# In Silico Prediction of Buffer Solubility Based on Quantum-Mechanical and HQSAR- and Topology-Based Descriptors

Andreas H. Göller,*,† Matthias Hennemann,‡ Jörg Keldenich,† and Timothy Clark‡

Bayer Healthcare AG, Aprather Weg 18a, 42096 Wuppertal, Germany, and Friedrich-Alexander-Universität
Erlangen-Nürnberg, Computer-Chemie-Centrum, Nägelsbachstrasse 25, 91052 Erlangen, Germany

Received August 11, 2005

We present an artificial neural network (ANN) model for the prediction of solubility of organic compounds in buffer at pH 6.5, thus mimicking the medium in the human gastrointestinal tract. The model was derived from consistently performed solubility measurements of about 5000 compounds. Semiempirical VAMP/AM1 quantum-chemical wave function derived, HQSAR-derived logP, and topology-based descriptors were employed after preselection of significant contributors by statistical and data mining approaches. Ten ANNs were trained each with 90% as a training set and 10% as a test set, and deterministic analysis of prediction quality was used in an iterative manner to optimize ANN architecture and descriptor space, based on Corina 3D molecular structure and AM1/COSMO single point wave function. In production mode, a mean prediction value of the 10 ANNs is created, as is a standard deviation based quality parameter. The productive ANN based on Corina geometries and AM1/COSMO wave function gives an $r^2_{cv}$ of 0.50 and a root-mean-square error of 0.71 log units, with 87 and 96% of the compounds having an error of less than 1 and 1.5 log units, respectively. The model is able to predict permanently charged species, e.g. zwitterions or quaternary amines, and problematic structures such as tautomers and unresolved diastereomers almost as well as neutral compounds.

## INTRODUCTION

Solubility is an important issue for the design of new drugs with favorable ADMET (absorption, distribution, metabolism, excretion, toxicity) properties. Solubility modulates the bioavailability to a great extent. Poor solubility usually translates into limited absorption hampering the desired therapeutic outcome. Sometimes, this issue can be overcome by elaborate formulation of a compound but increasing the cost of goods for the market product. Hence, there is a need to address this topic as early as possible in the drug discovery process, either by high-throughput experimental or by in silico solubility prediction, which has the additional charm that it can be applied to virtual compounds.

Experimentally, one has to face the effect that compounds will give different solubilities depending on crystallization state and quality, i.e., purity, powder vs crystal, counterion effects, cocrystallized solvents, and also depending on experimental conditions such as temperature,[1] solvent, solvation or desolvation experiment, lab equipment, and probably even lab personnel.[1] These effects can only partially be addressed by a highly standardized apparatus and quality management.

Today, all practical in silico models dealing with druglike molecules still have to be based on experimental training sets, since ab initio treatments are not feasible due to the complexity of the phenomenon solvation.[2,3] The number of published models today does not allow for discussion of all of them in detail. The current situation was reviewed recently.[4]

The models currently available to us all show weaknesses to predict solubility values that are consistent with our experimental setup. The reasons are manifold and can be understood by analysis of the published models and are mainly as follows: (i) Our setting is a desolvation experiment from DMSO solution into buffer medium with physiological pH (see the section Methods), whereas most of the models try to predict aqueous solubility. (ii) Many of the models[5a] are based on the data set of Huuskoonen et al.[6] derived from the AQUASOL[7] and PHYSPROP[8] databases or other data sets,[9] which contain mostly nondruglike compounds. (iii) The Huuskoonen data set and many others stem from various experimental sources and therefore are not consistent.[1] (iv) Bergström et al.,[10] on the other hand, describe a model based on only 85 in-house compounds. Even though this set is diversity-selected by ChemGPS one cannot expect that the data basis is large enough to build a global model.

Current models are based on a great variety of descriptors, even including experimentally based models, that move the problem from measuring solubility to measuring logP or crystallization energies, for instance.[11] The descriptors used range from structure-fingerprint[12] to group-contribution[13] to topological/Estate[6] and atom/molecule-based[14] ones such as molecular weight, polarizability, or radial distribution functions[5] and are derived from 2D and 3D structures and by quantum-mechanical calculations. Interestingly, many of

* Corresponding author phone: ++49-202-365442; e-mail: andreas.goeller@bayerhealthcare.com

† Bayer Healthcare AG.

‡ Friedrich-Alexander-Universität Erlangen-Nürnberg.

10.1021/ci0503210 CCC: $33.50 © 2006 American Chemical Society
Published on Web 01/19/2006

these models employ some logP as an additional descriptor, which is clear from the older observations that logP and solubility are strongly correlated.[11]

The authors employ a variety of linear and nonlinear approaches to build their models, like artificial neural nets (ANN),[5,6,9] multiple linear regression (MLR),[9,14,15] support vector machines (SVM),[12] decision trees (recursive partitioning RP),[5b] or partial least squares (PLS).[5b] MLR and RP automatically yield the subset of the predictive descriptors from the sometimes overwhelmingly large starting collection; other models need preselection. Therefore, genetic algorithms (GA),[9,16] simulated annealing,[9] self-organizing maps (SOM),[5a] or principal components analysis (PCA),[10] or MLR, RP, PLS are employed as a prefilter.[5b]

Thus, all models commercially available to us will, due to the data sets they are built upon, not be able to give predictions of drug candidate solubilities in the media found in the human gastrointestinal tract, as they are measured in our fully automated platform in buffer at pH 6.5. This is illustrated by Figure 2, which shows predicted in-house solubilities by a model based on the Huuskoonen data set. We therefore were in need for an in silico prediction model of “our” solubility, i.e., a decision-driving tool for medicinal chemists’ daily compounds based on our experimental setup.

In this work, we describe our in-house solubility model that is based on a mixed set of semiempirical AM1-derived, HQSAR- and topology-based descriptors as input parameters for a back-propagation feed-forward artificial neural network (ANN) model for in silico solubility prediction for drug-candidate molecules in buffer medium at pH 6.5. We also describe the descriptor selection process by various data mining techniques and the model’s predictive power.

## METHODS

**Data Sets.** The starting point of our investigations was an unpublished ANN model obtained by Clark et al.,[17] based on the data set **DS0**, a structurally diverse subset from the Aquasol[7] data set. The data set consisted mainly of non-druglike compounds, and the ANN was not suitable for our purposes.

Therefore, we created our own data set starting from 5928 compounds with in-house experimental aqueous solubilities ranging from 0.02 mg/L to 2000 mg/L. The structural representation of the compounds was handled as described in the following: We stripped off counterions and neutralized all compounds. To create an unambiguous data set with as little noise as possible, by visual inspection and filtering we cleaned up the data set by removing (i) all registered compounds containing more than one large molecular fragment, like e.g. stereoisomer mixtures, (ii) permanently charged molecules such as quaternary amines, (iii) zwitterions, (iv) molecules with ambiguous tautomerism, and (v) several molecules which failed automatic preprocessing by CORINA[18,19] 3D generation or VAMP.[20] The final training data set **DS1** contained 4889 compounds.

A smaller training data set **DS2** was formed containing those 2163 compounds from **DS1** that were measured experimentally by the HT experimental procedure described in the following, since we observed some shift in the measured value range and distribution with this subset.

By visual inspection of the outliers it became obvious, that some values, especially if compared to structural analogues, were probably wrong, i.e., deviated by about 0.5 to 2 log units from their close analogues. We remeasured about 40 of these compounds and corrected 16 strongly deviating values. We removed another 83 compounds that had suspicious values but where we had no substance available anymore for experimental retesting, yielding **DS3** with 4806 compounds.

The value distribution in **DS3** is much nearer to a uniform distribution than expected. Binned into 0.5 log units sets, we find subsets of 158 molecules (<6.5), 235 (<6.0), 455 (<5.5), 766 (<5.0), 899 (<4.5), 783 (<4.0), 597 (<3.5), 717 (<3.0), 196 (<2.0), respectively. Nevertheless, since our criteria for a predictive model are the percentages of compounds with maximal deviations from experiment lower than trust regions of 1 and 1.5 log units, we did the following test: we assigned any compound the mean value of $-4.55$ mol $L^{-1}$ resulting in 63% of the compounds in the 1 log unit and 86% in the 1.5 log units trust regions, respectively. The final model is performing considerably better.

Two test sets were constructed and applied on the models in addition: the first one, **DS4**, contains all 1064 compounds that were removed by the procedures denoted before, and the second one, **DS5**, contains a set of 7222 compounds that were measured after model creation. Models were built based on the common logarithm of the solubilities in units of mol $L^{-1}$.

**Descriptor Calculation.** Descriptor calculation involved the following steps:

(i) Compounds were converted from two-dimensional to single conformation three-dimensional (3D) structures using CORINA.[18]

(ii) The AM1[21] electron wave function was calculated on the Corina 3D structure by COSMO[22] single point calculations. The COSMO approach (solvent water, $\epsilon = 78.0$, effective solvent radius 1.0, effective charge radius of the solvent 1.0, distance up to which segment−segment interactions are evaluated 2.0, number of segments per atom 42) as implemented into VAMP[20] was applied to account for the influence of solvent-induced polarization on the molecular electrostatics and hence the wave function based descriptors.

(iii) The AM1 electronic wave function information as encoded by the natural atomic orbital based point charge (NAO-PC)[23] model and stored in VAMP result SD-files is transformed by PROPGEN1.0[24] into a set of 65 molecular descriptors (see Table 1).

(iv) Via our in-house property calculation engine PILO we calculated eight additional 2D structure derived descriptors. These are two descriptors for the estimation of pH dependent logP at pH = 2.3 and pH = 7.5.[25] The other descriptors are flexibility, numbers of ring atoms, sum of OH and NH groups (H-donors), sum of O and N atoms (H-acceptors), and polar and lipophilic surface areas.[26]

All models described in this publication are based on AM1/COSMO wave functions and Corina geometries. Alternative models were built based on Corina 3D coordinates and VAMP[20] AM1 gas-phase single points, AM1 gas-phase optimizations (GNORM = 0.4 kcal $mol^{-1}$ Å$^{-1}$, EF optimizer), and on COSMO/AM1 optimizations but were unpractical due to the reasons described in the following.

**Table 1.** List of VAMP/Propgen and PILO Descriptors[c]

| no. | acronym | description | model[a] | ref |
|---|---|---|---|---|
| 1 | $\mu$ | total molecular dipole moment | 3 | 21 |
| 2 | dipden | dipole density: total dipole moment/molar volume | 3 | 36 |
| 3 | dipden2 | (total dipole moment)$^2$/molar volume | | 36 |
| 4 | α | total polarizability (original variational) | | 37 |
| 5 | m0pol | total polarizability (parametrized model 0) | 3 | 38 |
| 6 | m4pol | total polarizability (parametrized model 4) | | 38 |
| | | Sums of the Electrostatic Potential-Derived Atomic Charges on | | |
| 7 | QsumH | H atoms | 1,2,3 | 39 |
| 8 | QsumN | N atoms | 1,2,3 | 39 |
| 9 | QsumO | O atoms | 1,3 | 39 |
| 10 | QsumP | P atoms | | 39 |
| 11 | QsumS | S atoms | | 39 |
| 12 | QsumHal | halogen atoms | | 39 |
| 13 | QsumF | F atoms | | 39 |
| 14 | QsumCl | Cl atoms | | 39 |
| 15 | QsumBr | Br atoms | | 39 |
| 16 | QsumI | I atoms | | 39 |
| 17 | $nV_S+$ | no. of triangles on the surface with a + MEP[b] | | 40 |
| 18 | $nV_S^-$ | no. of triangles on the surface with a − MEP | | 40 |
| 19 | $V_{S,max}$ | max MEP on the surface | 3 | 40 |
| 20 | $\underline{V}_{S,min}$ | min MEP on the surface | 3 | 40 |
| 21 | $\overline{\underline{V}}_S+$ | mean + MEP | 1,2,3 | 40 |
| 22 | $\overline{\underline{V}}_S^-$ | mean − MEP | 3 | 40 |
| 23 | $\overline{V}_S$ | mean MEP | 3 | 40 |
| 24 | $\sigma^2{}_+$ | variance of + MEP | 2,3 | 40 |
| 25 | $\sigma^2{}_-$ | variance of − MEP | 3 | 40 |
| 26 | $\sigma^2{}_{tot}$ | total variance ($\sigma^2{}_+ + \sigma^2{}_-$) | 3 | 40 |
| 27 | $\nu$ | balance parameter | 3 | 40 |
| 28 | $\nu\ \sigma^2{}_{tot}$ | balance × total variance | 1,3 | 40 |
| 29 | $\pi$ | average deviation of the MEP: | 3 | 40 |
| 30 | $\epsilon_A$ | ‘covalent’ H bond acidity | 1,3 | 41 |
| 31 | $\epsilon_B$ | ‘covalent’ H bond basicity | 1,3 | 41 |
| 32 | q+ | ‘electrostatic’ H bond acidity | 1,2,3 | 41 |
| 33 | q- | ‘electrostatic’ H bond basicity | 1,3 | 41 |
| 34 | nAcc | no. of H bond acceptor groups | 3 | 24 |
| 35 | nDon | no. of H bond donor groups | 3 | 24 |
| 36 | nAryl | no. of aromatic rings | 1,3 | 24 |
| 37 | cohindex | cohesive index (nAcc × nDon$^{1/2}$/surface) | 2,3 | 42 |
| 38 | MW | molecular weight | 2,3 | 24 |
| 39 | volume | molecular volume | 3 | 43 |
| 40 | surface | molecular surface area | 3 | 43 |
| 41 | globularity | Globularity | 1,3 | 44 |
| | | Sum of E-States Based on QM-Calculated Bond Orders on | | |
| 42 | EstateN | N atoms | | 45 |
| 43 | EstateO | O atoms | | 45 |
| 44 | EstateP | P atoms | | 45 |
| 45 | EstateS | S atoms | | 45 |
| 46 | EstateHal | halogen atoms | | 45 |
| 47 | EstateF | F atoms | | 45 |
| 48 | EstateCl | Cl atoms | | 45 |
| 49 | EstateBr | Br atoms | | 45 |
| 50 | EstateI | I atoms | | 45 |
| | | Sum of E-States Based on Distances on | | |
| 51 | Estate2N | N atoms | | 45 |
| 52 | Estate2O | O atoms | | 45 |
| 53 | Estate2P | P atoms | | 45 |
| 54 | Estate2S | S atoms | | 45 |
| 55 | Estate2Hal | halogen atoms | | 45 |
| 56 | Estate2F | F atoms | | 45 |
| 57 | Estate2Cl | Cl atoms | | 45 |
| 58 | Estate2Br | Br atoms | | 45 |
| 59 | Estate2I | I atoms | | 45 |
| 60 | $nV_{S,<-30}$ | no. of triangles on the surface with a MEP $< -30$ | 2,3 | 24 |
| 61 | $nV_{S,[-30,30]}$ | no. of triangles on the surface with a MEP between −30 and +30 | 3 | 24 |
| 62 | $nV_{S,>30}$ | no. of triangles on the surface with a MEP $> 30$ | 3 | 24 |
| 63 | $pV_{S,<-30}$ | partition of triangles on the surface with a MEP $< -30$ | 1,3 | 24 |
| 64 | $pV_{S,[-30,30]}$ | partition of triangles on the surface with a MEP between −30 and +30 | 3 | 24 |
| 65 | $pV_{S,>30}$ | partition of triangles on the surface with a MEP $> 30$ | 1,3 | 24 |
| 66 | $logP_{pH2.3}$ | estimated logP at pH 2.3 | 1,2,3 | c) |
| 67 | $logP_{pH7.5}$ | estimated logP at pH 7.5 | 1,3 | c) |
| 68 | flexibility | flexibility: | 1,2,3 | c) |
| 69 | nRingAtoms | no. of ring atoms | 1,2 | c) |
| 70 | OH+NH | no. of OH and NH | | c) |
| 71 | O+N | no. of O and N atoms | | c) |
| 72 | PSA | polar surface area | | c) |
| 73 | LSA | lipophilic surface area | | c) |

[a] Numbers of models where descriptors appear. [b] MEP = molecular electrostatic potential. [c] PILO descriptors; PILO is a Bayer in-house application for the calculation of molecular descriptors and physicochemical parameters.

Quantum-mechanical gas-phase geometries are usually collapsed conformations with maximized intramolecular interactions. Corina, on the other hand, is parametrized based on small molecule X-ray structures. In crystals, small molecules balance the inter- and intramolecular interactions which in turn favor more elongated geometries. In a polar solvent like water, on the other hand, intermolecular hydrogen bonds to solvent compete with intramolecular ones, also leading to elongated structures. It is a common experience that Corina derived structures will be more similar to AM1 optimized geometries with the inclusion of continuum solvation models than to AM1 gas-phase derived ones. The large increase in computer time per molecule and additionally risks such as convergence problems, after some early tests, let us decide that AM1/COSMO geometries are not feasible for a high-throughput model.

For the same reason, conformational sampling, Boltzmann averaging, or global minima searches were not attempted. In the course of other ongoing projects, we have analyzed the influence of conformation on the variability of descriptor values. For flexible molecules, there is a certain deviation between global minimum, Boltzmann average, and "random" local minimum. Pharmaceutical molecules are mostly not as flexible, and therefore the deviations are lower and acceptable, especially when looking into the other sources of uncertainty in the model-building process.

**Descriptor Selection.** Removal of insignificant descriptors improves the model quality via the noise-to-information ratio, though a neural net from its construction should be able to handle a certain amount of insignificant descriptors. There are various methods for descriptor selection described in the literature, some of which were already mentioned in the Introduction, like genetic algorithms, PCA, SOM, or stepwise regression. We prefer to use a nonautomatic descriptor selection as an nonautomatic iterative procedure, with learning cycles of descriptor set selection, ANN training, ANN application, and outlier analysis. We have tested four different techniques for descriptor-preselection, namely analysis of pairwise descriptor-property correlations and following stepwise linear regression (SLR),[27] multiple linear regression analysis (MLR),[27] and nonbinary decision tree software "Formal Inference-based Recursive Modeling" (FIRM).[28,29] FIRM came up as superior to SLR and MLR and was solely applied.

For any promising set of descriptors derived from a FIRM decision tree, we trained several ANNs, systematically increasing the number of hidden nodes from two to the number of the input nodes. The best ANN based on predictivity for the test set was then re-evaluated. Following this, we analyzed the residual values for any descriptor in this ANN for any compound using FIRM again, which enabled us to remove descriptors adding noise to the model. It was also useful to check the direct correlation of each possible descriptor with the residual values obtained by the ANN.

After removal of these insignificant descriptors, new ANNs were trained. If the quality of the ANN obtained was as good as before, we again removed one or more descriptors and trained the ANN with different numbers of hidden nodes once again. This procedure was repeated until the predictive power of the ANN declined.

**Feed-Forward Neural Nets.** Three-layer feed-forward artificial neural nets (ANN) using sigmoid transfer functions and trained with the back-propagation of errors algorithm[30,31] were applied in this work.

We used a 10-fold cross-validation strategy for the ANN models. For each model, the data set was split into 10 equal subsets, yielding 10 training sets of each 90% of the compounds, and their corresponding test sets consisting of 10% of the compounds. Then, 10 separate ANN with random starting weights were trained with these training and test sets, chosen such that each molecule appears in the test set for one and only one network. The cross-validated result for any given molecule is then the prediction by the net where it appeared as test set molecule. This should therefore give a worst case prediction for the given molecule in this model.

Training of the nets is halted when the root-mean-square error (RMSE) of the cross-validated predictions for the entire data set reaches a minimum. The prediction of a model is given by the mean of the results for the 10 nets. Empirically we find that a reliable estimate of the prediction error[32,33] of an individual molecule can be given by 3 times the standard deviation of the results of the 10 neural nets for that species.

As a result, all compounds can be used for training and testing, and by applying 10 ANNs to any compound of a validation set and by analyzing the deviations of the predictions of the 10 ANNs, one gets a quality parameter that is an indicator for molecules that are far from the training set compounds.

## EXPERIMENTAL METHODS

We apply two assays for solubility, a manual solubility assay and an automated high-throughput (HT) solubility assay: In the manual assay, the amount of compound used depends on the maximal solubility to be measured, usually up to 500 mg/L. From a solution of 50 $\mu$g of compound per $\mu$L of DMSO, a 10 $\mu$L aliquot is diluted with an aqueous buffer, usually PBS (phosphate buffer saline) at pH 6.5 to end up with 1% of DMSO in solution, hence giving a nominal solubility of 500 $\mu$g/mL. This solution is agitated at room temperature for 24 h.

After ultracentrifugation at 220 000g for 30 min, an aliquot of the supernatant is diluted with DMSO (1/5 and 1/100) and analyzed in a HPLC system together with four calibration points of the compound in DMSO at 0, 0.1, 2.5, and 20 mg/L. Individual HPLC methods are developed for each compound.

For the HT solubility assay, solutions of 50 $\mu$g/$\mu$L in DMSO are prepared. A 10 $\mu$L aliquot of this solution is diluted with an aqueous buffer, usually PBS (phosphate buffer saline) at pH 6.5 to end up with 1% of DMSO in solution, and the solution again is agitated at room temperature for 24 h.

After ultracentrifugation at 220 000g for 30 min, an aliquot (150 $\mu$L) of the supernatant is diluted with DMSO (1/5 and 1/100) and analyzed by HPLC together with three calibration points of the compound in DMSO at 0, 2.5, and 20 mg/L. The acidic or basic nature of the compound is assessed prior to the analysis by a computer algorithm. The algorithm

**Table 2.** ANN Performance Statistics Summary for the ANN Models[b]

| | | | | | test set | | | validation set | | |
|---|---|---|---|---|---|---|---|---|---|---|
| model | data set | no. of compds | descr. select. | ANN geometry | $r^2_{cv}$ | RMSE | MaxUE | $r^2$ | RMSE | MaxUE |
| 0 | DS0 | 559 | all[a] | 15−10−1 | 0.88 | 0.51 | 1.67 | | | |
| 0 | DS1 | 4889 | all[a] | 15−10−1 | | | | 0.03 | 0.44 | 5.78 |
| 1 | DS1 | 4889 | FIRM | 17−15−1 | 0.48 | 0.74 | 2.93 | 0.55 | 0.69 | 2.52 |
| 2 | DS2 | 2163 | FIRM | 11−10−1 | 0.44 | 0.65 | 2.52 | 0.50 | 0.51 | 2.32 |
| 3 | DS3 | 4806 | FIRM + manual | 38−15−1 | 0.50 | 0.78 | 3.11 | | | |
| 3 | DS3 | 4806 | FIRM + manual | 38−15−1 | 0.50 | 0.72 | 2.96 | 0.50 | 0.55 | 2.96 |
| 3 | DS4 | 1064 | manual | 38−15−1 | | | | 0.27 | 0.82 | 3.54 |
| 3 | DS5 | 7222 | manual | 38−15−1 | | | | 0.38 | 0.73 | 2.72 |

[a] AM1 gas-phase geometries. [b] Given are the descriptor selection method, the ANN geometry (input, hidden, output layer nodes), and correlation coefficient $r$, root-mean-square error RMSE, and maximum unsigned error MaxUE. For the definition of test and validation sets see text.

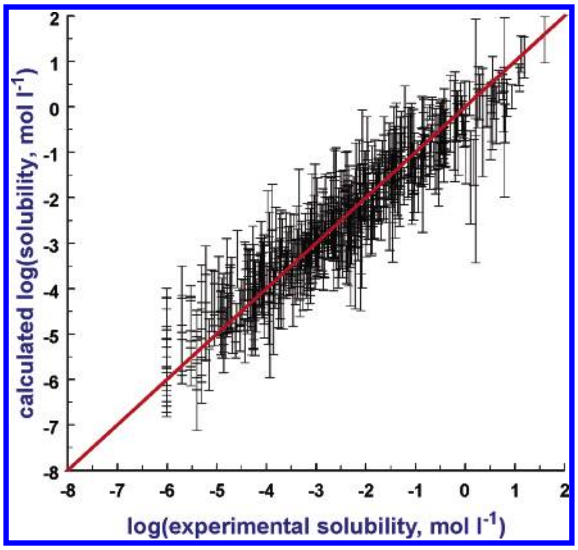


**Figure 1.** Model 0 correlation plot for 559 training set compounds from the Aquasol database ($r^2_{cv}$ = 0.88, RMSE 0.51, MaxUE 1.67).

counts the number of strongly acidic and basic functionalities in the molecule via a set of Unity substructure queries. Based on this, two gradient HPLC methods are used: one for acidic and neutral compounds with phosphoric acid in the water phase (pH: 2) and one for basic compounds with an aqueous phase containing 5 mL/L perchloric acid as counterion.

Experimental solubilities are given in units of mg/L.

## RESULTS AND DISCUSSION

**Application of an Existing Solubility ANN.** The starting point for our investigations was an unpublished ANN solubility model (Model 0, Table 2) by Clark et al.[17] based on semiempirical descriptors and **DS0**, created analogously to the models described herein (Figure 1). We expected Model 0 to be a good starting point for our purposes. However, this was not the case. On the training data set, Model 0 performed extremely well, with a correlation coefficient $r^2_{cv}$ of 0.88, a root-mean-square error (RMSE) of 0.51, a maximum unsigned error (MaxUE) of 1.67 log units, and a slope of 1.64. The same ANN applied to **DS1** with 5675 in-house structures revealed the disastrous cloud shown in Figure 2 with an $r^2$ of only 0.03 and a MaxUE of 5.78 log units. (Model statistics are given in Tables 2 and 3.) A critical re-evaluation beyond simple statistical descriptors reveals that even for the training set compounds of Model 0, the prediction quality is not convincing. Thirty-five percent of the training set compounds lie outside the error bars that define consensus score reliability. Moreover, for the validation set only 52% of the compounds lie in a trust region < 1 log unit.

We suspected that protonation states would be a major source of unreliability of Model 0, since most of its training set compounds were neutral, which is not true for our compounds. Therefore, we measured buffer solubilities for 19 out of 22 drug compounds from the Yalkowsky data set that were available to us. For the 11 neutral compounds (Table 4, Chart 1) we find an excellent agreement between water and buffer solubilities. Compounds charged under physiological pH are much more soluble than in water, and therefore we find strong deviations between the two experimental results, as can be seen in Figure 3. Since about 63% of the drug candidates are charged species,[34] we had to create our own model to predict solubilities of compounds with pharmaceutically relevant properties.

**ANN Model Creation.** In the first attempt (Model 1) we applied the iterative decision-tree based descriptor selection process which gives a minimum number of predictive descriptors. The best ANN model based on **DS1** with 4889 compounds consists of 17 descriptors (derived from Corina 3D structure and AM1/COSMO wave function) that are described in Table 1:

$$\log S_{pred}\,(\text{Model 1}) = f_{ANN}(\text{QsumH}, \text{QsumN}, \text{QsumO}, \bar{V}_S^{+}, \sigma^2_{tot}, \epsilon_A, \epsilon_B, q+, q\text{-}, \text{nAryl}, \text{globularity}, pV_{S,<-30}, pV_{S,>30}, \log P_{pH2.3}, \log P_{pH7.5}, \text{flexibility}, \text{nRingAtoms})$$

On the test set, it has an $r^2_{cv}$ of 0.48 with an RMSE of 0.74 and a maximum unsigned error of 2.93 log units. Using the same compounds in the validation mode, we obtain a correlation coefficient $r^2$ = 0.55, RMSE = 0.69, and a MaxUE of 2.52 log units (Table 2).

The improvement becomes even more obvious when looking at the trust regions. For test and validation runs, the percentages of compounds with a prediction error less than 1 log unit are 83 and 86% and less than 1.5 log units are 96 and 97%, respectively. More than 99% are always predicted better than 2 log units, which would still allow for classification into high, medium, and low solubility.

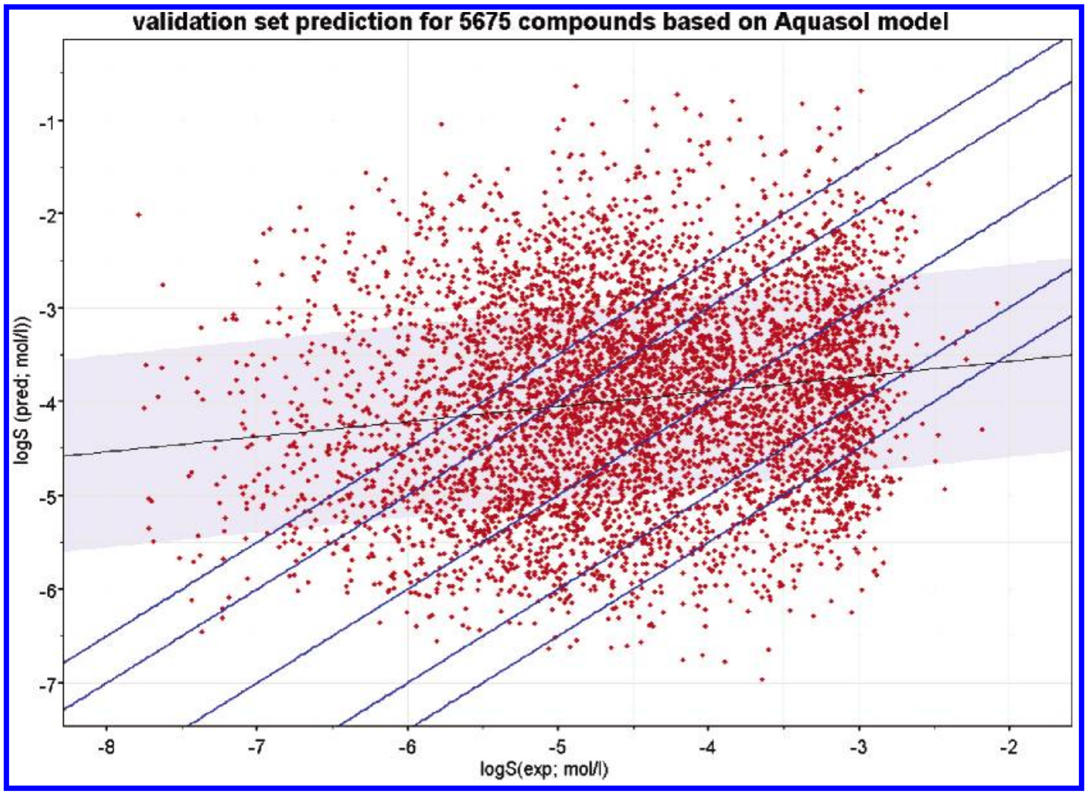


**Figure 2.** Model 0 correlation plot predicted vs experimental logS for 4889 compounds ($r^2 = 0.03$, RMSE = 0.44, MaxUE = 5.78). The blue lines indicate perfect correlation and the trust regions of “1.0 and ”1.5 log units, respectively. The black line is the actual regression, and the region of two standard deviations is shown as the gray area.

**Table 3.** Percentages of Compounds in Trust Regions of Lower than 1, 1.5, and 2 log Units

| model | mode | no. of compds | ΔlogS < 1.0 log unit | ΔlogS < 1.5 log units | ΔlogS < 2 log units |
|---|---|---|---|---|---|
| 0 | validation | 5675 | 52% | 71% | 84% |
| 1 | test | 4889 | 83% | 96% | 99% |
| 1 | validation | 4889 | 86% | 97% | 99% |
| 2 | test | 2163 | 87% | 98% | 99% |
| 2 | validation | 2163 | 88% | 98% | 99% |
| 3 | test[a] | 4806 | 83% | 96% | 99% |
| 3 | test[b] | 4806 | 86% | 98% | 99% |
| 3 | validation[b] | 4806 | 87% | 96% | 99% |
| 3 | validation | 1064 | 79% | 93% | 98% |
| 3 | validation | 7222 | 83% | 96% | 99% |

[a] Original experimental data. [b] Remeasured experimental data. For the definition of test and validation sets see text.

We then tried to optimize the predictability with the aim to optimize the percentages of compounds in the trust region, the correlation coefficient, and the standard deviation. For that purpose, we systematically analyzed the influence of ANN geometries by varying the numbers of descriptors (via different variable selection methods) and hidden layer nodes and of the source of 3D-coordinates (Corina, AM1, AM1/COSMO), as described in the Methods section. Overall, there was no major improvement achievable. ANNs based on gas-phase wave function and/or suboptimal descriptor sets performed with $r^2_{cv}$ of maximal 0.40, RMSE of at least 0.81, MaxUE higher than 3.8, and a maximal 80% of the compounds in the trust region less than one log unit (which still is considerably better than by assigning the mean value to any compound). On the other hand, there are some ANNs with a very similar performance to Model 1. The reason for not being considerably better than Model 1 is to be attributed to the uncertainty of the experimental data, the inherent errors of the quantum chemical method, and the descriptors itself. Improvements of semiempirical methodology and our future ability to derive more precise experimental data allows us to realistically plan for a next-generation model with improved predictability, as will be shown in the last paragraph.

During the systematic search just described we became aware of a subset **DS2** consisting of the 2163 compounds measured after the move from manual to automated high-throughput (HT) solubility assay, with worse model 1 predictivity. We were able to obtain Model 2 based on **DS2**, Corina 3D structure and AM1/COSMO wave function which has an 11−10−1 ANN architecture:

$$\text{logS}_{\text{pred}}\,(\text{Model 2}) = f_{\text{ANN}}\,(\text{QsumH, QsumN}, \bar{V}_S^{\,+}, \sigma^2_{+}, q+, \text{cohindex, MW}, nV_{S,<-30}, \text{logP}_{\text{pH2.3}}, \text{flexibility, nRingAtoms})$$

Its statistics are $r^2_{cv}$ of 0.44 and 0.50, RMSE of 0.65 and 0.51, and MaxUE of 2.52 and 2.32 for test and validation mode, respectively. Eighty-seven (test) and 88% (validation) of the compounds lie in the trust region lower than one and 98 and 98% lower than 1.5 log units.

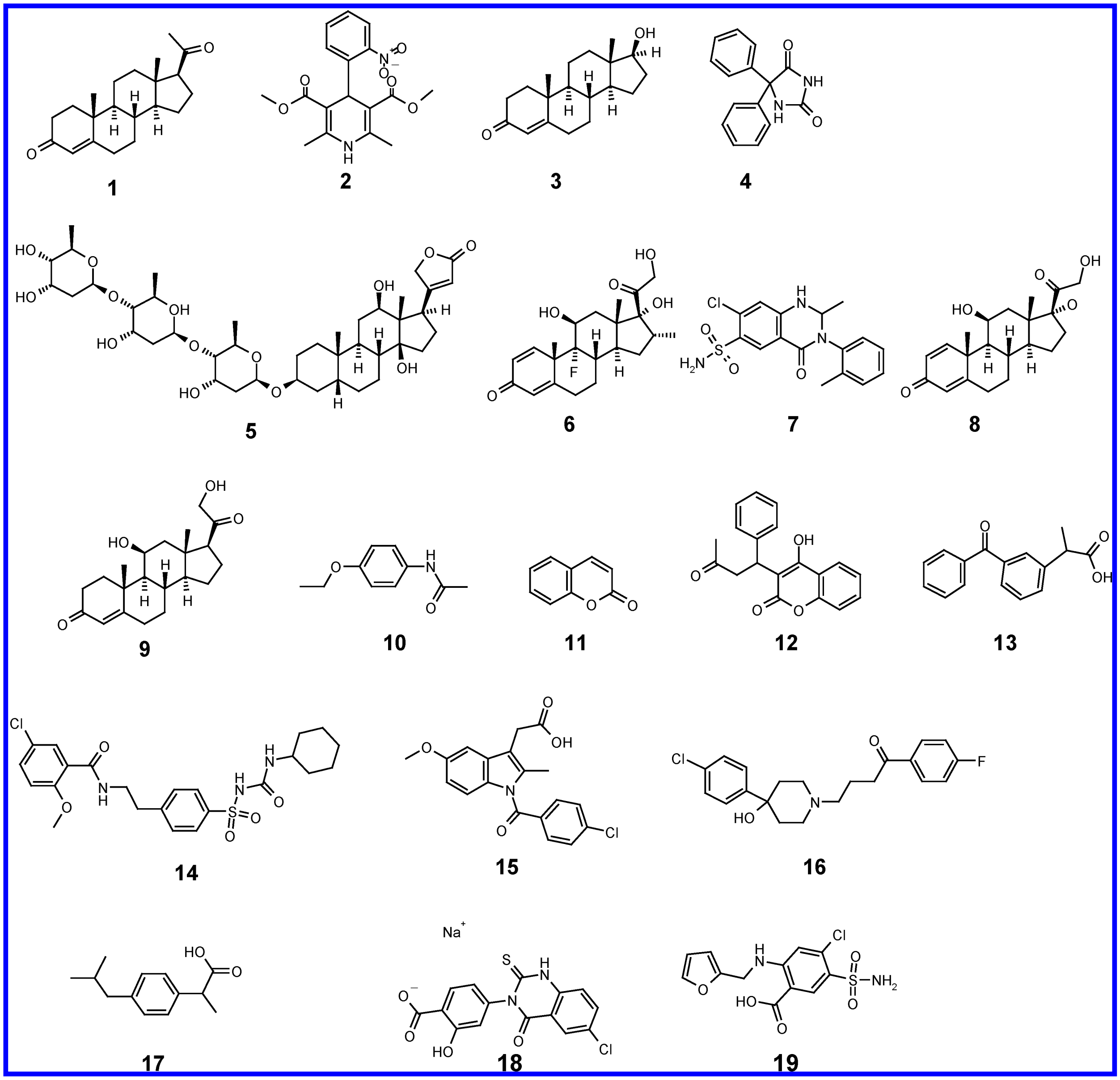
Model 2 is superior to Model 1 only for this subset. For the other experimental data that were derived semiautomatically, the RMSE is even worse with 0.80, as is the overall performance of Model 2 for **DS1**. Deeper analysis of the compounds in **DS2** revealed that not the change in the

**Table 4.** Aqueous and Buffer Solubilities (In-House) for 19 Drug Compounds from the Publication of Yalkowsky et al.[11a] [a]

| compd | name | water sol. | buffer sol. neutral compds | compd | name | water sol. | buffer sol. charged compds |
|---|---|---|---|---|---|---|---|
| 1 | progesterone | 12.0 | 6.0 | 11 | coumarin | 2700.0 | 2900.0 |
| 2 | nifedipine | 6.0 | 9.0 | 12 | warfarin | 40.0 | 235.0 |
| 3 | testosterone | 28.0 | 20.0 | 13 | ketoprofen | 140.0 | 275.0 |
| 4 | phenytoin | 26.0 | 23.0 | 14 | glyburide | 4.0 | 1.1 |
| 5 | digoxin | 54.0 | 31.0 | 15 | indomethacin | 8.6 | 240.0 |
| 6 | dexamethasone | 96.0 | 59.0 | 16 | haloperidol | 14.0 | 180.0 |
| 7 | metolazone | 60.0 | 88.0 | 17 | ibuprofen | 36.0 | 290.0 |
| 8 | prednisolone | 240.0 | 140.0 | 18 | piroxicam | 23.0 | 92.0 |
| 9 | corticosterone | 200.0 | 280.0 | 19 | furosemide | 72.0 | 200.0 |
| 10 | phenacetin | 800.0 | 1080.0 | | | | |

[a] Eleven compounds are neutral, and 8 compounds are charged under physiological conditions.

**Chart 1**



experimental setup made up the difference but a shift in project portfolio and following a shift in chemical structure classes. Since our model has to be as general and flexible as possible to cope with hopefully any new structure class in the future, Model 1 is better suited than Model 2. On the other hand, we were able to demonstrate that our approach is able to generate not only global solubility models but also local substance class based ones with slightly superior predictability.

The next step was outlier analysis as described in the Data Sets section. By this, we remeasured 40 compounds and excluded another 83 compounds due to obviously wrong

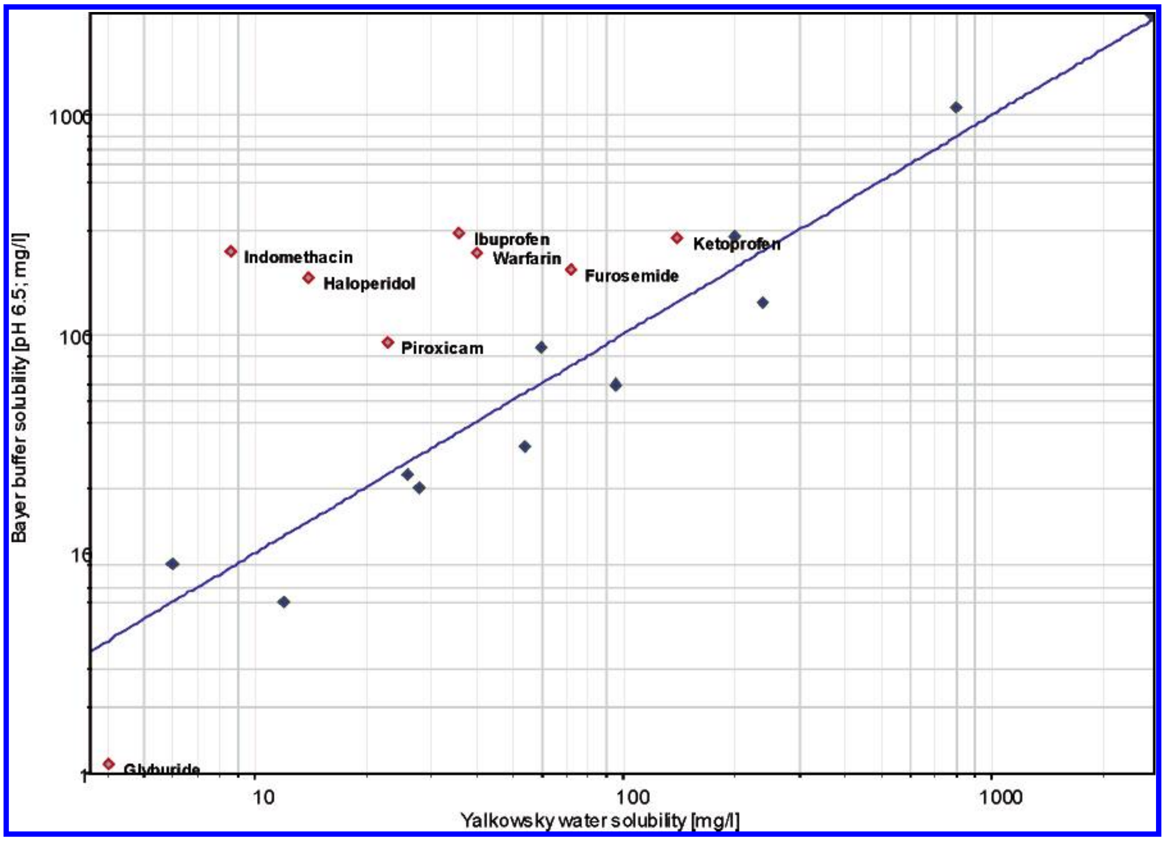


**Figure 3.** Correlation between experimental solubilities in water and buffer medium (pH 6.5) for neutral (blue) and charged (red) compounds. The respective structures and values are given in Table 4.

values, yielding 16 new strongly deviating values and **DS3** with 4806 compounds.

$$\log S_{pred}\ (\text{Model 3}) = f_{ANN}(\mu, \text{dipden}, \text{m0pol}, \text{QsumH}, \text{QsumN}, \text{QsumO}, V_{S,max}, V_{S,min}, \bar{V}_S{}^{+}, \bar{V}_S{}^{-}, \bar{V}_S, \sigma^2{}_{+}, \sigma^2{}_{-}, \sigma^2{}_{tot}, \nu, \nu\,\sigma^2{}_{tot}, \pi, \epsilon_A, \epsilon_B, q+, q\text{-}, \text{nAcc}, \text{nDon}, \text{nAryl}, \text{cohindex}, \text{MW}, \text{volume}, \text{surface}, \text{globularity}, nV_{S,<-30}, nV_{S,[-30,30]}, nV_{S,>30}, pV_{S,<-30}, pV_{S,[-30,30]}, pV_{S,>30}, \log P_{pH2.3}, \log P_{pH7.5}, \text{flexibility})$$

We found two ANN by FIRM and manual descriptor selection, with 20−18−1 and 38−15−1 architectures, respectively. The latter was more stable in test runs, with $r^2_{cv} = 0.50$, RMSE = 0.72, and MaxUE = 2.96, compared to RMSE = 0.78 and MaxUE = 3.11 for the original experimental values. In the validation run, this comes down to $r^2 = 0.50$, RMSE = 0.55, and MaxUE = 2.96. Figure 4 shows the test run correlation. The yellow squares symbolize the 16 compounds' original values, and the red diamonds symbolize their new values. For some compounds the shifts are dramatic.

The percentages of compounds inside the trust regions of less than 1 and 1.5 log units are 83 and 96% for the test run and 86 and 98% for the validation run, respectively.

ANNs nonlinearly connect the input values to the dependent variable. Nevertheless, a closer look at the descriptors for Model 3 resembles that they can be grouped and be connected to the molecular properties that influence solubility.[9] Molecular shape, which affects packing and solvent interactions, can be described through the geometry descriptors molecular weight, volume, surface, globularity, and flexibility. The charged partial surface areas that play an important role in solvent−solute interactions are manifested by the surface area counts for the molecular electrostatic potential ($V_{S,max}$, $V_{S,min}$, $\bar{V}_S{}^{+}$, $\bar{V}_S{}^{-}$, $\bar{V}_S$, $V_{S,<-30}$, $nV_{S,[-30,30]}$, $nV_{S,>30}$, $pV_{S,<-30}$, $pV_{S,[-30,30]}$, $pV_{S,>30}$) and the Qsum parameters. Permanent and induced polarity is encoded in dipole moment $\nu$, dipole density dipden, and polarizability m0pol. The variance and balance of charge distribution in the

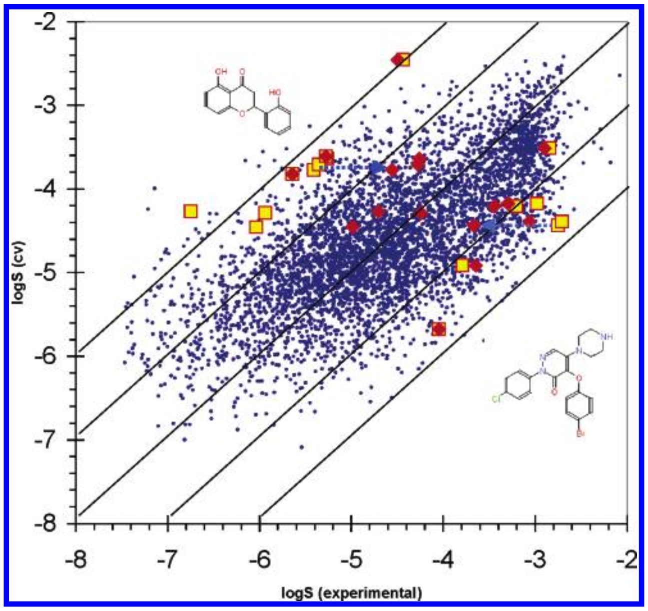


**Figure 4.** Model 3 predicted (test mode; $r^2_{cv} = 0.50$, RMSE = 0.72, MaxUE = 2.96) vs. experimental logS. Original values are shown as yellow boxes and remeasured ones as red diamonds.

molecule, which accounts for locality of strongly polar centers is given via $\sigma^2_{+}$, $\sigma^2_{-}$, $\sigma^2_{tot}$, $\nu$, $\nu\ \sigma^2_{tot}$, $\pi$, whereas hydrogen acceptor and donor capabilities are given via the counts for hydrogen donors (nDon) and acceptors (NAcc) and the covalent and electrostatic hydrogen bond acidities and basicities ($\epsilon_{A,}$ $\epsilon_{B,}$ q+, q−). Additionally, we have included an indirect experimentally based descriptor for the protonation state, namely the $\log P_{pH2.3}$ and $\log P_{pH7.5}$. Finally, the hydrophobic interactions are accounted for by the cohesive index (cohindex) and the number of aryl rings (nAryl).

None of the descriptors of any of these groups alone seems to be sufficient to account for features describing solubility, but combinations of these which always describe a part of the physical phenomenon are able to yield into a predictive model.

The relatively low slopes of all models' curves result in a model prediction of low experimental solubilities as too high and high experimental solubilities as too low. Low experimental values have a higher experimental error, especially in the HT apparatus, which has an inherent low boundary of 0.1 mg/L (any value below will be reported as 0.1 mg/L). Therefore, the low values in our training set stem from manual measurement, in fact. High values are, on the other hand, very much influenced by the actual crystallization state of the compound, the presence of counterions, or cocrystallized solvents. Therefore, the model gives a value also for low soluble compounds, where the experiment gives only the lower bound, thus yielding sometimes large MaxUE and lowering the overall correlation, even if the model-derived value is reasonable. For the highly soluble compounds, our model will always predict a value for the "clean" compound and will not be able to predict the additional effects from experimental conditions. This will even become clearer when looking at the predictions made on the two further validation sets.

We find two other subsets of compounds with a much higher MaxUE than on average: first, compounds that have a counterion or salts or mixtures and second, compounds that contain the element sulfur, with almost any chemical group such as sulfide, sulfonamide, sulfoxide, or thiophene. For the sulfur problem we have taken various unsuccessful attempts such as adding sulfur count descriptors or building sulfur submodels. This can be attributed to a known deficiency of classical semiempirical NDDO approaches such as AM1 or PM3 to cope with hypervalent atoms from the third row of the PSE. We will now attack this problem again with the newly developed AM1* Hamiltonian,[35] an extended implementation of AM1 with an explicit inclusion of d-orbitals.

**Application to Validation Sets.** We now applied our model on two additional data sets. **DS4**, 1064 compounds intentionally removed from the original data set, consists of permanently charged species such as quaternary amines, zwitterions, ambiguous tautomers, or diastereomers with unknown absolute stereochemistry. The predictivity with Model 3 is somewhat lower with $r^2 = 0.27$, RMSE = 0.82, MaxUE = 3.54, and the trust region populations are worse with 79, 93, and 98% inside 1, 1.5, 2 log units but still good enough to at least classify these compounds and find out the ones which bear the biggest ADME risks in a lead optimization project.

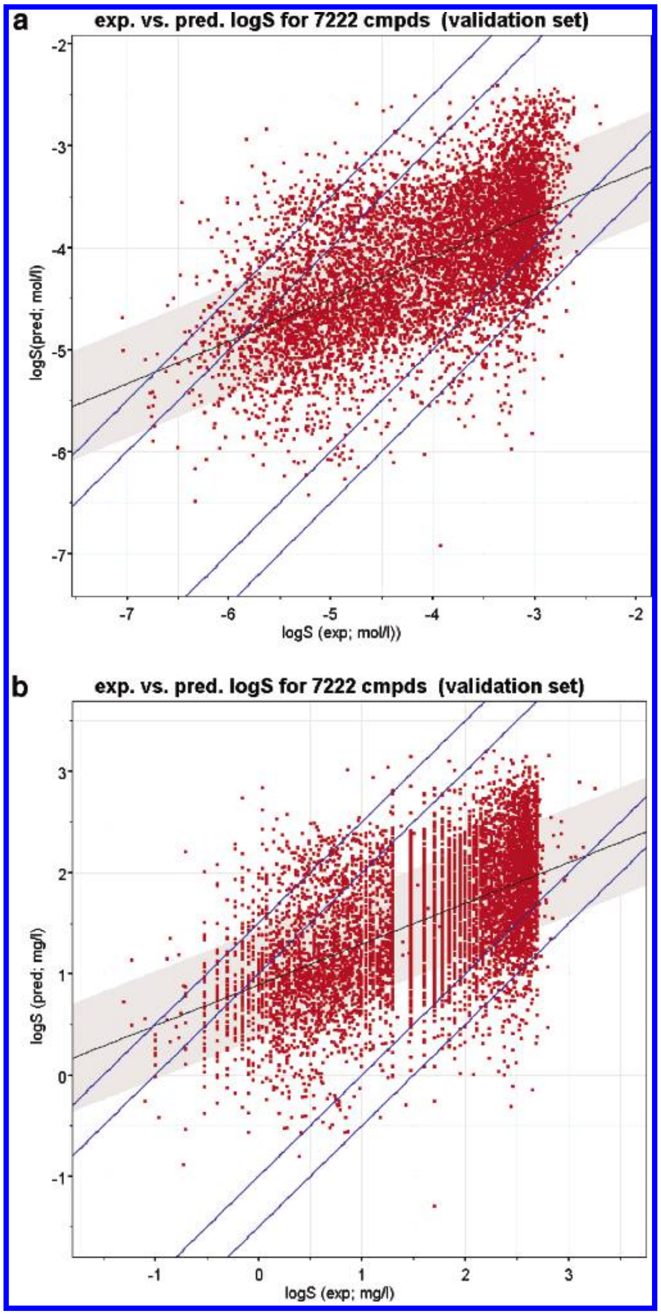


**Figure 5.** (a) Model 3 predicted vs experimental logS for **DS5** with 7222 compounds ($r^2_{cv} = 0.38$, RMSE = 0.73, MaxUE = 2.72). The blue lines indicate the trust regions of less than 1.0 and 1.5 log units, respectively. The black line is the actual regression, and the region of two standard deviations is shown as the gray area. (b) Model 3 predicted vs experimental logS for **DS5** with 7222 compounds, logS derived from S in units of mg/L as used at Bayer by default ($r^2_{cv} = 0.31$, RMSE = 0.53, MaxUE = 3.00). The blue lines indicate the trust regions of less than 1.0 and 1.5 log units, respectively. The black line is the actual regression, and the region of two standard deviations is shown as the gray area.

Figure 5a finally gives the analogous information for **DS5**, 7222 compounds measured after finalization of Model 3. The statistical parameters are $r^2 = 0.38$, RMSE = 0.73, MaxUE = 2.72, and the trust region populations are good with 83, 96, and 99% inside 1, 1.5, and 2 log units.

Since the chemists are used to values in mg/L, thus taking into account the molecular weight, the model should be able to predict reliably in this scale, too, as demonstrated in Figure 5b. The statistical parameters are somewhat lower with $r^2$

= 0.31, RMSE = 0.53, MaxUE = 3.00. The trust region populations are 81, 95, and 99% inside 1, 1.5, and 2 log units, clearly suggesting to retrain the model in the logS (mg/L) scale instead of the current logS (mol/L) scale.

Therefore, Model 3 is a solid basis for the in silico prediction of buffer solubilites even for the bad cases, i.e., the permanently charged, zwitterionic, and ambiguous structures as well as sulfur containing compounds, with a probability of larger than 80% for being not worse than a factor of 10, about 96% for being not worse than a factor of 32, and 99% to be able to classify and automatically sort out the "stones" (insoluble compounds) early.

Model 3 is a fast and robust method for buffer solubility prediction, taking about 15 s per molecule on an Intel Xeon P4 2.8 GHz CPU. Applied to our corporate database, it has a success rate of about 99.9%, the failures mostly due to inconsistencies in the registered structures.

## CONCLUSIONS

In silico ADME models can help to filter out, prioritize, and risk-assess compounds in the early stage of drug discovery, following the "fail early, fail cheap" principle. One of the critical parameters for this assessment is solubility. The published as well as the commercially available solubility models deal with the prediction of aqueous solubility, in contrast to the buffer solubilities needed by the pharmaceutical industry to mimic the human intestinal tract medium. Additionally, many of these models are based on nondruglike compounds and data sets derived from various sources with differing experimental standards, adding some more noise to the already sparse data.

We therefore developed a broadly applicable buffer solubility model on the basis of about 5000 experimental values derived in a consistent manner from the same laboratory under identical conditions.

The model is based on descriptors derived from AM1 semiempirical quantum chemical calculations, together with two in-house developed HQSAR-based logP descriptors for pH 2.3 and 7.5, and topological descriptors accounting for molecular features such as hydrogen bond donors and acceptors and polar and lipophilic surface areas. Backpropagation artificial neural network (ANN) models were built by systematic variations of input and hidden layer architecture after iterative application of manual and automatic descriptor selection schemes. The ANN training procedure automatically divides the data set into 10 equal subsets, training 10 ANNs with always nine subsets as training and the remaining subset as a test set. The final model gives 10 predictions and results in a mean solubility value and a quality parameter for overall predictivity of that set of ANNs.

We were able to show that Corina 3D structures and AM1/COSMO solvent polarized wave functions are adequate to describe the chemical entities. Conformational flexibility was explicitly not taken into account. The model has proven to work in high-throughput, with a mean calculation time of about 15 s per molecule and a failure rate of about 0.1%.

We were able to derive a global model which is able to predict more than 80% of the compounds with an accuracy of lower 1 log unit and about 96% with an accuracy of lower 1.5 log units. More than 99% are better than 2 log units. The model will always be at least as accurate to classify into low, medium, and high solubility. The model is able to predict buffer solubilites for permanently charged molecules and formally neutral salts and mixtures. Thereby, it implicitly takes into account charged states of compounds that were presented to the ANN as formally neutral, as are about 63% of current druglike compounds. The successful implicit inclusion of protonation states can also be seen in the case of permanently charged quaternary amines and zwitterions, which are predicted quite well (83% within 1 log unit), but not as good as more obvious cases.

To summarize, we have established a high-quality buffer solubility model for the in silico prediction of the solubilities derived from our experimental setup.

## REFERENCES AND NOTES

(1) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of Aequeous Solubility Based on Large Datasets Using Several QSPR Models Utilizing Topological Structure Representation. *Chem. Biodiversity* **2004**, *1*, 1829−1841.

(2) (a) Jensen, F. In *Introduction to computational chemistry*; Wiley: Weinheim, 1999. (b) Leach, A. R. In *Molecular Modelling: Principles and Applications;* Eddison Wesley: Longham, Harlow, 1996.

(3) Yalkowsky, S. H. In *Solubility and Solubilization in Aqueous Media*; American Chemical Society and Oxford University Press: Washington, DC, New York, Oxford, 1999.

(4) (a) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25. (b) Jorgensen, W. L.; Duffy, E. M. Prediction of Drug Solubility from Structure, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 355−366. (c) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A. A Fuzzy ARTMAP Base on Quantitative Sructure-Property Relationships (QSPRs) for predictive Aqueous Solubility of Organic Compounds *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177−1207.

(5) (a) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organci Compounds based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434, and references therein. (b) Butina, D.; Gola, J. M. R. Modelling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837−841, and references therein.

(6) Huuskoonen, J.; Salo, M.; Taskinen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 473−477.

(7) Yalkowsky, S. H.; Damnelfelser, R. M. *The ARIZONA dATAbASE of Aqueous Solubility*; College of Pharmacy, University of Arizona: Tuscon, AZ, 1990.

(8) Syracuse Research Corporation. *Physical/Chemical Property Database (PHYSPROP)*; SRC Environmental Science Center: Syracuse, NY, 1994.

(9) (a) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247. (b) 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. (c) Danauskas, S. M.; Jurs, P. C. Prediction of C60 Solubilities from Solvent Molecular. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 419−424.

(10) Bergström, C. A. S.; Wassvik, C. M.; Norinder, U.; Luthman, K.; Artursson, P. Global and Local Computational Models for Aqueous Solubility Prediction of Drug-like Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1474−1488.

(11) (a) Yalkowsky, S. H.; Valvani, S. C.; Rosemann, T. J. Solubility and Partitioning VI: Octanol Solubility and Octanol−Water Partition Coefficients. *J. Pharm. Sci.* **1983**, *72*, 866−870. (b) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Octanol/Water Partition Coefficients. *Environ. Toxicol. Chem.* **1996**, *15*, 100−106.

(12) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855−1859.

(13) 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.

(14) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.

(15) Cheng, A.; Merz, K. M., Jr. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure−Property Relationships. *J. Med. Chem.* **2003**, *46*, 3572−3580.

(16) 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.

(17) Beck, B.; Clark, T. unpublished results.

(18) Sadowski, J.; Schwab, C.; Gasteiger, J. *Corina 3.1*; Molecular Networks GmbH Computerchemie: Nägelsbachstrasse 25, 91052 Erlangen, Germany.

(19) The deficiencies of Corina 2.64 used during during model creation are fixed with the current version 3.1.

(20) Clark, T.; Alex, A.; Beck, B.; Burckhardt, F.; Chandrasekhar, J.; Gedeck, P.; Horn, A.; Hutter, M.; Martin, B.; Rahut, G.; Sauer, W.; Schindler, T.; Steinke, T. *VAMP 8.0*; Erlangen, 2001.

(21) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. A New Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(22) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(23) (a) Rauhut, G.; Clark, T. Multicenter Point-Charge Model for High-qulaity Molecular Electrostatic Potentials from AM1 Calculations. *J. Comput. Chem.* **1993**, *14*, 503−509. (b) Beck, B.; Rauhut, G.; Clark, T**.** The natural atomic orbital point charge model for PM3: Multipole moments and molecular electrostatic potentials**.** *J. Comput. Chem.* **1994**, *15*, 1064−1073.

(24) (a) Beck, B.; Horn, A.; Carpenter, J. E.; Clark, T. Enhanced 3D-Databases: A Fully Electrostatic Database of AM1-Optimized Structures. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1214−1217.

(25) Unpublished models established at Bayer AG, main contributors Beck, M. E.; Bürger, T. (a) logP@pH = 2.3: HQSAR (Tripos, see citation 26) model based on 70 000 compounds with experimental data in buffer at pH = 2.3; $r^2 = 0.76$, STD = 0.60, $q^2 = 0.76$, $\mathrm{STD_{CV}} = 0.60$. (b) logP@pH = 7.5: HQSAR (Tripos, see citation 26) model based on 7000 compounds with experimental data in buffer at pH = 7.5; $r^2 = 0.85$, STD = 0.62, $q^2 = 0.83$, $\mathrm{STD_{CV}} = 0.67$.

(26) *Sybyl 6.9.2, Unity 4.4.2*; Tripos Inc.: 1699 South Hanley Rd., St. Louis, Missouri, 63144 U.S.A.

(27) *Cerius$^2$ 4.9*; Accelrys Inc.: 9685 Scranton Rd., San Diego, CA 92121, 2003.

(28) (a) Hawkins, D. M. *FIRM*. http://www.stat.umn.edu/users/FIRM/index.html; (b) DIVA, Accelrys Ltd., 10199 Telesis Court, Suite 100, San Diego, CA 92121, U.S.A.

(29) FIRM is different from common recursive partitioning methods in that it can divide on any hierarchy into up to 10 subsets (default), therefore better representing the nonbinary nature of most data sets. It can work on continuous as well as categorical independent and dependent variables. FIRM tries to split at any level based on multibin separability, subset size, and subset significance. It applies statistical analyses on variance and relevance of the resulting sets. In any subset, it will stop if the resulting sets on the next level would be too small to be significant.

(30) (a) Müller, B.; Reinhardt, J.; Strickland, M. T. In *Neural Networks - An Introduction*, 2nd ed.; Springer-Verlag: Berlin, Heidelberg, 1995. (b) Pao, Y.-H. *Adaptive Pattern Recognition and Neural Networks*; Addison-Wesley Publishing Co.: Reading, MA, 1989. (c) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Verlag: Weinheim, Germany, 1993.

(31) Chalk, A. J.; Beck, B.; Clark, T. A Temperature-dependent Quantum Mechanical/Neural Net Model for Vapour Pressure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1053.

(32) Beck, B.; Breindl, A.; Clark, T. QM/NN QSPR Models with Error Estimation: Vapor Pressure and LogP. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1046−1051.

(33) Chalk, A. J.; Beck, B.; Clark, T. A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 457−462.

(34) Box, K.; Comer, J.; Hill, A.; Tam, K.; Trowbridge, L. High-throughput physicochemical profiling Part 1: Rapid Ionization Constants (pKa) Determination; Poster at logP Symposium, Lausanne, Switzerland, 2000.

(35) Winget, P.; Horn, A. H. C.; Selcuki, C.; Martin, B.; Clark, T. AM1* parameters for phosphorus, sulfur and chlorine *J. Mol. Model.* **2003**, *9*, 408.

(36) Mu, L.; Drago, R. S.; Richardson, D. E. A model based QSPR analysis of the unified nonspecific solvent polarity scale. *J. Chem. Soc., Perkin Trans. 2* **1998**, 159−167.

(37) (a) Rinaldi, D.; Rivail, J. L. *Theor. Chim. Acta* **1974**, *32*, 243−251. (b) Rinaldi, D.; Rivail, J. L. *Theor. Chim. Acta* **1974**, *32*, 57−70.

(38) Schürer, G.; Gedeck, P.; Gottschalk, M.; Clark, T. Accurate Parametrized Variational Calculations of the Molecular Electronic Polarizability by NDDO-Based Methods. *Int. J. Quantum Chem.* **1999**, *75*, 17−31.

(39) B.; Glen, R. C.; Clark, T. VESPA: A new, fast approach to electrostatic potential-derived atomic charges from semiempirical methods. *J. Comput. Chem.* **1997**, *18*, 744−756.

(40) (a) Murray, J. S.; Politzer, P. Statistical analysis of the molecular surface electrostatic potential: an approach to describing noncovalent interactions in condensed phases. *J. Mol. Struct. (THEOCHEM)* **1998**, *425*, 107−114. (b) Murray, J. S.; Lane, P.; Brinck, T.; Paulsen, K.; Grince, M. E.; Politzer, P. Relationships of critical constants and boiling points to computed molecular surface properties. *J. Phys. Chem.* **1993**, *97*, 9369−9373.

(41) Cronce, D. T.; Famini, G. R.; DeSoto, J. A.; Wilson, L. Y. Using theoretical descriptors in quantitative structure−property relationships: some distribution equilibria. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293−1301.

(42) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

(43) Pascual Pascual-Ahuir, J. L.; Silla, E.; Tunon, I. Incorporation of bond-length constraints in Monte Carlo simulations of cyclic and linear molecules: Conformational sampling for cyclic alkanes as test systems. *J. Comput. Chem.* **1994**, *15*, 1127−1138.

(44) Meyer, A. Y. *Chem. Soc. Rev.* **1986**, *15*, 449−475.

(45) Beck, B. Unpublished results.

CI0503210