; and (c) combinations of 2D and 3D descriptors. The result from the averaged consensus modeling approach is presented in graph d. The following symbols are used: ■ = training set ($n = 17$) and × = external test set ($n = 16$).

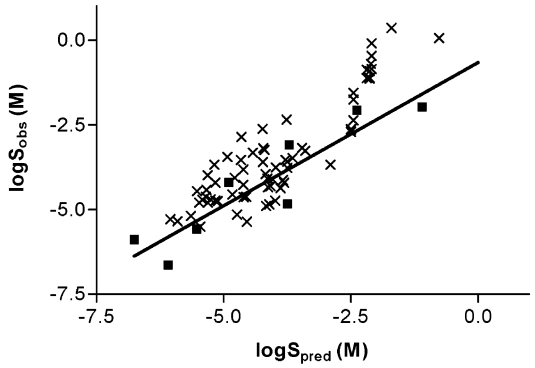

Figure 10. Solubility of nonproteolytes predicted from ClogP alone. The following symbols are used: ■ = training set ($n = 8$) and × = external test set ($n = 70$).

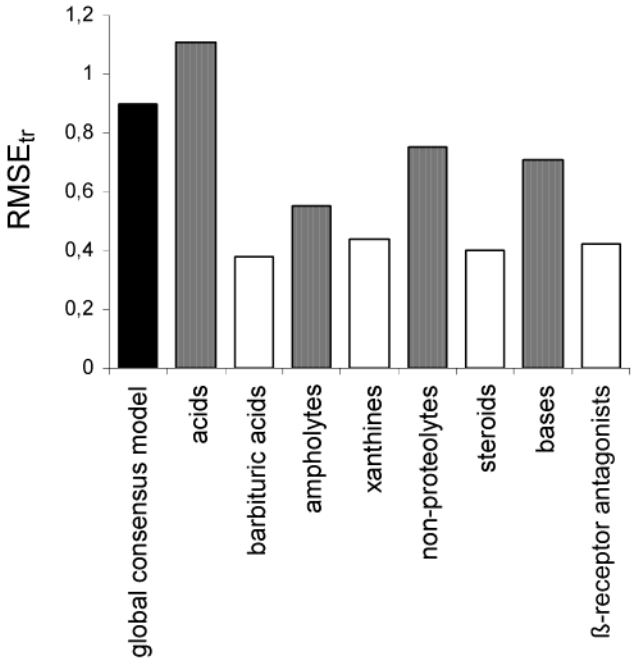

Figure 11. Compilation of the statistics of the generated models. RMSE-values for the training set are reported for the global model (black column) in comparison to the subset specific models based on a 2D and 3D combined matrix (gray columns) and the models for homologous series (white) obtained from each subset. Since no homologous series was found for the bases, β-receptor antagonists found in the training and test set were compiled with solubility data of β-receptor antagonists determined in-house to test if better predictions were obtained for a homologous series of bases than for the local model.[41]

acids and steroids, were highly successful using 2D descriptors alone, resulting in $RMSE_{barbiturates}$ and $RMSE_{steroids}$ of 0.38 and 0.40, respectively (Figure 11). The prediction of homologous series of ampholytes (xanthine derivatives) and bases (β-adrenoreceptor antagonists),[41] resulted in *RMSE* values of 0.42 and 0.44, respectively (Figure 11). These models show that solubility will be more easily predicted if models are established for homologous series occupying a limited volume of the drug-like space. This is also in agreement with a study of homologous series performed by Huuskonen et al.,[42] who obtained excellent models for three homologous series of barbituric acids, steroids, and reverse transcriptase inhibitors. Moreover, when we generated a model based on 2D descriptors for the compounds of the external test set (every third substance when listed in alphabetical order was used as a test set), $RMSE_{tr}$ and $RMSE_{te}$ of 0.73 and 0.86, respectively, were obtained. These results show that the accuracy of the solubility models as assessed by the use of test sets will be higher when structurally related compounds are included in the training set. This also explains the impressive statistics found for models based on this particular external test set.[4,33−35]

The analysis of the subset specific models revealed that none of the selected variables from either the 2D- or the 3D-generated descriptors were successful in the prediction of isomeric compounds found in the external test set, for example, 2-, 4-, 6-, or 7-hydroxypteridines. The models obtained in this study predicted that these compounds would have equal solubilities, although the experimentally determined solubility could differ more than 10-fold (see Supporting Information). To produce successful models that discriminate between closely related structures, we believe that two approaches have to be taken. First, the training set should include isomeric structures to allow variables describing the change in the electron distribution between different analogues to influence the multivariate data analysis. Second, our approach so far has been to predict solubility from the solvation effects only, since several studies have shown that these are of major importance for solubility.[1−9] For instance, the inclusion of ClogP in our models indicates the importance of solvation, since measured logP_{oct}-values have been found to predict the solubility of liquids in liquids.[43] However, the present work shows that solvation descriptors alone are insufficient to discriminate between the solubility of isomeric compounds. Hence, we believe that incorporation of solid-state characteristics will result in models that better predict the aqueous solubility of such structures.

In conclusion, we present a detailed analysis of in silico predictions of aqueous drug solubility based on a structurally diverse and drug-like dataset. The drug-like dataset used for the generation of the global model was structurally demanding and resulted in predictions with good accuracy, as assessed with two different test sets. However, the analysis of the external test set highlights the difficulties in generating truly global models for the prediction of drug solubility, since the training set should involve large numbers of structures, have considerable structural diversity, and be based on experimental solubility determinations of high quality. Our results indicate that even though 2D descriptors are better predictors of solubility than 3D descriptors, models based on information obtained from both types of descriptors will result in the best predictions. Our studies of local models for proteolytic subsets showed that descriptors generated from a 2D representation of the molecule were most suitable for the prediction of ampholytes, whereas surface area descriptors generated from a 3D structure of the molecule best predicted bases. Rather uncomplicated computational protocols could be applied to accurately predict the aqueous solubility of series of homologous compounds and specific subsets. Finally, we conclude that in order to improve the theoretical models for drug solubility further, even larger training sets should be used and descriptors for solid state characteristics and the crystal structure should be included. The rate-limiting step in this procedure will most probably be the generation of a large experimental database of highest possible quality.

ACKNOWLEDGMENT

This work was supported by grant No. 9478 from the Swedish Research Council, the Swedish Foundation for Strategic Research, the Knut and Alice Wallenberg Foundation, and GlaxoSmith Kline, P.A. We are grateful to Dr. Alex Avdeef, pION inc., MA, Dr. Mikael Bisrath, Pharmacia, Stockholm, and Dr. Bertil Abrahamsson, AstraZeneca R&D Mölndal for releasing solubility data for the model development.

Supporting Information Available: The following material is included as Supporting Information: tables of the experimentally determined solubility data and the results of the global and the subset specific models, a table of the PLS coefficients obtained for the global models and VIP plots of the subset specific models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Nelson, T. M.; Jurs, P. C. Prediction of aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* **1993**, *34*, 601−609.
(2) Sutter, J. M.; Jurs, P. C. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.
(3) 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.
(4) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(5) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(6) Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **2000**, *35*, 1081−1088.
(7) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.
(8) 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.
(9) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A fuzzy ARTMAP based on quantitative structure−property relationships (QSPRs) for predicting aqueous solubility of organic compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177−1207.
(10) Cheng, A.; Merz, K. M. Prediction of aqueous solubility of a diverse set of compounds using quantitative structure−property relationships. *J. Med. Chem.* **2003**, *46*, 3572−3580.
(11) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. Aquafac 3: Aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(12) AquaSol database compiled by S. H. Yalkowsky. For further information: www.pharm.arizona.edu/aquasol/index.html.
(13) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for prediction of aqueous drug solubility. *Pharm. Res.* **2002**, *19*, 182−188.
(14) Bergström, C. A. S.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption classification of oral drugs based on molecular surface properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(15) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177−1185.
(16) Solubility data were obtained from Dr Bertil Abrahamsson, AstraZeneca R&D Mölndal, Dr Michael Bisrath, Pharmacia, Sweden and Dr Alex Avdeef, pION Inc., MA.
(17) Avdeef, A. pH-metric solubility. 1. Solubility-pH profiles from Bjerrum plots. Gibbs buffer and pK_a in the solid state. *Pharm. Pharmacol. Commun.* **1998**, *4*, 165−178.
(18) Avdeef, A.; Berger, C. M.; Brownell, C. pH-Metric Solubility. 2: Correlation between the acid−base titration and the saturation shake-flask solubility-pH Methods. *Pharm. Res.* **2000**, *17*, 85−89.
(19) Avdeef, A.; Berger, C. M. pH-metric solubility. 3. Dissolution titration template method for solubility determination. *Eur. J. Pharm. Sci.* **2001**, *14*, 281−291.
(20) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3*, 157−166.
(21) *Molconn-Z v. 3 15S*; Hall Associates Consulting: Quincy, MA.
(22) Selma is an AstraZeneca in-house software package. For further information contact Olsson, T.; Sherbukhin, V. Synthesis and Structure Administration (SaSA), AstraZeneca R&D Mölndal.

(23) Kier, L. B.; Hall, L. H. *Connectivity in Structure−Activity Analysis*; Research Studies Press: John Wiley and Sons: Letchwort, 1986.
(24) Pearlman, R. S.; Smith, K. M. *New Software Tools for Chemical Diversity*; Kluwer Academic: Dordrecht, 1997.
(25) *MAREA*, version 2.4; The program MAREA is available upon request from the authors. The program is provided free of charge for academic users. Contact Johan Gråsjö (e-mail johan.grasjo@farmaci.uu.se).
(26) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211−228.
(27) Jackson, E. J. *A User's Guide to Principal Components*; Wiley: New York, 1991.
(28) The dataset was divided into the following subsets, which were used in the local models: acids ($n = 22$), ampholytes ($n = 17$), bases ($n = 38$), and nonproteolytes ($n = 8$).
(29) The external test set taken from the literature was divided into the following subsets, which were used to challenge the obtained local models: acids ($n = 78$), ampholytes ($n = 16$), bases ($n = 43$), and nonproteolytes ($n = 70$).
(30) Yalkowsky, S. H.; Banerjee, S. *Aqueous solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker Inc.: New York, 1992.
(31) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(32) Hasselbalch, K. A. Die Berechnung der Wasserstoffzahl des Blutes aus der freien und gebunden Kohlensäure desselben, und die Sauerstoffbindung des Blutes als Funktion der Wasserstoffzahl. *Die Biochem. Z* **1916**, *78*, 112−144.
(33) Wanchana, S.; Yamashita, F.; Hashida, M. Quantitative structure/property relationship analysis on aqueous solubility using genetic algorithm-combined partial least-squares method. *Pharmazie* **2002**, *57*, 127−129.
(34) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550−558.
(35) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(36) Kier, L. B.; Hall, L. H. An electrotopological-state index for atoms in molecules. *Pharm. Res.* **1990**, *7* (8), 801−807.
(37) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(38) Avdeef, A. Physicochemical profiling (solubility, permeability and charge state). *Curr. Top. Med. Chem.* **2001**, *1*, 277−351.
(39) Neuhoff, S.; Ungell, A. L.; Zamora, I.; Artursson, P. pH-dependent bidirectional transport of weakly basic drugs across Caco-2 monolayers: implications for drug-drug interactions. *Pharm. Res.* **2003**, *20*, 1141−1148.
(40) This is in contrast to the investigation of steroids performed by Huuskonen et al.,[47] who found only a weak correlation between the lipophilicity and solubility of steroids. Hence, our result was somewhat unexpected as steroids constitute a large part of the nonproteolytes of the external test set.
(41) No homologous series of bases was found within the external test set to generate a model for a series of homologous bases. Hence, the solubility values of β-receptor antagonists found in the training and test sets combined with shake flask solubility values of such analogues determined in our laboratory were used in the development of the homologous model of a series of bases. Except for the 8 compounds in the training and test sets the following compounds and solubility values (given as logS (M)) were used: bunitrolol −1.57, carvedilol −6.14, celiprolol −1.94, oxprenolol −2.44, pafenolol −2.78, timolol −1.45.
(42) Huuskonen, J.; Salo, M.; Taskinen, J. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharm. Sci.* **1997**, *86* (4), 450−454.
(43) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.

CI049909H