Prediction of pH-Dependent Aqueous Solubility of Druglike Molecules

Niclas Tue Hansen,[†] Irene Kouskoumvekaki,[†] Flemming Steen Jørgensen,[‡] Søren Brunak,[†] and Svava Ósk Jónsdóttir*,[†]

Center for Biological Sequence Analysis, BioCentrum, Technical University of Denmark, DK-2800 Lyngby, Denmark, and Department of Medicinal Chemistry, The Danish University of Pharmaceutical Sciences, Universitetsparken 2, DK-2100 Copenhagen, Denmark

Received July 12, 2006

In the present work, the Henderson−Hasselbalch (HH) equation has been employed for the development of a tool for the prediction of pH-dependent aqueous solubility of drugs and drug candidates. A new prediction method for the intrinsic solubility was developed, based on artificial neural networks that have been trained on a druglike PHYSPROP subset of 4548 compounds. For the prediction of acid/base dissociation coefficients, the commercial tool Marvin has been used, following validation on a data set of 467 molecules from the PHYSPROP database. The best performing network for intrinsic solubility predictions has a cross-validated root mean square error (RMSE) of 0.70 log *S*-units, while the Marvin pK_a plug-in has an RMSE of 0.71 pH-units. A data set of 27 drugs with experimentally determined pH-solubility curves was assembled from the literature for the validation of the combined pH-dependent model, giving a mean RMSE of 0.79 log *S*-units. Finally, the combined model has been applied on profiling the solubility space at low pH of five large vendor libraries.

INTRODUCTION

Aqueous solubility is an important determinant of the usefulness of a drug candidate that may have a significant impact on the whole process of drug discovery and development. Poor aqueous solubility is likely to hamper the uptake, distribution, transport, and, eventually, the bioavailability of a drug. Thus, it is of high interest to estimate the aqueous solubility of new drug candidates at an early stage in the drug design process.

As high throughput screening (HTS) focuses primarily on biological activity, lead compounds originating from it have tended toward higher molecular weights and lipophilicity, which are properties often associated with low solubility and poor absorption of orally administrated drugs.[1] The solubility of a neutral compound or of a compound in its nonionized form is named the intrinsic solubility and normally represented as log *S*, where *S* is the concentration of the compound in mol/L in a saturated aqueous solution in equilibrium with the solid phase. In practice, about 85% of drugs have log *S* between −1 and −5, and virtually none have values below −6. Values above −1 are not problematic, though they are often associated with highly polar molecules such as sugars and small peptides that may have low membrane permeability in the absence of active transport. Empirically, it is apparent that the target log *S* range of −1 to −5 for most drugs reflects a compromise between the polarity needed for reasonable aqueous solubility and the hydrophobicity needed for acceptable membrane passage.[2]

The solubility of a compound in water is determined by a synergy of crystal packing, cavitation, and solvation energies. Variations in solubility is mainly caused by differences in chemical structural parameters such as size, shape, steric effects, polarity, and the ability to form hydrogen bonds. The solubility of a compound is strongly influenced by the packing of the crystal lattice and thus intermolecular interactions in the solid phase. The strength of interactions within the crystalline phase plays a very important role. Therefore, much attention has been paid to the modeling of the relationship between chemical structure and solubility of organic compounds.[3]

A large number of in silico models have been developed for the prediction of the aqueous intrinsic solubility, as discussed in recent reviews by Blake,[4] Huuskonen,[5] and Jorgensen.[2] Table 1 gives an overview of the most significant contributions in the area that include both general models[3,8−13,15] as well as models that have been developed with the focus on drug and druglike compounds.[2,14,18−20] The models rarely perform well when they are faced with the complex drug candidates from the pharmaceutical industry.

In a review from 2003, Taskinen[6] states that the existing methods for predicting aqueous solubility are based on training sets of inadequate data to make them reliable in a drug discovery environment, due to the fact that experimental data sets with adequate size, distribution, diversity, and quality have not been publicly available.

The redundancies and biases of the selected training and validation sets during the development of each model have a significant impact on the performance that is depicted in the reported RMSE values. As a consequence, the models shown in Table 1 cannot be directly comparable with each other unless they are evaluated again on a common test set. Delaney[7] has recently summarized the performance of some selected models on a common test set of 21 drugs (last column of Table 1), which reveals that the models of

* Corresponding author phone: +45 45256164; fax: +45 45931585; e-mail: svava@cbs.dtu.dk.
[†] Technical University of Denmark.
[‡] The Danish University of Pharmaceutical Sciences.

10.1021/ci600292q CCC: $33.50
Published on Web 11/07/2006

Table 1. Comparison of Models for the Prediction of the Intrinsic Aqueous Solubility in Terms of Modeling Method, Complexity (Type and Number of Descriptors), General Performance (RMSE in Training and Validation Sets), and Predictive Ability on Drugs and Druglike Compounds (RMSE in Common Test Set)[a]

				training set		validation set		common test set
author	modeling method	type of descriptors	no. dscr.	*N*	RMSE	*N*	RMSE	RMSE
Huuskonen[8]	ANN	2D T	30	884	0.60	413	0.60	0.72
Hou[9]	MLR	2D T	78	1290	0.61	120	0.79	0.84
Jorgensen[2]	MLR and MC	2D T	11	317	0.63	20	1.01	
Yan[3]	ANN	3D	40	797	0.50	496	0.59	0.94
Wegner[10]	GA and ANN	2D T	9	1016	0.52	253	0.54	0.91
Delaney[11]	MLR	2D T	4	2874	0.98	528	0.96	0.87
Klopman[12]	MLR	2D S	46	469	0.46	13	0.58	1.32
Liu[13]	ANN	2D T	7	1033	0.70	258	0.71	1.01
Klamt[14]	MLR	QM	3	150	0.66	107	0.61	
McFarland[18]	MLR	2D	3	22	0.70			
Ran[20]	GSE	exp MP and K_{ow}	2			19[b]	0.72	
Huuskonen[19]	ANN	S	31	160	0.76	51	0.53	1.25
Tetko[15]	ANN	E-state indices	33	879	0.47	412	0.61	0.76

[a] ANN: artificial neural networks, MLR: multilinear regression, MC: Monte Carlo simulation, GSE: general solubility equation, T: topological, S: structural, QM: quantum mechanically derived, MP: melting point, K_{ow}: octanol/water partition coefficient, exp: experimental, 2D: two-dimensional; and 3D: three-dimensional. (2D T thus means topological descriptors derived from the two-dimensional molecular structure.) [b] The model was tested against the common test set of the 21 compounds, but only 19 results were reported.

Huuskonen[8] and Tetko[15] are those that perform best. However, the common test set is a small data set, likely not fully representative for drugs and in particular future drugs, as it includes also compounds of environmental interest (e.g. insecticides). Thus, this test performance of the models should not be over interpreted but rather taken as one subjective indicator of their predictive capabilities.

Although several models have been developed for the prediction of the intrinsic solubility, not much computational research has been done on ionizable compounds, which comprise 60% of the total number of drugs[16] and whose solubility varies considerably, often up to 1000-fold,[17] with the pH-value. The pH can adopt values from 1 to 8 throughout the different compartments of the gastrointestinal tract (GIT), and if the solubility of the drug is too low in any of the compartments, it will most likely be excreted without the possibility of passage from the gastrointestinal tract into the cardiovascular system.

The aim of this study was to develop a reliable model for predicting pH-dependent solubility of druglike compounds. For this purpose a data set of drug and druglike compounds was collected and used to develop a model based on artificial neural networks (ANN) for the prediction of both the intrinsic solubility as well as the pH-profile of the compounds. The data collection and the results of the ANN training and testing as well as the evaluation of the predictive ability of the combined model are described in this paper. To avoid confusion, in the rest of this paper S_0 is used for the description of the intrinsic solubility, while *S* is used to describe of the pH-dependent solubility.

METHODS

Data Sets. Two different data sets have been used to train the ANNs.

•A small data set assembled from several published studies (set 1-A), consisting of 378 drugs and druglike compounds with experimental intrinsic solubility data in the temperature range of 25 ± 5 °C and with an average experimental error of 0.43 log *S*-units: Those compounds found in the literature with experimental deviations larger than 1.25 log *S*-units were not included in the data set. A detailed list with all the compounds, experimental solubilities, and respective sources is available in the Supporting Information.

•A second, larger data set (set 1-B) was derived from the PHYSPROP database (www.syrres.com). PHYSPROP is a commercial data set of experimental data for physical properties, with reference to the original papers that the data have been collected from. It has previously been used successfully by other researchers for the development of aqueous solubility models.[21] Furthermore, it has the advantage of containing a large number of compounds, which provides great diversity in the data set that it is difficult to obtain otherwise and has shown to be essential for the development of good, predictive models. PHYSPROP is a database of 41 040 compounds, many of them not relevant for the drug discovery process. Therefore, a series of filters has been applied, to obtain a data set representing a druglike chemical space, with compounds expected to be present in the drug manufacturing pipeline. The following rules for the elimination of undesired compounds have been used, inspired by the work of Engkvist and Wrede:[21]

i. Available experimental intrinsic solubility data at a temperature range of 25 ± 5 °C.

ii. Molecular weight (MW) must be between 80 and 800.

iii. The compound must contain at least two carbon atoms and one of the three atom types (nitrogen, oxygen, or sulfur).

iv. The compound must have at most six halogen atoms (Cl, Br, and I) in total.

v. The compound should not contain any other atom types besides H, C, N, P, S, O, F, Cl, Br, and I.

vi. Permanently charged compounds are excluded.

vii. The following reactive groups are excluded: acylhalides, phosphanes, peroxides, isocyanates, anhydrides, acylcyanides, azides, cyanides, and diazonium compounds.

viii. Compounds present in one of the three external validation sets (see below) have been removed.

The resulting training data set consists of 4548 diverse drugs and druglike compounds with experimental intrinsic solubility data within a temperature range of 25 ± 5 °C. It should be noted here that the distribution of solubility values

in the training set influences the reliability of prediction. The larger the concentration of data points the more reliable is the prediction.

Therefore, we have analyzed the solubility and weight distributions of our data sets, and it appears that the PHYSPROP data set is slightly skewed toward more soluble compounds. The majority of the solubilities lie within a log S range of -5 to 1, with a peak around -2.5. The direct consequence is that the predictions in this range will be more reliable than predictions for extremely soluble or insoluble compounds.

Furthermore, three different data sets have been used for the final validation of the ANNs.

- 21 compounds from Ran et al.[20] were used as the common validation test (set 2).
- 239 compounds from set 1-A that are not included in set 1-B as well as 21 compounds of pharmaceutical interest research, provided to us by Lundbeck A/S (set 3).
- 22 compounds with high quality experimental pH-dependent aqueous solubility data based on the potentiometric method[22−23] as well 5 compounds with pH-dependent solubilities measured with the saturated shake-flask method[24−25] were used the validation set for the combined model for pH-dependent solubility (set 4).

Descriptors. Aiming at creating an automated solubility prediction tool, only 2D descriptors that can be derived from the molecular structure were considered. 3D descriptors have not been used as they depend on the conformation of the molecule, for which information is not stored in the canonical SMILES. Although 3D-descriptors in principle contain more accurate structural information than 2D descriptors, their conformational dependence can by default add an extra source of error. Furthermore, the calculation of 2D descriptors is less time-consuming than the calculation of the 3D descriptors, which—especially in the case of very large databases—is a significant parameter. The importance of the above observations is reflected through the trend from the literature to the use of 2D descriptors (see Table 1). The initial set of 251 descriptors calculated by MOE (Molecular Operating Environment, www.chemcomp.com) has been filtered in a two-step process: Initially, the 67 3D descriptors were removed. Furthermore, 13 descriptors that showed either no variance or represented practical rules of thumb (e.g. Lipinski's rule of five) were also removed. A detailed list with the remaining 171 descriptors and their definitions can be found in the Supporting Information.

Modeling Tools. Artificial Neural Network (ANN) Model for S_0. For modeling the intrinsic solubility a feed-forward fully connected ANN with one hidden layer a sigmoid transfer function[26] was chosen, which has been trained with the program GA-HOWLIN developed at the Centre for Biological Sequence Analysis, Technical University of Denmark. The ANN was optimized using three-part cross-validation, according to which, the data set is divided to three subsets of roughly the same size. In the case of the small-size data set, set 1-A, and in order to ensure similar property distributions, the subsets are randomized with the first principal component equally distributed, thus no parameter dominates over the others in a particular subset. The model was trained on the two subsets and tested on the remaining third in cycles. For selecting the most optimal set of descriptors used as input nodes (parameters) in the ANN the GA-HOWLIN program uses a composite algorithm that utilizes a heuristic optimization to kick-start a genetic algorithm.

The heuristic algorithm trains a network with one descriptor at a time in an initial descriptor set and selects the 30 descriptors that best fit the S_0 data. New networks are trained with all possible combinations of these 30 descriptors, and the 30 best combinations are selected for the next iteration.

The genetic algorithm starts with a population of descriptor combinations that have resulted from the heuristic algorithm. In each generation, the population is recombined and mutated. The models with the best Pearson Correlation Coefficient (PCC) values are selected and included in the next generation. This process is repeated for 18 generations.

To evaluate the performance of each algorithm separately, it has been decided to proceed with two versions of the model in parallel: one based on the heuristic algorithm (ANN_H) alone and one based on the composite heuristic and genetic algorithm (ANN_{HG}).

pK_a Model. The acid−base dissociation coefficient, pK_a, is one of the most important parameters of a drug and has given rise to a variety of predictive software, including the Marvin software developed by ChemAxon (www.chemaxon.com) and the software by ACD/Labs (www.acdlabs.com). Marvin is free of charge for academic use, and, therefore, it represents a more preferable solution over the commercial software of ACD/Labs, provided that it has also a satisfactory performance. To evaluate Marvin, a subset of 467 molecules with available experimental pK_a values was extracted from the PHYSPROP database (set 5). For reasons of consistency with the model for the prediction of intrinsic solubility, the experimental values were limited in the temperature range of 25 ± 5 °C. Furthermore, the data set has been filtered for pK_a values outside the pH range of the pH-dependent aqueous solubility data, i.e., only experimental pK_a values within a range from 0 to 13 are included in the validation. The pK_a model was also validated against the 27 compounds from set 4.

Combined Model for Prediction of pH-Dependent Aqueous Solubility. The combined model is based on the Henderson−Hasselbalch (HH) equation,[17,27] which in the case of a monoprotic acid has the form

$$\log S = \log S_0 + \log (1 + 10^{\mathrm{pH}-\mathrm{p}K_a}) \qquad (1a)$$

in the case of a monoprotic base becomes

$$\log S = \log S_0 + \log (1 + 10^{\mathrm{p}K_a-\mathrm{pH}}) \qquad (1b)$$

and for an ampholyte, the above two equations are combined to give

$$\log S = \log S_0 + \log (1 + 10^{\mathrm{pH}-\mathrm{p}K_a}(\mathrm{acid}) + 10^{\mathrm{p}K_a(\mathrm{base})-\mathrm{pH}}) \qquad (1c)$$

where S is the solubility at a given pH value, S_0 is the intrinsic solubility, and pK_a is the dissociation coefficient of the acid or the base.

The HH equation follows the degree of ionization of a compound in water and describes solubility as a function of pH. It contains a constant term, log S_0, and a term that asymptotically approaches zero as the pH decreases (for

Table 2. Performance (in Terms of R^2 and RMSE) on the Training Sets and the External Validation Set of ANN_H (Heuristic) and ANN_{HG} (Heuristic and Genetic) Versions of Model 1-A and Optimum Number of Descriptors

	training set (set 1-A)			common validation set (set 2)	
model 1-A	no. dscr.	R^2	RMSE[a]	R^2	RMSE
ANN_H	9	0.86	0.97	0.75	1.16
ANN_{HG}	>60	0.90	0.87	0.64	1.33

[a] The ANNs have been trained with the three-part validation method, which implies that the resulting RMSE describes both the correlative (training on the two subsets) and the predictive (validation on the third subset) capabilities of the model.

Table 3. Performance (in Terms of R^2 and RMSE) on the Training Sets and the External Validation Set of ANN_H and ANN_{HG} Versions of Model 1-B and Optimum Number of Descriptors

	training set (set 1-A)			common validation set (set 2)	
model 1-B	no. dscr.	R^2	RMSE[a]	R^2	RMSE
ANN_H	9	0.94	0.70	0.85	0.97
ANN_{HG}	>70	0.94	0.70	0.82	1.08

[a] The ANNs have been trained with the three-part validation method, which implies that the resulting RMSE describes both the correlative (training on the two subsets) and the predictive (validation on the third subset) capabilities of the model.

acids) or increases (for bases). The equation predicts that the solubility of an acid will rise indefinitely as the pH increases (with the opposite valid for a base); although in practice a limit is reached at the salt solubility.

To evaluate the theoretical validity of the HH equation on druglike compounds, it was tested by correlating the experimental solubility-pH profiles of the compounds in set 4, where eqs 1a−c were fed with the experimental values of S_0 and pK_a for each of the drug compounds.

The predictive capability of the combined model was evaluated on set 4, using S_0 values calculated with the developed ANN and pK_a values computed with Marvin, respectively.

RESULTS AND DISCUSSION

The Predictive Model for S_0. Two different models have been developed for the prediction of the intrinsic solubility, one based on a small, drug-focused data set (model 1-A/set 1-A) and one based on a larger and more general data set (model 1-B/set 1-B). Scanning different architectures, the optimal number of hidden neurons for the ANNs was found to be 5 and 9 hidden neurons, respectively, whereas the learning rate was found to be optimal within an interval from 0.05 to 0.5 for both models. A larger number of hidden neurons enable the model to take into account higher order correlations between input descriptors and output.

As shown in Table 2, the training of model 1-A results in a ANN_H (heuristic) version with a correlation coefficient of 0.86 with a set of 9 descriptors and an ANN_{HG} (heuristic and genetic) version with a correlation coefficient of 0.90 with a set of more than 60 descriptors. Interestingly, even though the ANN_{HG} version performs better than the ANN_H for the training set, it performs worse in the external validation set (set 2). The ANN_{HG} version has a correlation coefficient of 0.64 (29% decrease), while, in the case of the ANN_H, the decrease of the correlation coefficient is in the range of 13%. A possible interpretation could be that the genetic algorithm fits the noise as it uses a large number of descriptors to model the solubility, while the heuristic algorithm gives results of more general applicability. Compared to the best performing models in the literature (see Table 1), the descriptor selection algorithms of GA-Howlin in combination with the small data set, set 1-A, result in either a small model (ANN_H, 9 descriptors) or a very large one (ANN_{HG}, 60−80 descriptors) that is overtrained on the data set. Therefore the latter model correlates the data used for the model development accurately but has less predictive power than the ANN_H version of the model.

Due to the shortcomings of model 1-A, a second model was developed, based on a much larger data set of 4578 compounds, with the aim to achieve better predictive capabilities in both the training and the validation sets. As shown in Table 3, model 1-B has better performance both for the ANN_H and the ANN_{HG} versions, with a correlation coefficient of 0.94 in both cases. Like for model 1-A 9 descriptors are selected by the ANN_H algorithm. As shown in Figure 1, the number of descriptors increases rapidly from nine to ca. 70−100 as soon as the genetic algorithm starts,

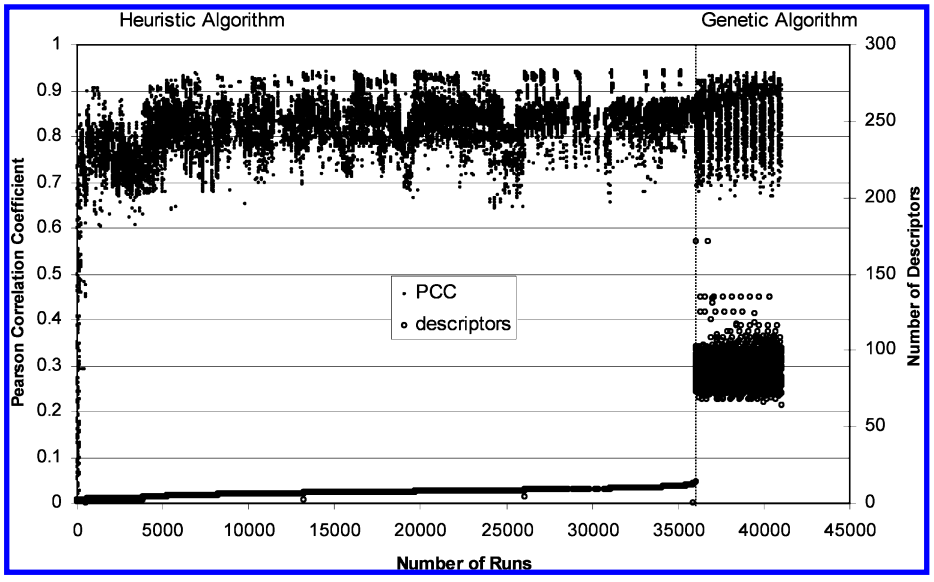

Figure 1. Pearson's correlation coefficient (PCC) and number of descriptors vs number of runs for the heuristic and the genetic algorithms of the GA-HOWLIN ANN of model 1-B.

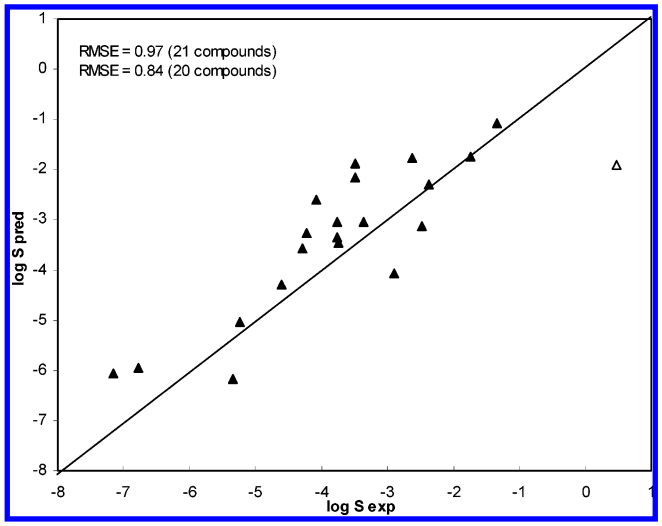

Figure 2. Calculated intrinsic solubility values by the neural network for set 2 vs experimental values. Key: (△) for the outlier and (▲) for the remaining 20 compounds.

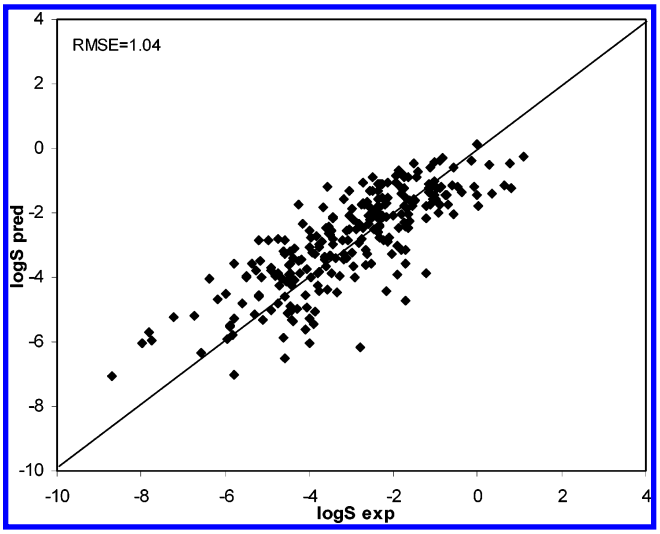

Figure 3. Calculated intrinsic solubility values by the neural network for set 3 vs experimental values.

without adding significant improvement in the PCC values. The predictive capabilities of model 1-B in the validation set have also been significantly improved compared to model 1-A. For the common validation set, the ANN_H version has a PCC of 0.85, while, in the case of the ANN_{HG}, the PCC is 0.82. The predictive power is thus not significantly reduced by having added the extra set of descriptors by the genetic algorithm, and problems with overtraining the network have been reduced by using a larger data set in the model development.

According to the results given in Tables 2 and 3, the most optimal model is the ANN_H version of model 1-B, a neural network based on 9 2D descriptors selected with the heuristic algorithm. This model was therefore chosen for the prediction of the pH-dependent aqueous solubility in the final combined model.

Model 1-B performs better than model 1-A probably due to the larger and more diverse data set (set 1-B) that has been used for its training, which contains a broad range of data, from very soluble to almost insoluble compounds that can be present in a drug manufacturing pipeline.

The performance of the model is demonstrated in Figure 2, where the experimental values of the intrinsic solubility for the compounds of set 2, the common validation set, are plotted against the predictions. An RMSE of 0.97 for the whole set of 21 compounds is a value close to the lower end of the performances of similar models from the literature (last column of Table 1). However, a closer inspection of the data set reveals one outlier, which is responsible for a large part of the resulted prediction error: Antipyrine is an exceptionally soluble drug, the only one from the common test set with a positive log *S* value. When this compound is removed form set 2, the performance of the model is significantly improved, resulting in a decrease of the RMSE to the value of 0.84.

The performance of model 1-B can be also demonstrated through the evaluation of the model in set 3 (Figure 3), which contains most of the drug and druglike molecules present in set 1-A as well as compounds of pharmaceutical interest provided to us by Lundbeck A/S. The intrinsic solubility of the compounds is satisfactorily predicted with an RMSE of 1.04, which indicates that this validation set resides in a part of the chemical space that is well described by the training set, set 1-B.

Table 4. Descriptors—Ranked in Order of Significance—in the Neural Network Based on 2D Descriptors from MOE Selected with the Heuristic Algorithm

descriptor	definition
PEO_VSA_NEG	total negative van der Waals surface area
Apol	sum of atomic polarizabilities (inducing implicit hydrogens) with polarizabilities taken from ref 28
log P(o/w)	log of the octanol/water partition coefficient calculated from a linear atom type model[29]
a_hyd	number of hydrophobic atoms
Slog P	log of the octanol/water partition coefficient calculated using an atomic contribution model[30]
SMR	molecular refractivity (including implicit hydrogens)[30]
vsa_acid	approximation to the sum of van der Waals surface areas of acidic atoms
a_nC	number of carbon atoms
a_acid	number of acidic atoms

Table 4 lists the 9 descriptors in the selected model together with their definition. Compared to the best models of Table 1, we are using less descriptors, and the performance on the validation sets (set 2, set 3) is very satisfactory. Furthermore, the selected descriptors represent physical properties such as the octanol/water partition coefficient and the van der Waals surface area, which are known to be highly correlated with the aqueous solubility. The descriptors log P(o/w) and Slog P are predicted values of the octanol—water partition coefficient based on two different models and therefore provide supplementary information to the neural network. These facts strengthen the theoretical basis of the model.

Evaluation of the pK_a Model. The results of the evaluation of the Marvin model are shown in Figure 4, where the predicted values of the dissociation coefficient are plotted against the experimental values for the compounds of set 5.

With a PCC equal to 0.94, Marvin is thus sufficiently reliable to be included in the combined prediction model. It should be mentioned, however, that the validation set (set

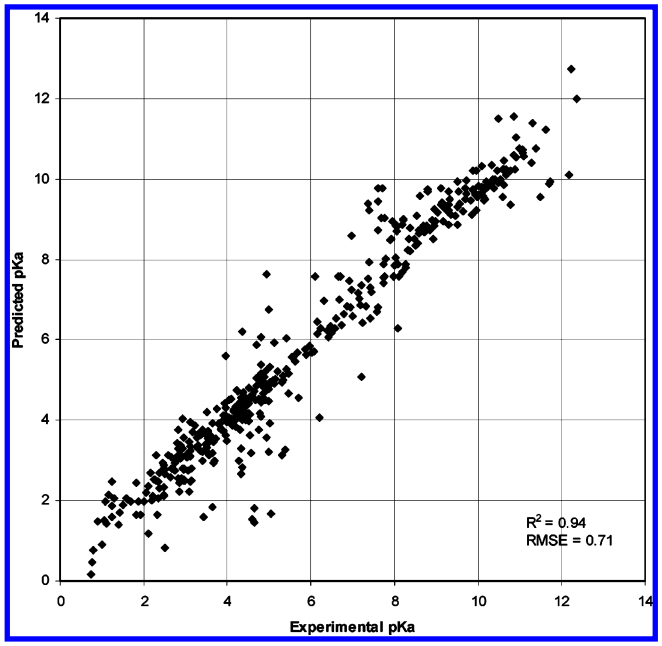

Figure 4. Predicted pK_a values with Marvin vs experimental data.

5) has not been filtered for nondruglike compounds. As a result, the complexity of the validation set may be inadequate in our case, and Marvin's performance might be somewhat less accurate for druglike compounds, but as demonstrated in Figure 4 there is an overall good agreement between predicted and experimental values. Using set 4 for an extra evaluation of Marvin, the model gives an RMSE of 0.77 pH-units. Compared to the RMSE of 0.71 pH-units on set 5, thus it seems that the original validation results overestimate slightly the actual performance of the model on the druglike set.

Correlated pH Profiles Using the HH Equation. The results on the evaluation of the HH equation are shown in the third column of Table 5. The equation is very good for correlating the solubility-pH profiles of all but two compounds, as also shown in Figures 5−7. A closer inspection of the experimental solubility-pH data of isoxazolyl-naphthoquinone-1 and isoxazolyl-naphthoquinone-4 show that their source of error is due to the fact that the experimental pK_a values have been determined from a nonlogarithmic solubility-pH curve and thus are not expected to be sufficiently accurate. For this reason, isoxazolyl-naphthoquinone-1 and isoxazolyl-naphthoquinone-4 were removed from the correlation set but not from the prediction set—as is discussed in the next paragraph—where predicted pK_a values are used instead. The mean RMSE for the remaining 25 compounds is 0.07 log *S*-units, which confirms that the HH equation provides a solid theoretical foundation for a model for the prediction of the pH-dependent aqueous solubility.

Combined Model for the Prediction of pH-Dependent Aqueous Solubility. Having selected models for the prediction of both the intrinsic solubility and the pK_a values, it is possible to develop the combined pH-dependent aqueous solubility prediction model.

The automated predictor works in the following way: a structure of the compound in canonical SMILES format enters the model and is converted into a 2D structure data

Table 5. Experimental and Predicted Disassociation Coefficient (pK_a) Values by Marvin, Experimental and Predicted Intrinsic Solubility (log S_0) Values by the Neural Network, and RMSE of the Henderson−Hasselbalch (HH) Equation Correlation and the Combined Model (Prediction) on the Experimental Solubility-pH Profiles of Set 4

	pK_a (Marvin)		log S_0 (ANN)		RMSE (HH)	
compound	exp	pred	exp	pred	corr	pred
atenolol[b,23]	9.54	9.87	−1.30	−1.94	0.03	0.56
benzoic acid[b,23]	3.99	4.18	−1.59	−1.43	0.03	0.13
cimitidine[b,23]	6.93	6.65	−1.43	−2.06	0.02	0.54
diclofenac[b,23]	3.99	4.00	−5.59	−4.48	0.13	1.21
diltiazem[b,23]	8.02	8.43	−2.95	−4.20	0.05	1.11
enalapril maleate[b,23]	2.92/5.42	3.18/5.49	−1.33	−3.34	0.11	1.90
famotidine[b,23]	11.19/6.74	9.70/3.29	−2.48	−2.62	0.01	0.72
fentanyl[a,24] (dp:11)	8.99	9.38	−4.53	−4.25	0.24	0.63
flurbiprofen[b,23]	4.03	4.42	−4.36	−3.42	0.06	0.80
furosemide[b,23]	10.63/3.52	9.91/4.25	−4.75	−2.97	0.05	1.50
hydrochlorothiazide[b,23]	9.96/8.87	9.96/9.24	−2.63	−1.98	0.10	0.60
ibuprofen[b,23]	4.42	4.85	−3.62	−2.92	0.03	0.54
isoxazolyl-naphthoquinone-1[a,c,25] (dp:6)	8.77	5.14	−3.90	−2.86	1.68	0.38
isoxazolyl-naphthoquinone-4[a,c,25] (dp:5)	7.72	5.30	−3.31	−3.37	2.89	1.61
ketoptofen[b,23]	4.05	3.88	−3.33	−3.31	0.09	0.21
labetolol[b,23]	7.48/8.05	9.42/9.75	−3.45	−3.50	0.06	0.41
metroprolol[b,23]	9.52	9.87	−1.20	−2.30	0.04	0.83
morphine[a,c,24] (dp:14)	8.08	8.97	−3.57	−1.95	0.52	3.19
nadolol[b,23]	9.69	9.96	−1.57	−2.03	0.04	0.39
naproxen[b,23]	4.18	4.19	−4.21	−3.12	0.02	1.10
phenytoin[b,23]	8.21	9.13	−4.13	−3.28	0.01	0.70
propoxyphene[b,23]	9.06	9.47	−5.01	−4.89	0.07	0.33
propranolol[b,23]	9.53	9.87	−3.62	−2.81	0.08	0.96
quinine[b,23]	8.53/4.24	8.39/4.44	−2.82	−3.33	0.01	0.55
sufentanil[a,24] (dp:11)	8.51	9.04	−5.51	−4.25	0.35	1.81
terfenadine[b,23]	9.53	10.20	−6.69	−6.58	0.05	0.58
trovafloxacin[b,23]	5.87/8.03	5.11/9.79	−4.53	−3.59	0.01	0.56
av RMSE					0.25	0.88
av RMSE (outliers removed)					0.07	0.79

[a] Shake-flask experiment. [b] Potentiometric experiment (pSOL). [c] Outliers dp: number of data points.

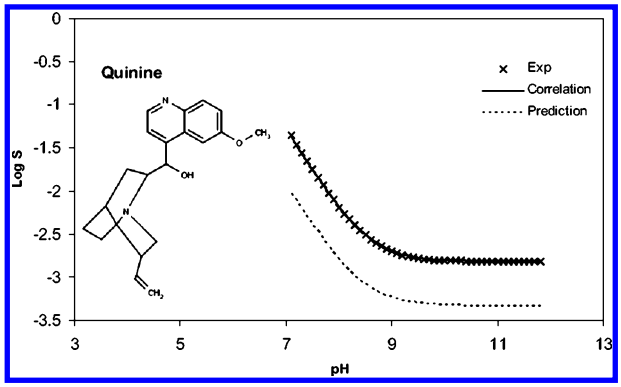

Figure 5. pH-dependent solubility curves for quinine: experimental data (crosses), the correlation with the HH equation (continuous line), and the prediction with the combined model (discontinuous line).

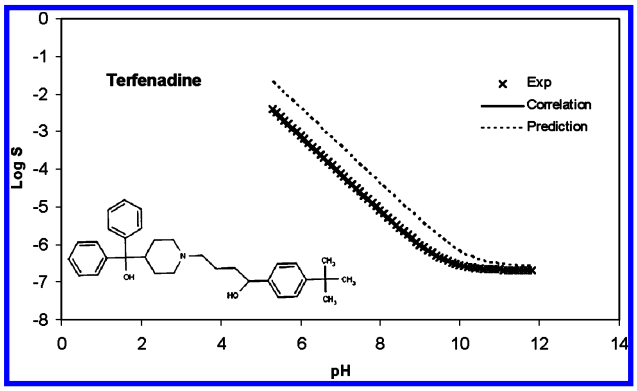

Figure 6. pH-dependent solubility curves for terfenadine: experimental data (crosses), the correlation with the HH equation (continuous line), and the prediction with the combined model (discontinuous line).

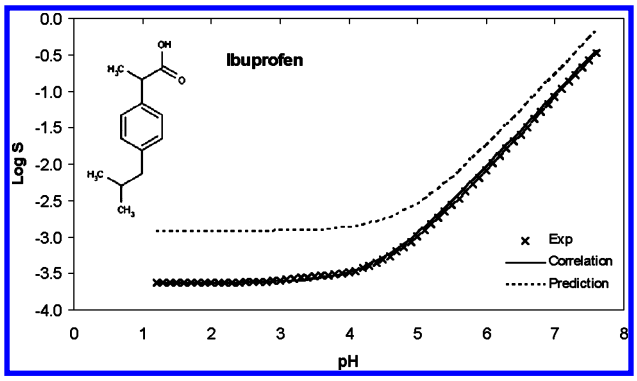

Figure 7. pH-dependent solubility curves for ibuprofen. Experimental data (crosses), the correlation with the HH equation (continuous line), and the prediction with the combined model (discontinuous line).

file (sdf). Through a 'svl' script, MOE is activated and calculates all necessary descriptors for the input compound. The descriptor values are then normalized and fed to the ANN for the prediction of the intrinsic solubility. The solubility predictor returns a value between 0 and 1, and subsequently this value is transformed back to the normal solubility scale with the help of an inverse sigmoid function. At the same time, the 2D structure is fed to the Marvin pK_a prediction plug-in. The output from Marvin is parsed, thus only returning an array of pK_a values. The predicted values for the intrinsic solubility and the dissociation constants are combined through the HH equation, which returns the predicted curve for the pH-dependent aqueous solubility.

Table 5 summarizes the performance of the combined model in predicting the solubility-pH profiles of the compounds in set 4. The resulting mean RMSE of the combined model is 0.79 log *S*-units, which shows that it is a reliable predictive tool for the pH-dependent solubility for drug and druglike compounds. Morphine has been removed as the only outlier, since the high deviation of the prediction ($RMSE_{Prediction} = 3.19$) is to a large extent due to the correlation error ($RMSE_{Correlation} = 0.52$) of the experimental data.

Figures 5−7 show the experimental and predicted pH-dependent aqueous solubility curves for three selected drug compounds from set 4: quinine, terfenadine, and ibuprofen. The experimental data are matched almost perfectly with the correlated curves obtained with the HH equation. According to eqs 1a−c, the prediction error of the combined model equals to the sum of the errors from the prediction of the intrinsic solubility and the prediction of the dissociation coefficient. The error in the prediction of the intrinsic solubility displaces the solubility-pH curve vertically (Figure 5, quinine), while the error in the prediction of the dissociation coefficient corresponds to an uncertainty in the horizontal placement of the solubility-pH curve and affects only the graded part of the curve (Figure 6, terfenadine). In other words, inaccurate values of the dissociation coefficients do not add error to the flat part of the curve (that of constant solubility), but inaccurate values of the intrinsic solubility add error to the whole curve.

In Figure 7, the combined model that we have developed overestimates the intrinsic solubility of ibuprofen, as seen in the flat part of the curve, but captures the pH-dependency with good accuracy. In this case the errors in the prediction of the intrinsic solubility and of the disassociation coefficients are of opposite sign and thus are canceling out, which explains the very accurate prediction of the pH-dependency despite the overestimation of the value of the intrinsic solubility.

Finally, we used the combined model to depict the solubility space of chemical compounds provided by five big vendors taken from the ZINC library,[31] namely ChemBridge (339 209 unique compounds), ChemDiv (251 694 unique compounds), NCI (188 315 unique compounds), Maybridge (53 038 unique compounds), and Sigma-Aldrich (49 017 unique compounds). Figure 8 shows the intrinsic solubility profile of each library, while Figure 9 shows the solubility profile at pH = 2.

It is worth noticing in both figures that the smaller libraries have the higher proportion of soluble compounds, both at intrinsic and ionized state. As seen in Figure 8, the majority of the compounds of ChemBridge and ChemDiv libraries have intrinsic solubilities in the range −6 to −5, whereas in Sigma-Aldrich most of the compounds are evenly spread in a solubility range of −5 to −2.

When it comes to solubility at pH = 2, there are fewer nonsoluble compounds in all libraries and, at the same time, significantly more soluble compounds up to solubility values of −1. Moreover, NCI seems to be the library with the most soluble 'character', both for compounds in their neutral state and for ionized molecules at low pH. NCI is a library of molecules experimentally tested as potential drug candidates, and Sigma-Aldrich is the one that looks more similar to it. Sigma-Aldrich contains a high number of molecules within the solubility range of −5 to −1, which is the desired range

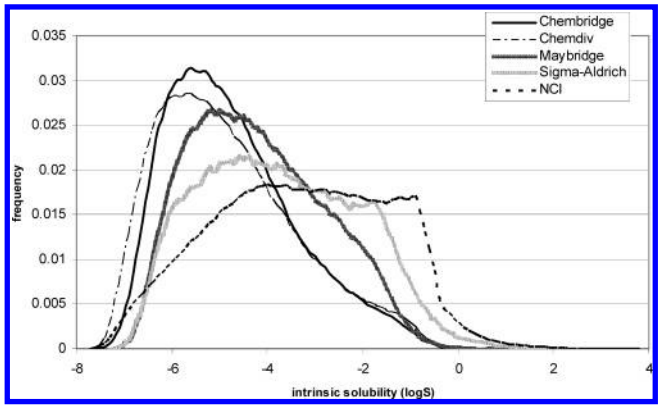

Figure 8. Intrinsic solubility profile of ChemBridge. ChemDiv, NCI, Maybridge, and Sigma-Aldrich libraries.

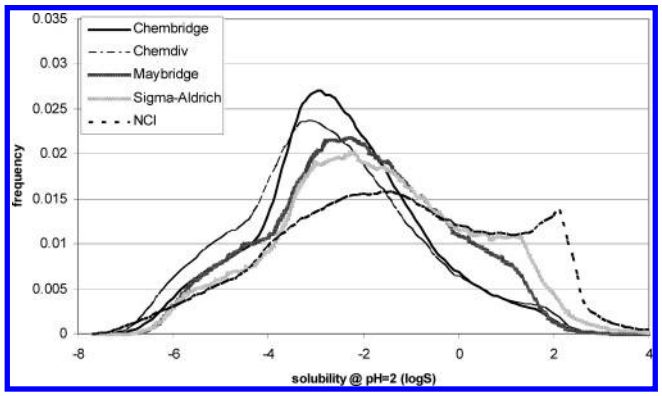

Figure 9. Solubility profile at pH = 2 of ChemBridge. ChemDiv, NCI, Maybridge, and Sigma-Aldrich libraries.

in a drug development pipeline. Figures 8 and 9 demonstrate the usefulness of our tool for evaluating existing commercial and in-house libraries as well as for composing a new library of a desired solubility distribution.

One major weakness of the combined model is that it is not temperature-dependent. The current model is optimized for a temperature range in the vicinity of 25 °C, whereas applications in the pharmaceutical industry would require the model to work at 37 °C. However, such an extension of the model would require the availability of sufficient amount of experimental data in the temperature range of interest, which is not the case at present. One other area for future improvement is the extension of the combined model to take salt solubilities into account, i.e., to model at which pH the ionized drugs precipitate as salts. This will be a novelty in the area of available prediction tools, and it would automatically broaden its applications substantially. However, such a step cannot be taken unless trustworthy experimental data on salt solubilities involving drugs and druglike compounds will become available.

CONCLUSION

In this work we have developed a model for prediction of pH-dependent aqueous solubility of drugs and druglike compounds. The model is a combination of two separate modules: (1) an ANN for the prediction of the intrinsic solubility which has been developed as a part of this work and (2) an already available tool, Marvin developed by ChemAxon, for the prediction of the dissociation coefficients. The performance of the developed ANN is of comparable accuracy to the best tools found in the open literature, and our model uses significantly fewer descriptors compared to other models. To our knowledge nothing has been published up to now regarding a tool for the prediction of the pH-dependence of the aqueous solubility of complex compounds (although commercial software like ACD/Labs include such a tool in their package.)

One of the most interesting applications of the combined model for prediction of pH-dependent aqueous solubility is to use it in compartment-based predictive bioavailability models taking the pH-levels in the different compartments into consideration. Furthermore, the combined model can be used for evaluating existing commercial and in-house libraries as well as for composing new libraries of a desired solubility distribution at specific pH levels.

ACKNOWLEDGMENT

The work is supported by a grant from the Program Commission on Nanoscience, Biotechnology and IT (NABITT), and the Danish Research Council for Technology and Production Sciences. The authors thank Jens Pontoppidan Larsen and Thomas Sicheritz-Ponten from CBS/BioCentrum for the assistance on GA-Howlin and Olivier Taboureau in assistance in descriptor generation. We should also like to thank Lundbeck A/S, Valby, Denmark, and, in particular, the head of the computational chemistry group at Lundbeck A/S, Klaus Gundertofte, for providing information on the 21 compounds included in set 3.

Supporting Information Available: Lists of the compounds of data set set 1-A with experimental intrinsic solubility data and literature source and 171 descriptors calculated by MOE and their definition. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lipinski, C. A. Drug-like properties and the Causes for poor Solubility and poor Permeability. *J. Pharm. Tox. Methods* **2000**, *22*, 387−398.
(2) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.
(3) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds by Topological Descriptors. *QSAR Comb. Sci.* **2003**, *22*, 821−829.
(4) Blake, J. F. Chemoinformatics − predicting the Physicochemical Properties of Drug-like Molecules. *Curr. Opin. Biotech.* **2000**, *11*, 104−107.
(5) Huuskonen, J. Estimation of Aqueous Solubility in Drug Design. *Comb. Chem. High Throughput Screening* **2001**, *4*, 311−316.
(6) Taskinen, J.; Yliruusi, J. Prediction of Physicochemical Properties based on Neural Network Modeling. *Adv. Drug Delivery Rev.* **2003**, *55*, 1163−1183.
(7) Delaney, J. S. Predicting Aqueous Solubility from Structure. *Drug Discovery Today* **2005**, *10*, 289−295.
(8) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(9) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility based on Atom Contribution. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.
(10) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient optimized by a Genetic Algorithm-based Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(11) Delaney, J. S. Esol: Estimating Aqueous Solubility directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(12) Klopman, G. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(13) Liu, R.; So, S. Development of Quantitative structure−property Relationship Models for early ADME Evaluation in Drug Discovery: 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633−1639.

(14) Klamt, A. Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS. *J. Comput. Chem.* **2002**, *23*, 275−281.
(15) Tetko, I. V.; Tanchuk, V. Y.; Kasheva T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds using E-state Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(16) Comer, J. E. A. In *Drug Bioavailability*; van de Waterbeemd, H., Ed.; Wiley-VCH: New York, 2003; Vol. 1, Chapter 2, pp 21−45.
(17) Bergström, C. A. S.; Luthman, K.; Artursson, P. Accuracy of calculated pH-dependent Aqueous Drug Solubility. *Eur. J. Pharm. Sci.* **2004**, *22*, 387−398.
(18) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the Water Solubilities of Crystalline Compounds from their Chemical Structures alone. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355−1359.
(19) 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.
(20) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.
(21) Engkvist, O.; Wrede P. High-Throughput, *In Silico* Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247−1249.
(22) 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.
(23) *Compendium of Solubility Standards, Standard Compounds for the pSOL-3 Instrument*; *p*ION Inc.:, 2003; Vol. 1.
(24) Roy, S. D.; Flynn, G. L. Solubility Behavior of Narcotic Analgesics in Aqueous Media: Solubilities and Dissociation Constants of Morphine, Fentanyl, and Sufentanil. *Pharm. Res.* **1989**, *6*, 147−151.
(25) Granero, G.; de Bortello, M. M.; Brinõn, M. C. Solubility profiles of some isoxazolyl-naphthoquinone derivatives. *Int. J. Pharm.* **1999**, *190*, 41−47.
(26) Baldi, P.; Brunak, S. In *Bioinformatics: The Machine Learning Approach*, 2nd ed.; The MIT Press: 2001; Chapter 5, pp 91−104.
(27) Hasselbalch, K. A. Calculation of blood pH based on the free and bound carbonic acid, and oxygen binding of blood as function of pH. *Die Biochem. Z.* **1916**, *78*, 112−144.
(28) *CRC, Handbook of Chemistry and Physics*; CRC Press: 1994.
(29) Labute, P. source code in MOE logP model, unpublished.
(30) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(31) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

CI600292Q