Experimental and Computational Screening Models for Prediction of Aqueous Drug Solubility

Christel A. S. Bergström,[1] Ulf Norinder,[2] Kristina Luthman,[3] and Per Artursson[1,4]

Received July 2, 2001; accepted November 7, 2001

Purpose. To devise experimental and computational models to predict aqueous drug solubility.

Methods. A simple and reliable modification of the shake flask method to a small-scale format was devised, and the intrinsic solubilities of 17 structurally diverse drugs were determined. The experimental solubility data were used to investigate the accuracy of commonly used theoretical and semiexperimental models for prediction of aqueous drug solubility. Computational models for prediction of intrinsic solubility, based on lipophilicity and molecular surface areas, were developed.

Results. The intrinsic solubilities ranged from 0.7 ng/mL to 6.0 mg/mL, covering a range of almost seven $\log_{10}$ units, and the values determined with the new small-scale shake flask method agreed well with published solubility data. Solubility data computed with established theoretical models agreed poorly with the experimentally determined solubilities, but the correlations improved when experimentally determined melting points were included in the models. A new, fast computational model based on lipophilicity and partitioned molecular surface areas, which predicted intrinsic drug solubility with a good accuracy (R^2 of 0.91 and $RMSE_{tr}$ of 0.61) was devised.

Conclusions. A small-scale shake flask method for determination of intrinsic drug solubility was developed, and a promising alternative computational model for the theoretical prediction of aqueous drug solubility was proposed.

KEY WORDS: shake flask method; drug solubility; molecular surface area; solubility prediction.

INTRODUCTION

Combinatorial chemistry and high-throughput screening techniques have increased the number of candidate drugs produced annually. Unfortunately, these techniques do not generally provide compounds with optimal biopharmaceutical and pharmacokinetic properties. The drug candidates are often poorly soluble in water, which results in low drug concentrations in the gastrointestinal fluids and, hence, unacceptably low drug absorption (1). Therefore, studies of drug solubility early in the drug development process are motivated. Indeed, several experimental methods (1–3) and computational models (4,5) of varying accuracy and complexity have been developed for the prediction of aqueous drug solubility.

The most reliable and commonly used experimental method for determining intrinsic aqueous drug solubility is the shake flask method (6,7). However, this method is time-consuming and a single solubility experiment can be ongoing for several days to weeks, which limits its usefulness (6,8). Moreover, accurate determination of lipophilic, insoluble substances may be troublesome because of loss of substance in the filtration step (6). In addition, the experiments are traditionally performed on a large scale, and large amounts (grams) of substances are required. However, preliminary, solubility studies in a microtiter plate format have been reported (9).

A more rapid, but less reliable, alternative to the shake flask method is based on precipitation of the drug after serial dilution of DMSO stock solution (1). This turbidimetric method is used as a screening tool in the drug discovery process, because the compounds generated by combinatorial chemistry are stored in DMSO stock solutions. It is rapid at the expense of accuracy, and, because the solubility is determined from DMSO solutions, no consideration is taken of the influence of the solid state. Rapid and reliable methods to determine drug solubility on a small-scale basis are, therefore, warranted.

Recently, a variety of computational models have been devised to predict the aqueous solubilities of homologous and heterologous series of compounds. Simple linear regression models using physicochemical descriptors for size, lipophilicity, and/or hydrogen-bonding capacity (5,10–12) as well as more complex neural network models based on, *e.g.,* electrotopologic descriptors (13,14) have resulted in fairly good predictions. Unfortunately, the data sets used to build the computational models generally contain only a small number of druglike molecules. Therefore, the usefulness of these models in drug discovery remains to be shown.

Uncertainties in the experimental data used may contribute to large errors in the solubility predictions made by computational models (11,15). We have generated our own reliable solubility data and used these data to predict aqueous drug solubility from molecular descriptors. We refined the shake flask method for rapid and reliable solubility determinations. The filtration step was replaced by an ultracentrifugation step to minimize loss of substance and sample volumes. Furthermore, the suspension volumes were reduced to limit the consumption of compounds to the microgram range and to be readily adaptable to a microtiter plate format. We used the solubility data obtained to evaluate available theoretical and semiexperimental models for prediction of aqueous drug solubility. Finally, we developed a new computational model that is mainly based on lipophilicity and non-polar surface area descriptors.

MATERIALS AND METHODS

Selection of Drugs

Seventeen structurally diverse compounds that were chosen so that a number of physicochemical properties, such as size, lipophilicity, and melting point, covered a large range were studied (Fig. 1) (Table I). Both protolytes and nonprotolytes were investigated. The approximate aqueous solubilities of these compounds were predicted on the basis of their octanol-water partition coefficient and molecular

[1] Department of Pharmacy, Uppsala University, Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden.
[2] Department of Medicinal Chemistry, AstraZeneca Research and Development, SE-151 85 Södertälje, Sweden.
[3] Medicinal Chemistry, Department of Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden.
[4] To whom correspondence should be addressed. (e-mail: per.artursson@farmaci.uu.se)

0724-8741/02/0200-0182/0

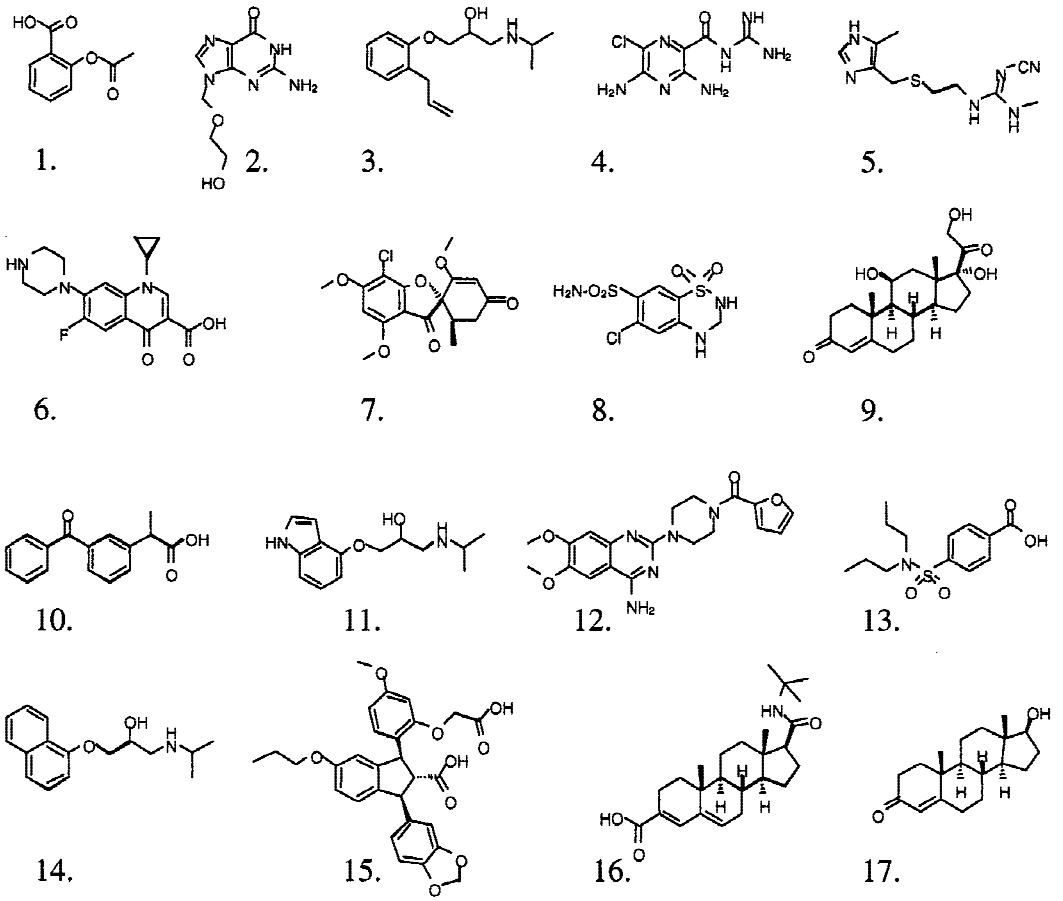

Fig. 1. Chemical structures of the compounds studied. 1. acetylsalicylic acid, 2. acyclovir, 3. alprenolol, 4. amiloride, 5. cimetidine, 6. ciprofloxacin, 7. griseofulvin, 8. hydrochlorothiazide, 9. hydrocortisone, 10. ketoprofen, 11. pindolol, 12. prazosin, 13. probenecid, 14. propranolol, 15. SB209670, 16. SKF105657, 17. testosterone.

weights (10). The range of the compounds predicted as the least soluble and the most soluble in water covered more than seven $\log_{10}$ units. Acyclovir, ciprofloxacin, SB209670, and SKF105657 were gifts from SmithKline Beecham (Philadelphia, PA). All other compounds were purchased from Sigma (St. Louis, MO). Alprenolol, amiloride, prazosin, and propranolol were used as the corresponding HCl salts.

Solid-State Characterization

The crystallinity, purity, water content, and melting point of each compound were determined by differential scanning calorimetry (DSC) using a Mettler DSC 20 TC10A/15 (Switzerland). Insufficient amounts of SB209670 were available for DSC analysis. None of the compounds underwent glass transitions or crystal transitions during the DSC experiment, which confirmed that all compounds were crystalline. Thus, the solubility was determined of the most stable form of the solid state. Amiloride was an exception: amiloride dihydrate was used, and the crystal water may have influenced the solubility value.

Solubility Studies by Shake Flask

Each drug was added in excess to 1000, 500, 200, 100, or 50 μL Milli-Q water in a test tube at room temperature (22.5 ± 1°C), and the test tubes were placed on a plate shaker, which agitated the suspensions at 300 rpm. Compounds used as HCl salts were studied with different amounts of excess solid present at the establishment of the equilibrium, to evaluate if the solubility values determined were affected by the salt. The pH of each drug suspension was adjusted by using 1 M HCl or 1 M NaOH to a value at least 1 pH unit below (acids) or above (bases) the pK_a value of the drug. This allowed the solubilities of uncharged compounds to be determined. The pH values of ampholyte suspensions were adjusted to the isoelectric point of the compound to determine the solubilities of the zwitterionic species. Samples were withdrawn after 24, 48, and 72 h for drugs that were expected to dissolve rapidly and at 24, 72, and 144 h for drugs that were expected to dissolve slowly. The samples were centrifuged in an Eppendorf centrifuge model 5043 at 23,000g for 15 min to separate the solid material from the solution. The temperature of the samples remained at 22.5 ± 1°C during the centrifugation. The supernatants were stored at −20°C pending analysis with reversed phase HPLC. Acetylsalicylic acid was hydrolyzed after 24 h at the pH used (pH 2), and results from time points after this could not be obtained. Therefore, only the 24 h value is reported. Two compounds, SKF105657 and SB209670, showed no detectable solubility in water with use of the small-scale shake flask (SSF) method and HPLC analysis; therefore, these compounds were studied by using metha-

Table I. Aqueous Solubilities and Physicochemical Properties of the Compounds Studied[a]

Substance	S_o(μg/mL) 24 h	S_o(μg/mL) 72 h	lit. values	Mw (g/mol)	ClogP_{oct}	#H_{tot}	PSA (Å^2)	NPSA (Å^2)	pK_a[b]	mp (°C)
SKF105657	0.0007 ± 0.0002[c]	n.d.[d]		399.6	5.04	9	70	414	5.4	254
SB209670	0.088 ± 0.025[c]	n.d.[d]		520.5	3.64	20	134	448	3.9, 5.6	n.d.[g]
Prazosin	3.1 ± 1.1	3.2 ± 1.4		383.4	1.50	15	106	342	7.0	285
Probenecid	3.8 ± 0.8	3.6 ± 0.8		285.4	3.37	10	75	270	3.3	198
Griseofulvin	4.6 ± 0.3	5.2 ± 0.9	6.6	352.8	1.75	12	77	316	n.a.	219
Testosterone	22 ± 6	18 ± 3	24	288.4	3.22	5	42	315	n.a.	153
Pindolol	29 ± 3	33 ± 2		248.3	1.67	9	58	282	9.6	170
Propranolol	30 ± 4	31 ± 2		259.4	2.75	7	40	316	9.5	162
Ciprofloxacin	60 ± 5	54 ± 7	86	331.4	–1.93	11	79	293	6.3, 8.5	266
Ketoprofen	85 ± 4	94 ± 1	51	254.3	2.76	7	61	260	4.0	93
Amiloride	191 ± 51	150 ± 12		229.6	–0.79	17	152	79	8.7	291, 299
Hydrocortisone	207 ± 88	294 ± 14	285	362.5	1.70	13	89	311	n.a.	223
Aprenolol	377 ± 28	367 ± 34		249.3	2.65	7	39	324	9.5	109
Hydrochlorothiazide	587 ± 8.5	595 ± 40	609	297.7	–0.40	15	132	133	7.9, 9.2	267
Acyclovir	1170 ± 86	1213 ± 66	1400	225.2	–2.30	15	128	134	2.4, 9.2	255, 265
Acetylsalicylic acid	3198 ± 228	n.d.[e]	3405	180.2	1.02	9	44	196	3.5	142
Cimetidine	4945 ± 356	6046 ± 72[f]		252.3	0.35	9	86	243	7.1	141

[a] Aqueous solubilities at 24 h and 72 h (S_o), lipophilicity (ClogP_{oct}), number of hydrogen bonds (#H_{tot}), dynamic surface areas (PSA and NPSA) and melting point (mp) were obtained as described in Materials and Methods. Literature values given for solubility were used for evaluation of the SSF method (33).

[b] For compounds that are nonprotolytes, the pK_a value is nonapplicable (n.a.).

[c] The solubility has been determined by using methanol as cosolvent as described in Materials and Methods.

[d] Not determined. The end point was set to 24 h in the cosolvent determinations.

[e] Not determined at 72 h because of time-dependent hydrolysis.

[f] A statistically significant difference was obtained between the 24 and 72 h solubility determinations ($p < 0.05$).

[g] Not determined because insufficient quantities of the substance were available to allow analysis of the compound with DSC.

nol as cosolvent. Briefly, the solubility was determined at four different concentrations of methanol in water (11.4, 17.4, 23.5, and 29.8% w/w, n = 3). After 24 h, the suspensions were centrifuged and analyzed by HPLC. The intrinsic solubilities were then determined by linear regression and extrapolation to aqueous solubility (0% w/w methanol) (16,17). The linear regression gave coefficients of determination (R^2) of 0.98 and 0.99 for SB209670 and SKF105657, respectively.

Molecular Descriptors

Lipophilicity was calculated by using the ClogP program from BioByte Corp. (Claremont, CA), and the numbers of hydrogen bond acceptors, the numbers of hydrogen bond donors, and the total number of hydrogen bonds were calculated according to Ren *et al.* (18). Monte Carlo conformational analysis (19) was performed by using the BatchMin program and the MM2 force field, as implemented in MacroModel version 5.0 (20). The conformational analysis of cimetidine was performed by using the MMFF force field instead of MM2, because the latter does not contain the necessary parameters. In this case, MacroModel version 6.5 was used. The conformational analysis was performed in simulated water with the compounds in their unionized state. Molecular surface areas were calculated from the results of the conformational analysis by using an in-house program, MAREA (21).

The composite property polar surface area (PSA) was defined as the surface area occupied by oxygen and nitrogen, and hydrogen atoms bound to these heteroatoms. The composite property non-polar surface area (NPSA) was defined as the total surface area minus the PSA. In addition, these composite descriptors were divided into the partitioned total surface area descriptors (PTSA) (22,23). Briefly, a molecule has a number of PTSAs, each one of which corresponds to a certain atom type. For example, the PSA originating from oxygen atoms can be partitioned into the surface areas of single-bonded oxygen, double-bonded oxygen, and hydrogen atoms bound to single-bonded oxygen atoms.

The surface areas were calculated both as absolute and relative numbers compared to the total surface area (*i.e.,* PSA and %PSA). Both static and dynamic surface areas were calculated. Dynamic surface areas were calculated according to a Boltzmann distribution at 22.5°C, where every low-energy conformation ($E \leq 2.5$ kcal) is weighted by its probability of existence (24,25). Static surface areas were calculated for the global minimum conformation.

Analytic Methods

Drug concentrations were analyzed by using a reversed phase HPLC system that consisted of the following components: a PerkinElmer isocratic LC pump 250, a PerkinElmer advanced LC sample processor ISS-200, and a Spectra-Physics UV100 detector. The analytical columns used were a Hichrom Partisil ODS3 (10 × 3.2 mm) guard column and a Becker Ultrasphere ODS (250 × 5.6 mm) analysis column, both with a mean particle size of 5 μm. The composition of the mobile phases were designed to allow the detection of the compound within 10 min.

Data Analysis

The solubility determinations were performed in triplicates on at least two occasions. Values are expressed as means

± 1 SD (Table I). ANOVA was used to test whether the difference between two mean values was statistically significant ($p < 0.05$). The coefficient of determination (R^2) was used to assess the goodness of fit of linear regressions when comparing methods of measuring solubility and when building predictive models using linear or multilinear regression.

Principal component analysis (PCA) (26) and partial least squares projection to latent structures (PLS) (27) were performed in Simca (28). The data set was divided into a training set and a test set of 12 and 5 compounds, respectively, by PCA. The training set was selected to cover a maximum range in descriptor space, which was achieved by selecting the extreme values from the first three components of the PCA. The test set was compiled of alprenolol, griseofulvin, ketoprofen, pindolol, and probenecid and was representative for the training set used. Correlations between the descriptors and solubilities were established by PLS. The descriptors were centered and scaled to unit variance. The numbers of PLS components computed were assessed by Q^2, the "leave-one-out" cross-validated R^2. Only PLS components resulting in a positive Q^2 were computed. The models were refined through stepwise selection of the descriptors. If the exclusion of the least important descriptor resulted in a more predictive model (higher Q^2), then that descriptor was permanently left out of the model. The descriptor selection procedure was repeated until no further improvement of the model was achieved. The predictive power of the models established was assessed by the root-mean-square error of the training set ($RMSE_{tr}$) and the test set ($RMSE_{te}$), respectively. A $RMSE_{te}$ of <1 $\log_{10}$ unit was regarded as acceptable for our small data set.

RESULTS AND DISCUSSION

We have developed an accurate small-scale method for determination of aqueous drug solubility. The method was successfully used in determinations of aqueous solubility of a set of 17 structurally diverse drugs, whose physicochemical properties and solubilities covered a wide range. We used the solubilities determined by our method to investigate the accuracies of established theoretical and semiexperimental computational models for predicting aqueous solubility, and we propose a new model for prediction of aqueous drug solubility.

Solubilities ranging from 0.7 ng/mL to 6.0 mg/mL, *i.e.,* almost 7 $\log_{10}$ units, could be determined with the new method (Table I). No effect of the HCl salt on the intrinsic solubility value was seen for the four compounds that were used as HCl salts, as assessed by different amounts of solid present at equilibrium.

The dissolving process was rapid in most cases, and the solubility values reached a plateau within 24 h. However, the kinetic behaviors of cimetidine and hydrocortisone differed from those of the other 15 substances. Cimetidine showed dissolution rate-limited kinetics and reached its solubility plateau after 48 h. The experimental values from hydrocortisone at 24 h had a much higher standard deviation than the results from 72 h; the latter results were considered more reliable (Table I). A possible explanation for this result may be that several crystal forms of hydrocortisone are taking part in the establishment of the solubility equilibrium. This could be achieved if hydrocortisone dissolves and precipitates several times in an oscillated manner, before the equilibrium is established. Thus, at early time points, several precipitated crystal forms, with different aqueous solubilities, interact with water, and the solubility value determined will have a larger variability. Further investigation of this issue was outside the scope of this article. However, the data analysis indicates that it is generally not necessary to perform solubility experiments for longer than 24 h, provided that the deviations discussed above are accepted.

The measured aqueous solubilities of eight compounds were compared with previously published solubility values from the traditional shake flask method (Fig. 2a). The agreement between the values obtained by the two methods was excellent ($R^2 = 0.98$, RMSE of 0.13 log unit). This clearly shows that the SSF method is as accurate as the traditional shake flask method.

Studies of the effect of sample volume showed that intrinsic solubility values can be determined in volumes as small as 50 μL (Fig. 2b). No statistically significant differences were observed in intrinsic solubility in the volume range 50–1000 μL and the SDs were <10%. We conclude that the SSF method can be used to determine intrinsic solubilities of solids using volumes that are suitable for the 96-well microtiter plate format. At this scale, only microgram quantities of samples are needed to determine the solubility accurately. Moreover, the microtiter plate format allows the solubility determinations to be automated, resulting in a higher throughput.

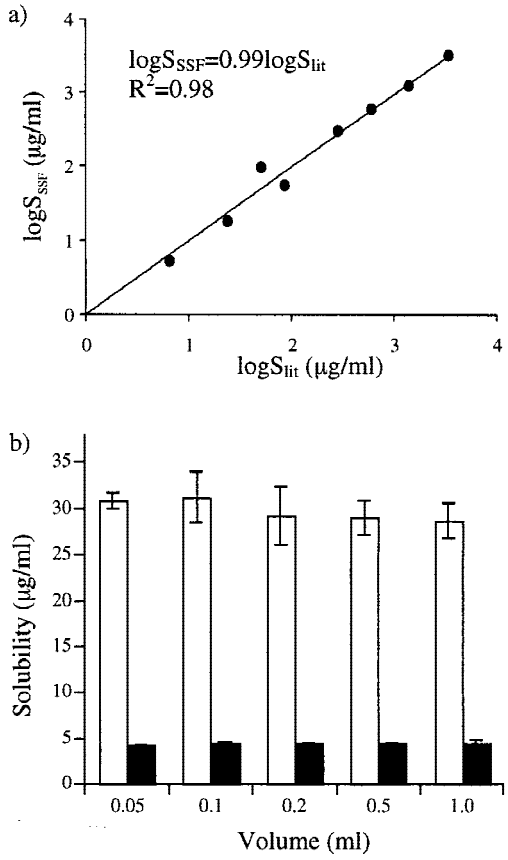

Fig. 2. Development of the small-scale shake flask method (SSF). (a) Correlation between the modified SSF method and the traditional large-scale shake flask method (33). (b) Determinations of aqueous solubility of pindolol (white columns) and probenecid (black columns) in volumes between 1000 and 50 μL.

We used our experimental results to investigate the accuracy of four simple and commonly used models for prediction of aqueous solubility (10,29,30). These models are based on simple physicochemical descriptors, and they were originally built by using data sets of compounds that are not druglike. Therefore, it was not surprising that the original models predicted the solubilities of the 17 substances poorly. For instance, using the equation derived by Yalkowsky and Valvani gave the best prediction of the four models tested. This prediction is made from lipophilicity (ClogP) and melting point (30) and resulted in a correlation with R^2 of 0.67 and RMSE of 1.27 for the 17 compounds. Therefore, we used our data set to build new models based on these descriptors. The results are summarized in Table II (models 1, 2, 6, and 7). Lipophilicity was the most important molecular property for predicting the solubility of a compound, and it was necessary to include a solid-state parameter to obtain acceptable models. This finding is in agreement with previous publications (29,30). The linear combination of ClogP and melting point now yielded a better model with R^2 of 0.85, $RMSE_{tr}$ of 0.74, and a $RMSE_{te}$ of 0.66. These results show that non-druglike compounds cannot be used to predict the aqueous solubilities of druglike molecules accurately. Furthermore, the models improved significantly if the experimentally derived solid-state parameter, the melting point, was included in the calculations. Therefore, we investigated whether an advanced purely theoretical model containing no experimentally determined properties gave better results. QikProp, a recently introduced commercial software, predicts solubility solely from theoretical descriptors calculated from the molecular structure (31). QikProp predicted aqueous solubilities for the substances in our test set poorly, with an $RMSE_{te}$ of 1.92 (Table II, model 3). In summary, these investigations show that further development of computational models for prediction of aqueous drug solubility is required.

Molecular surface descriptors were recently found to be of importance for the prediction of drug solubility from Monte Carlo simulations (5). Therefore, our initial approach was to investigate composite molecular surface area properties such as total surface area (SA), nonpolar (NPSA), and polar (PSA) surface areas as descriptors for aqueous drug solubility. A rough model was established after linear combination of the ClogP, PSA, and NPSA. When exchanging the PSA for the melting point, the model was significantly improved (Table II, models 4 and 8, respectively).

Our group recently showed that partitioned total surface area (PTSA) can provide a better predictive model for membrane permeability than composite surface areas such as PSA (22). These promising results encouraged us to study whether PTSA could be used to predict aqueous solubility. An initial input matrix was constructed, which contained all of the descriptors obtained (ClogP, PTSAs, composite surface areas, number of hydrogen bond donors, number of hydrogen bond acceptors, and total number of hydrogen bonds), and a stepwise selection of the important descriptors for solubility was performed by PLS. The best computational model was obtained from ClogP, PTSAs, and composite surface areas (Fig. 3a). This model gave R^2 of 0.91, $RMSE_{tr}$ of 0.61, and $RMSE_{te}$ of 0.90 (Table II, model 5). The model was based on two principal components, and the remaining descriptors after the variable selection are shown in Table III. The PLS analysis showed that properties that are negatively correlated with solubility, such as size, lipophilicity, and non-polar atoms, are the most important for solubility predictions of this data set. Only one hydrogen bond descriptor (surface area of hydrogen bound to nitrogen atoms) remained after the descriptor selection. This result is in contrast to previous publications that have shown that hydrogen bond descriptors improve solubility prediction models (5,11).

Static surface area can be computed rapidly, whereas dynamic surface area is computationally more demanding. We investigated the effect of using these two types of area on the accuracy of the model. The accuracy of the model was not affected when the Boltzmann distributed, dynamic surface areas were replaced with the static surface area calculated for the global minimum conformation ($R^2_{dyn} = 0.91$, $RMSE_{dyn,tr}$

Table II. Predictive Power of the Devised Models[a]

	Model	Descriptors	Method	R^2_{tr}	$RMSE_{tr}$	$RMSE_{te}$
Theoretical	1	ClogP	LR	0.53	1.36	1.24
	2	ClogP, MW	MLR	0.78	0.97	0.54
	3	# hydrogen bonds, interaction energy, size	MLR	—	—	1.92
	4	ClogP, NPSA, PSA	MLR	0.76	0.99	0.66
	5	ClogP, PTSA, composite SA	PLS	0.91	0.61	0.90
Semiexperimental	6	ClogP, mp	MLR	0.85	0.74	0.66
	7	ClogP, MW, mp	MLR	0.85	0.89	0.59
	8	ClogP, NPSA, mp	MLR	0.90	0.60	0.70
	9	ClogP, PTSA, composite SA, mp	PLS	0.91	0.55	0.80

[a] The predictive power of original models (1–3, 6, and 7) and the models developed in this work (4, 5, 8, and 9). The established models were evaluated by building new models for the data set by using the descriptors found to be of importance in the original publications. The training set (tr) consisted of 12 compounds (11 compounds in the models including melting point) and the test set (te) of 5 compounds. QikProp (model 3) has been evaluated by prediction of the test set only, because the model implemented in the QikProp software has been trained on a different data set. The methodology used in the models established were linear regression (LR), multilinear regression (MLR), and partial least square projection to latent structures (PLS). Original models are reviewed in Refs. 10, 29–31.

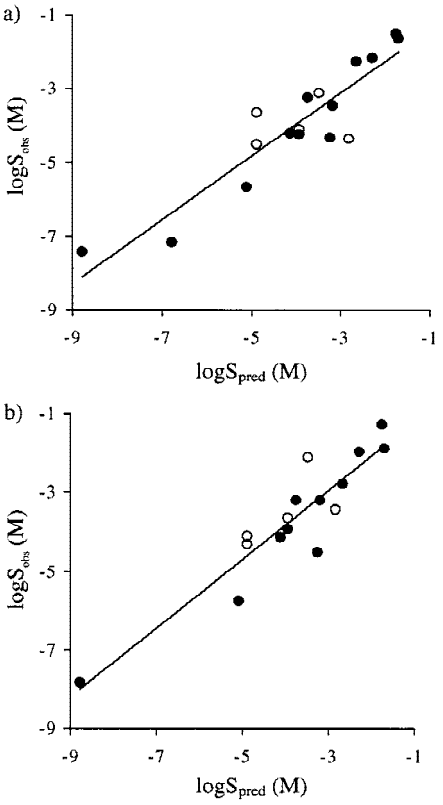

Fig. 3. Prediction of aqueous solubility of the data set using PLS methodology. The compounds were divided into a training set (•) and a test set (○). The descriptors used for prediction were (a) partitioned total surface areas, composite surface areas, and ClogP and (b) partitioned total surface areas, composite surface areas, ClogP, and melting point.

$= 0.61$ compared to $R^2_{stat} = 0.92$, $RMSE_{stat,tr} = 0.57$). Thus, calculation of dynamic surface areas was not necessary for this heterogenous data set. However, the drugs in our data set have a limited flexibility, and dynamic surface areas may be needed to provide good models for data sets of more flexible compounds (32).

The influence of the solid state on the solubility was investigated by using a semiexperimental PLS model combining the experimentally determined melting point with the theoretically calculated descriptors. A model based on three principal components was achieved (Fig. 3b), resulting in R^2 of 0.91, $RMSE_{tr}$ of 0.55 and $RMSE_{te}$ of 0.80 (Table II, model 9). The descriptors were arranged as shown in Table III. It was surprising that the inclusion of the melting point only improved the model marginally, suggesting that this descriptor is partly accommodated in the PTSAs. Preliminary computational modeling of the melting point supports this hypothesis (data not shown). Although the improvement in the prediction is small, the melting point is the second most important descriptor in this semiexperimental model, and the weighting of lipophilicity is higher than in the corresponding theoretical model (Table III). We speculate that the higher weighting of lipophilicity may compensate for the melting point that is mainly reflecting polar, hydrophilic groups. Hence, lipophilicity is needed to correct the balance between hydrophobic and hydrophilic descriptors. This finding is supported by the fact that the only molecular hydrogen bond descriptor included in both models is given less weight in the prediction when melting point is included as a descriptor. In summary, the results suggest that accurate predictive models for solubility can be developed by using rapidly calculated descriptors only. The findings in this article, that non-polar descriptors and the size of the molecule can be used to predict solubility, must be further investigated to address the issue of general application. Investigations of larger and even more structurally diverse druglike data sets will show if models based on PTSAs will be as successful as models devised from *e.g.,* electrotopologic descriptors (13).

Table III. Important Descriptors for Solubility Predictions by PLS[a]

No.	Theoretical Model	Semiexperimental
1	total SA	total SA
2	SA of H bound to N	mp
3	%neutral H of total SA	ClogP
4	saturated NPSA	saturated NPSA
5	ClogP	SA of H bound to N
6	SA of sp^3 hybridized C	SA of sp^3 hybridized C
7	%S of total SA	%neutral H of total SA
8		%S of total SA

[a] List of descriptors in order of importance (descending) for the PLS solubility prediction using a theoretical model (Table II, model 5) and a semiexperimental model (Table II, model 9).

In conclusion, we have modified the shake flask method to allow for reliable measurements of solubility in a small-scale format. Solubility data for a structurally diverse set of drugs generated with this method was used to evaluate common models used for prediction of aqueous solubility. These models were not as good in predicting the solubilities of our druglike molecules, as they were in predicting the solubilities of the original training sets. We attribute the poor performance to the fact that the models were trained on non-druglike compounds. We propose a new theoretical model for the prediction of drug solubility based on partitioned molecular surface areas and lipophilicity. This improved model shows a good accuracy for the relatively small data set in this study and does not require the use of the experimentally determined melting point. Further studies will show if the model is generally applicable to larger data sets of druglike compounds.

ACKNOWLEDGMENTS

This work was supported by Grant 9478 from The Swedish Medical Research Council, The Swedish Foundation for Strategic Research, and SmithKline Beecham. We thank Drs. Chao Pin Lee, Philip Smith, Dominic Ryan, and Harma Ellens at SmithKline Beecham for discussions and constructive criticism.

REFERENCES

1. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeny. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**:3–25 (1997).
2. A. Avdeef. pH-metric solubility. 1. Solubility-pH profiles from Bjerrum plots. Gibbs buffer and pKa in the solid state. *Pharm. Pharmacol. Commun.* **4**:165–178 (1998).
3. L. Pan, Q. Ho, K. Tsutsui, and L. Takahashi. Comparison of chromatographic and spectroscopic methods used to rank compounds for aqueous solubility. *J Pharm. Sci.* **90**:521–529 (2001).

4. Absolv solute property prediction version 1.2. For further information: http://www.sirius-analytical.com/absolv.htm.
5. W. L. Jorgensen and E. M. Duffy. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **10**: 1155–1158 (2000).
6. S. H. Yalkowsky and S. Banerjee. Aqueous Solubility: Methods of Estimation for Organic Compounds. S. H. Yalkowsky and S. Banerjee, editors. Marcel Dekker Inc., New York, 1992.
7. FDA. Guidance for Industry. Waiver of in vivo bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on a biopharmaceutics classification system. For further information: http://www.fda.gov/cder/guidance/index.htm.
8. S. Venkatesh, J. Li, Y. Xu, R. Vishnuvajjala, and B. D. Anderson. Intrinsic solubility estimation and pH-solubility behaviour of cosalane (NSC 658586), an extremely hydrophobic diprotic acid. *Pharm. Res.* **13**:1453–1459 (1996).
9. D. Roy, F. Ducher, A. Laumain, and J. Y. Legendre. Determination of the aqueous solubility of drugs using a convenient 96-well plate-based assay. *Drug Dev. Ind. Pharm.* **27**:107–109 (2001).
10. W. M. Meylan, P. H. Howard, and R. S. Boethling. Improved method for estimating water solubility from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **15**:100–106 (1996).
11. M. H. Abraham and J. Le. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **88**:868–880 (1999).
12. J. W. McFarland, A. Avdeef, C. M. Berger, and O. A. Raevsky. Estimating the water solubilities of crystalline compounds from their chemical structures alone. *J. Chem. Inf. Comput. Sci.* **41**:1355–1359 (2001).
13. J. Huuskonen, M. Salo, and J. Taskinen. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **38**:450–456 (1998).
14. B. E. Mitchell and P. C. Jurs. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **38**:489–496 (1998).
15. P. B. Myrdal, A. M. Manka, and S. H. Yalkowsky. Aquafac 3: aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* **30**:1619–1637 (1995).
16. M. Mizutani. Die Dissoziation der schwachen Elektrolyte in wässerig-alkoholischen Lösungen. IV. Die Dissoziation der schwachen Elektrolyte in Methylalkohol. *Z. Physik. Chem.* **119**: 318–326 (1925).
17. A. Li and S. H. Yalkowsky. Solubility of organic solutes in ethanol/water mixtures. *J. Pharm. Sci.* **83**:1735–1740 (1994).
18. S. Ren, A. Das, and E. J. Lien. QSAR analysis of membrane permeability to organic compounds. *J. Drug Target.* **4**:103–107 (1996).
19. G. Chang, W. C. Guida, and W. C. Still. An internal coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **111**:4379–4386 (1989).
20. F. Mohamadi, N. G. J. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, and W. C. Still. MacroModel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comp. Chem.* **11**:440–467 (1990).
21. 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@galenik.uu.se).
22. P. Stenberg, U. Norinder, K. Luthman, and P. Artursson. Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.* **44**:1927–1937 (2001).
23. The PTSAs investigated in this data set were: sp2 and sp3 hybridized carbons, sp2 hybridized nitrogens, double- and single-bonded oxygen, sulfur atoms and hydrogen atoms bound to nitrogen, oxygen, and carbon atoms.
24. K. B. Lipkowitz, B. Baker, and R. Larter. Dynamic molecular surface areas. *J. Am. Chem. Soc.* **111**:7750–7753 (1989).
25. K. Palm, K. Luthman, A.-L. Ungell, G. Strandlund, and P. Artursson. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **85**:32–39 (1996).
26. E. J. Jackson. *A User's Guide to Principal Components.* Wiley, New York, 1991.
27. A. Höskuldsson. PLS regression methods. *J. Chemometrics* **2**: 211–228 (1988).
28. Simca-P v. 8.0, Umetrics AB, Box 7960, SE-907 19 Umeå, Sweden.
29. C. Hansch, J. E. Quinlan, and G. L. Lawrence. The linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **33**:347–350 (1968).
30. S. H. Yalkowsky and S. C. Valvani. Solubility and partitioning. I. Solubility of non-electrolytes in water. *J. Pharm. Sci.* **69**:912–922 (1980).
31. QikProp program version 1.2. For further information: http://www.schrodinger.com/Products/qikprop.html.
32. K. Palm, K. Luthman, A. L. Ungell, G. Strandlund, F. Beigi, P. Lundahl, and P. Artursson. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.* **41**:5382–5392 (1998).
33. Reference data for solubility values were taken from the following sources: acetylsalicylic acid from Garret, *J. Am. Pharm. Assoc.* (Sci. ed.), **46**:584–586 (1957); acyclovir from Bundgaard *et al*, *Pharm. Res.* **8**:1087–1093 (1991); ciprofloxacin from Yu *et al*, *Pharm. Res.* **11**:522–527 (1994); griseofulvin from Mosharraff and Nyström, *Int. J. Pharm.* **122**:57–67 (1995); hydrochlorothiazide from Deppeler, *Analytical Profiles of Drug Substances* **10**:406–423 (1981); hydrocortisone and testosterone from Kabasakalian *et al*, *J. Pharm. Sci.* **55**:642 (1966); ketoprofen from Herzfeldt and Kummel, *Drug Dev. Ind. Pharm.* **9**:767–793 (1983).