Insolubility Classification with Accurate Prediction Probabilities Using a MetaClassifier

Christian Kramer,[†,‡] Bernd Beck,*[,†] and Timothy Clark*[,‡,§]

Department of Lead Discovery, Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany, Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Friedrich-Alexander Universität Erlangen-Nürnberg, Nägelsbachstrasse 52, 91052 Erlangen, Germany, and Centre for Molecular Design, University of Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth, PO1 2EG, United Kingdom

Received September 28, 2009

Insolubility is a crucial issue in drug design because insoluble compounds are often measured to be inactive although they might be active if they were soluble. We provide and analyze various insolubility classification models based on a recently published data set and compounds measured in-house at Boehringer-Ingelheim. The 2D descriptor sets from pharmacophore fingerprints and MOE and the 3D descriptor sets from ParaSurf and VolSurf were examined in conjunction with support vector machines, Bayesian regularized neural networks, and random forests. We introduce a classifier-fusion strategy, called metaclassifier, which improves upon the best single prediction and at the same time avoids descriptor selection, a potential source of overfitting. The metaclassifier strategy is compared to the simpler fusion strategies of maximum vote and highest probability picking. A prediction accuracy of 72.6% on a three class model is achieved with the metaclassifier, with nearly perfect separation of soluble and insoluble compounds and prediction as good as our calculated maximum possible agreement with experiment.

INTRODUCTION

Aqueous solubility (simply described as solubility in the following) is a key issue in the early drug-discovery phase because insoluble compounds are very likely to give negative test results. While the importance of insolubility varies from project to project, up to 77% of screening compounds have been reported to lack the solubility necessary for testing.[1] This leads to wasted time and money if insoluble compounds are not excluded from the screening data set or marked and interpreted taking solubility problems into account.

The importance of solubility has prompted many groups, both industrial and academic, to work on in silico models that predict solubility.[2–24] Simple relationships have been found for homologous series of similar compounds (e.g., alkanes), but no such simple treatment exists for the more general case of a wide variety of compounds. Many different common machine-learning methods have been used to predict solubility, from multiple linear regression[25,26] to partial least-squares,[27,28] neural networks and ensembles thereof,[23,29] support vector machines (SVM),[30] random forests (RF),[19] and Gaussian processes.[21] These algorithms have been combined with different descriptor sets calculated by programs such as Dragon[31] or the Molecular Operating Environment (MOE)[32] and a variety of descriptor-selection algorithms.

Machine-learning methods in computational chemistry have reached a point at which several different high-end algorithms are available that give similar overall performance but differ in detail. For example, RFs and SVMs have been shown to give similar correlation coefficients (R^2) between predicted and experimental results on various data sets; nevertheless they give different individual predictions.[33] The choice of descriptors has the same effect; without selection, different descriptor sets treated with the same method yield different predictions but similar measures of quality. This is seen, for example, for logP prediction, where many methods give very similar correlation coefficients.[34,35] The situation is similar for solubility predictions based on publicly available data sets.[21] There is a need to define methods for uniting single predictions into an overall one. This is especially necessary for cases with strongly different predictions when it is not obvious which classifier performs better for a specific compound.

A large number of regression models has been reported for solubility but only a few classification models. Manallack et al.[18] described a consensus neural network model based on BCUT descriptors for binary classification of solubility for a threshold of 100 μg/mL. They used a subset of the *PHYSPROP* database[36] and achieved an accuracy of approximately 87%. Lamanna et al.[15] recently published a recursive partitioning model based on molecular weight (MWt) and aromatic proportion (AP). They achieved an accuracy of 81% for binary classification of a test set that remained after the training set was chosen by space-filling design. The model was built and validated with company in-house data.

Nearly all of the models published so far were trained and/or validated on the *PHYSPROP* database or part thereof such as the so-called Huuskonen data set.[10] In 2007, Schwaighofer et al. showed that none of the commercial models available is able to predict in-house data satisfactorily, despite their

* To whom correspondence should be addressed. E-mail: clark@chemie.uni-erlangen.de (T.C.); E-mail: bernd.beck@boehringer-ingelheim.com (B.B.).
† Boehringer-Ingelheim Pharma GmbH & Co. KG.
‡ Friedrich-Alexander Universität Erlangen-Nürnberg.
§ University of Portsmouth.

10.1021/ci900377e
Published on Web 01/20/2010

generally good prediction performance for the Huuskonen data set.[21] In a previous paper,[37] we reported a new kinetic solubility data set with 711 compounds measured in-house at Boehringer-Ingelheim. In contrast to the Huuskonen data set, these data are suitable for early phase drug discovery solubility because the compounds were chosen to be druglike and the measurement protocol is very similar to most of the early phase preparation processes for compounds, i.e., kinetic solubility has been measured. Additionally, the data set is consistent because it was measured in one laboratory by one procedure. It is more suitable for classification models than for regression because nearly 50% of the values are given as "greater than" or "less than". Quantitative models are not necessary for early phase drug discovery. Identifying compounds that are likely to be insoluble and interpreting test results differently or even removing them from the screening data set is sufficient.

After introducing simple decision-tree models based on basic descriptors such as MWt, logP, and ionization state,[37] we now show how to improve predictions by nonlinear classifiers and their union into a metaclassifier. An important aspect of this work is to estimate the maximum possible prediction performance for a given data set. We therefore calculate this maximum possible performance as a function of the experimental standard error. We then develop single models for combinations of one algorithm from SVM, RF, and Bayesian regularized artificial neural networks (BRANN) and one descriptor set from MOE 2D descriptors (MOE2D),[32] pharmacophore fingerprints (FP),[38] and descriptors calculated by ParaSurf (PS)[39] and VolSurf (VS).[40] With 12 individual models at hand, we show how to unite these 12 models into a metaclassifier. This improves the overall prediction and removes the need to select one single classifier that might not be significantly better than the others. Finally, we show how to use the prediction probabilities obtained for in-depth analysis and "cherry picking" of predictions.

MATERIALS AND METHODS

Data Set. The training and test data set consists of 711 compounds published in ref 37 and another 131 Boehringer Ingelheim in-house compounds measured at the same time. The validation set consists of 747 in-house compounds measured in the past few years. Solubility was measured between 2 and 250 μg/mL. Values outside the range are assigned the values >250 μg/mL or <2 μg/mL. We have used this classification scheme throughout this work and also for the simple models reported with the details of the data set[37] in order to be able to judge the classification models objectively and impartially. We are thus able to judge the performance of the different classifiers quantitatively using the statistical criteria described below. Note also that the metaclassifier strategy neither changes the classification criteria nor introduces additional supervised learning steps so that it cannot lead to any kind of "self-fulfilling prophecy".

The compounds were chosen to be druglike, have a molecular weight of at least 300 g/mol, and have a ClogP lower than 6. For a complete description and analysis of the public part of the training data set see ref 37. Values of 20 and 200 μM are used as classification thresholds for the three classes of insoluble, moderately soluble, and soluble compounds used previously. Note that the characteristics and property distributions of the training data set have been described in detail.[37]

Class Membership Probabilities. QSAR/QSPR models employ estimates and strong simplifications. Thus, they provide statistical predictions and should ideally give error estimates. In the case of regression, these are standard deviations; in the case of classification, they are class membership probabilities (CMPs). These are simply the estimated probabilities that the compound has been assigned to the correct class.

Some classifiers are able to calculate probabilities for their predictions. The probability is indicative of the safeness of the prediction. Classification involves the assumption that classes really exist and an instance belongs to one class only. In the case of solubility, it is assumed that there is a real solubility, which means that a compound can belong to one solubility class only. The CMP is a measure of safeness of the prediction estimated by the classifier itself. CMPs of 0.8, 0.2, and 0.0 prediction for classes 1, 2, and 3, respectively, mean that 80 of 100 instances really should be class 1. Clearly, the reliability of the CMP increases with the number of instances. Thus, if an algorithm produces probabilities for a very small number of compounds it is usually impossible to verify them. It is highly desirable to have accurate class membership estimates, as this enables, for example, only predictions with high probabilities to be selected. Using different models that provide accurate CMPs allows predictions to be compared and the prediction with the highest probability selected.

Machine-Learning Methods. Three nonlinear machine-learning tools were used to classify insolubility. RFs, SVMs, and BRANNs have often proven to be superior to other tools, although they generally perform similarly to each other but with individual strengths and weaknesses.

SVM calculations were carried out using LibSVM[41] with an RBF kernel and the C-SVC routine. SVM as a standard algorithm in QSAR is described, for example, in ref 42. Parameters were optimized using a 50 step simplex algorithm, as this is more efficient and accurate than a grid search.[43] Scores and CMPs were evaluated based on the distance to the hyperplane. After creation of all the one-class-versus-one-class models, a sigmoidal function was fit to the distribution of distances to the hyperplane. CMPs were adjusted using 5-fold cross-validation with the training set. CMPs depend on the distance to the hyperplane; the further a substance is away from the hyperplane, the safer the prediction should be. For details on the probability fitting see ref 41. The SVM results are given for the test sets of 3-fold cross validation.

BRANN calculations were carried out using the flexible Bayesian modeling software from Radford Neal.[44] The BRANN algorithm is described in detail in ref 45. This Bayesian nature of the BRANN approach allows the CMP to be estimated directly for each compound. Automatic relevance detection is also available for BRANNs and has been used, for example, for QSAR studies of benzodiazepines and substances active at the muscarinergic receptor.[46] The BRANN results are given for the test sets of 3-fold cross-validation.

Random forest calculations were carried out using our own code. The random forest algorithm is described in detail in refs 47 and 48. Here we only briefly describe the major ideas:

In random forests, a multitude of decision trees is grown, each with a different bootstrap sample. For each split, only a subset of descriptors is evaluated, here the square root of the number of descriptors. Each tree is built until a specific stop criterion or until all leaves are pure, i.e., a maximum leaf size of one (in the original implementation). The gini-Index[49] is used as the criterion for evaluating all possible splits. For the final prediction, the votes from all decision trees are summed and the class with the most votes is chosen. We use a minimum leaf size of five (i.e., leaves that contain five or less compounds are not split further) because this avoids insignificant splits of nodes with very low populations. To unite the predictions from the different trees, we average over the training set occupation of the final leaves to derive CMPs. A total of 1 000 trees were generated for all the models described. The RF results are given for the out-of-bag test samples (i.e., ones not seen by the model before testing).

Descriptor Sets. Four different descriptor sets were used for generating the models.

Set one is the pharmacophore fingerprints descriptor set (FP) provided by ChemAxon.[38] Here, the occurrences of 210 basic pharmacophores are counted per compound.

Set two is the MOE 2D descriptor set, consisting of an internal selection of 184 different standard descriptors.[32] It comprises basic descriptors such as MWt, bond counts, atom counts, connectivity indices,[50] two logP and one logS model, BCUT descriptors,[51] GCUT descriptors, binned PEOE surface descriptors,[52] and more.

Set three is an augmented version of the ParaSurf08 descriptor set.[39] It consists of statistical measures of local surface properties plus some basic descriptors such as MWt and donor/acceptor counts, in total, 107 descriptors. The surfaces were derived from semiempirical AM1[53] wave-functions, Initial 3D molecular structures were obtained using CORINA[54] and were then geometry optimized with VAMP[55] using the AM1 Hamiltonian.[56]

Set four is the VolSurf descriptor set.[57] VolSurf descriptors are derived from interaction energies of different probes within a GRID field.[58] They are transformed to a smaller set of numerical descriptors by principal-component analysis[59] or partial least-squares. Additionally, VolSurf calculates logP and logD descriptors for different pH values and some classic descriptors such as molecular weight and polar surface area.[60] Altogether, this descriptor set consists of 90 descriptors.

Note that the ParaSurf and VolSurf descriptors, in contrast to the other descriptor sets, are based on 3D molecular structures and can therefore in principle treat intermolecular packing in crystal lattices. They can also encode information that might be used to detect polymorphism, which can influence thermodynamic solubility significantly.[17] However, polymorphism is less likely to be a problem for kinetic solubilities as Ostwald's rule[61] states that the least stable (i.e., most soluble) polymorphs will come out of solution first. This is a kinetic, rather than thermodynamic effect, so that we can expect that our data usually refer to the most soluble polymorphs. We also note that of 132 compounds in the recent solubility challenge,[17] only 6 (4.5%) were affected by polymorphism. Given that the maximum possible performance of our models lies around 75% (see below), we do not expect polymorphism to play a significant role in this work.

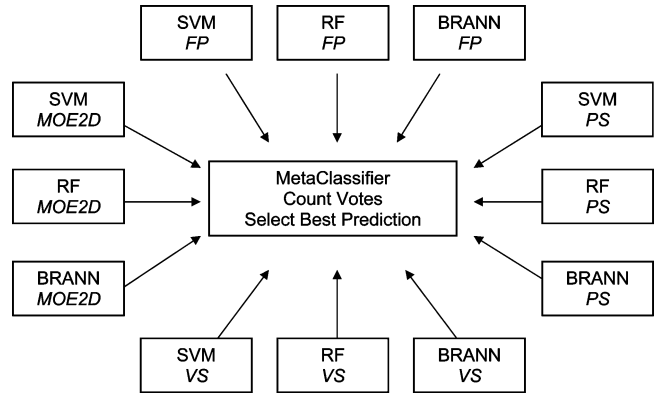

Figure 1. Different schemes for uniting predictions from single classifiers have been examined.

We have also only considered one conformation (that given by CORINA) in calculating the 3D-descriptors. These depend of course on the molecular conformation, but the training data is neither of sufficient accuracy nor quantity to construct a conformationally dependent model, as we have recently done for $\log P_{OW}$.[62] We have analyzed the effect of different conformations on the predictions of a boiling-point model[63] and found that the variations between very different conformations were smaller than the error limits of the predictions (see Figure 1).

Statistical Measures. We have used accuracy and the class-specific receiver operator characteristics ROC area to evaluate the performance of the models. The results are given as confusion matrices. The full results for the class-specific ROC areas are also shown as 3×3 matrices. For two-class classification problems, all ROC areas are the same but not so for multiclass classifications. The scheme for the confusion matrix in a three-class setting is shown in Table S1 of the Supporting Information. Quality measures such as the accuracy of prediction, true and false positive rates are defined in the Supporting Information.

RESULTS

Maximum Possible Performance. It is important to estimate the maximum possible performance of the model based on the likely experimental errors. We[64] have pointed out that the accuracy of solubility models based on the PHYSPROP data is probably limited by the quality of the experimental data themselves. The experimental error of the measurements limits the maximum possible prediction performance. The class assignment from the measurements can only be as good as defined by the standard deviation between the measured and true values. We have therefore calculated the maximum possible performance for all 436 compounds with solubility values that are assigned to bounded bins by assuming different standard errors for the experimental data. The details of this calculation are given in the Supporting Information.

We have analyzed the maximum accuracy achievable as the integral in the measured class divided by the total number of compounds measured for different experimental uncertainties. This is plotted in Figure 2.

As the standard deviation for solubility measurements has been estimated to be 0.46 logS units by Mueller et al[21] and between 0.4 and 0.5 logS units by Dearden,[4] the maximum achievable accuracy for the models presented below must

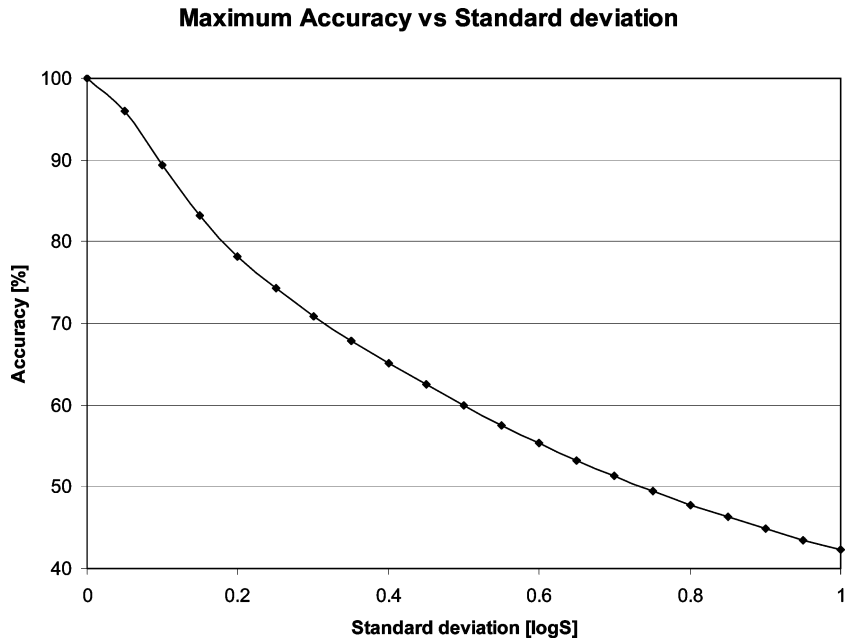

Figure 2. Maximum accuracy achievable as a function of the standard error of the experimental measurements for all bounded bins.

Table 1. Accuracies and ROC-AUCs for All Single Models

algorithm	descriptor set	accuracy	ROC-AUC		
			insoluble	moderately soluble	soluble
SVM	MOE2D	0.683	0.895	0.732	0.922
	VS	0.631	0.896	0.667	0.863
	PS	0.607	0.863	0.659	0.891
	FP	0.609	0.837	0.638	0.859
BRANN	MOE2D	0.715	0.912	0.774	0.932
	PS	0.648	0.881	0.707	0.901
	VS	0.656	0.877	0.703	0.902
	FP	0.612	0.862	0.691	0.877
RF	MOE2D	0.711	0.895	0.753	0.921
	VS	0.627	0.863	0.683	0.883
	FP	0.640	0.856	0.693	0.874
	PS	0.631	0.855	0.660	0.874

correspond to a similar experimental standard error. Assuming realistic experimental uncertainties for the compounds of this data set suggests a maximum achievable accuracy for compounds measured within bounded bins of at most 70%. To reverse the normal point of view, if we assume that the models perform as well as possible for the data,[64] our results allow us to estimate the true standard error of the measurements.

Results of the Single Models. SVMs, BRANNs, and RFs were trained using the four different descriptor sets described above. Descriptor selection was not carried out. The individual results are shown in Table 1.

The best performance is obtained with the BRANN algorithm and the MOE2D descriptor set, which contains a logS descriptor and several logP descriptors, while VS contains one logP and several logD descriptors. FP and PS do not contain either logS or logP descriptors. We have shown earlier[37] that commercial logS predictions are moderately good for this data set with a performance of 62% accuracy and worse (as shown in Table 8).

The BRANN algorithm performs better than the SVM algorithm. The BRANN algorithm is a little better than the RF algorithm with the MOE2D descriptors, worse with the FP descriptors, and clearly better with the PS and the VS descriptors. RF performs clearly better than SVM and BRANN on the FP descriptor set. FP descriptors are positive integers, in many cases only up to five. They are not continuous real numbers but rather a relatively small range of integers (i.e., they are effectively binned). This can be a problem for algorithms that interpolate using continuous functions. The random forest algorithm is able to cope very well with such discrete descriptors.

Metaclassifier. The question of how to unite the single models in order to obtain better predictions naturally arises. There are at least three different approaches. (1) Majority vote: Each classifier/descriptor set combination has one vote; the final prediction is the majority of all votes given. A potential drawback is that a good prediction can be overridden by many bad predictions. (2) Highest probability selection: If all single classifiers output probabilities of correct prediction, the prediction with the highest probability can be selected. Thus, the best single prediction is always used. However, the probability estimates must be very good for this approach to work. (3) Metaclassifier: There might be some structures in the data set like "if two specific models agree on high probability and one specific model disagrees, then this model is wrong" or "If one specific model gives a very high probability, then it is always right". For the metaclassifier, the individual class membership probabilities from the single models are used as a meta-descriptor set that consists of the outputs of the base classifiers. A new nonlinear classifier can then be trained with this meta-descriptor set as input. If the requirements for approaches 1 and 2 are not met, the metaclassifier can still improve or at least retain the performance of the best single model.

Majority vote, highest probability selection, and the metaclassifier were tested for the current data set. For the metaclassifier, we used both random forests and the BRANN algorithm because they perform best for the single models. The confusion tables for the three approaches are shown in Table 2.

Table 2. Confusion Table for the Different Fusion Approaches Described Above

predicted → measured ↓	insoluble	moderately soluble	soluble
Majority Vote			
insoluble	262	60	1
moderately soluble	61	201	32
soluble	0	62	112
Highest Probability			
insoluble	283	47	5
moderately soluble	112	142	60
soluble	3	37	150
Metaclassifier (RF)			
insoluble	272	60	3
moderately soluble	85	187	45
soluble	1	46	143
Metaclassifier (BRANN)			
insoluble	270	62	3
moderately soluble	69	204	44
soluble	1	52	137

Table 3. Confusion Matrix for the Metaclassifier Based on the RF Algorithm, Predicted Class versus Sum of CMPs

predicted → class probability ↓	insoluble	moderately soluble	soluble
All Compounds			
insoluble	267.5	61.7	4.2
moderately soluble	70.4	204.1	44.0
soluble	2.1	52.2	135.7

A total of 51 compounds have equal numbers of votes, so that the majority vote analysis cannot give a prediction. The accuracy for prediction of the remaining 791 compounds is 72.7%. However, all the compounds omitted are hard to predict. If they were included and the false prediction were used for each, the overall accuracy for all 842 compounds would be 70.2%. The performance of the highest probability fusion model is, however, by far the best for the insoluble compounds but far worse for the intermediate class.

No prediction could be obtained for three compounds using the highest probability criterion because they have two maximum predicted probabilities. The accuracy for selection according to the highest probabilities is 68.3%. Especially the prediction for moderately soluble compounds is worse in this case than in other models.

The accuracy for the BRANN metaclassifier is 71.5% and that for the RF metaclassifier 72.6% (marginally higher than the accuracies for fusion approaches 1 and 2). The RF algorithm performs better than the BRANN algorithm in the corresponding metaclassifier.

The confusion matrix for the expected performance, i.e., the distribution of predicted class versus the sum of CMPs is shown in Table 3. Ideally, this matrix should agree well with the confusion table for the metaclassifier based on the RF models shown in Table 2.

The expected distribution of predictions agrees extremely well with the real ratios of predictions obtained. The number of extreme outliers (i.e., ones that are in error by two classes) expected is overestimated, but this depends on the sum of very low probabilities, which are difficult to predict.

The accuracy for all compounds assigned to bins with finite bounds is 64.7%. This is lower than the overall accuracy obtained with all the training/test data because the compounds assigned to bins with finite bounds are closer to the thresholds than those assigned to the ">" or "<" bins. According to our estimation, the accuracy of 64.7% corresponds to a standard error of 0.4−0.45 logS units in the measured data. This is in perfect agreement with the estimates of experimental uncertainties for thermodynamic solubilities by Dearden and Mueller. The probability distribution of measurements versus the assigned class for $\sigma = 0.45$ logS units is shown in Table 4.

Table 4. Observed and Predicted Maximum Possible Performance (Assuming $\sigma = 0.45$ logS Units) for the Compounds in Bounded Bins

predicted → assigned ↓	insoluble	moderately soluble	soluble
Observed Performance			
insoluble	44	23	2
moderately soluble	69	204	44
soluble	0	16	34
Predicted for $\sigma = 0.45$ logS Units			
insoluble	45.09	23.56	0.35
moderately soluble	71.09	191.99	53.92
soluble	0.19	14.51	35.30

The accuracy calculated for $\sigma = 0.45$ is 62.5%. The distribution across the confusion matrix agrees very well with that found assuming that the measured values are exact. Given that the standard error of measurement is around 0.45 logS units, the model presented here is already optimal. If the predictions were better than the measurements, this would be impossible to see from comparing experimental data with the predictions. The ROC curves for the RF metaclassifier are shown in Figure 3.

The overall performance is highest for the RF metaclassifier model, both in terms of accuracy and ROC-AUC. The ROC-AUCs for the RF metaclassifier are shown in Table 5.

Insoluble compounds are best separated from soluble ones. Here, the ROC-AUC is nearly 1, which means that the separation is close to perfect, as expected because these two classes are physically the farthest apart. Also as expected, the worst separation is obtained for the moderately soluble compounds, although the ROC-AUC is still a respectable 0.79. The confusion table and, even more so the ROC-AUC matrix, illustrate that any improvement must come from identifying the moderately soluble compounds better. However, this class is also most susceptible to experimental noise, so that the performance of the model is likely to be limited by the accuracy of the experiments.

Systematic Trends. Predictions for compounds with very high and very low values of ClogP tend to be more accurate than those with intermediate values. However the extreme compounds are probably also easier to predict because solubility is usually considered to be directly related to logP. The quality of the predictions shows no discernible dependence on molecular weight, since the range of MWt is relatively small (∼350−500 g/mol) for this data set.

The prediction accuracy for neutral compounds (∼75%) is higher than that for charged ones (∼65%) This may be because there are more neutral than charged compounds in the data set but may also indicate an extra source of error introduced with the pK_a estimation module.

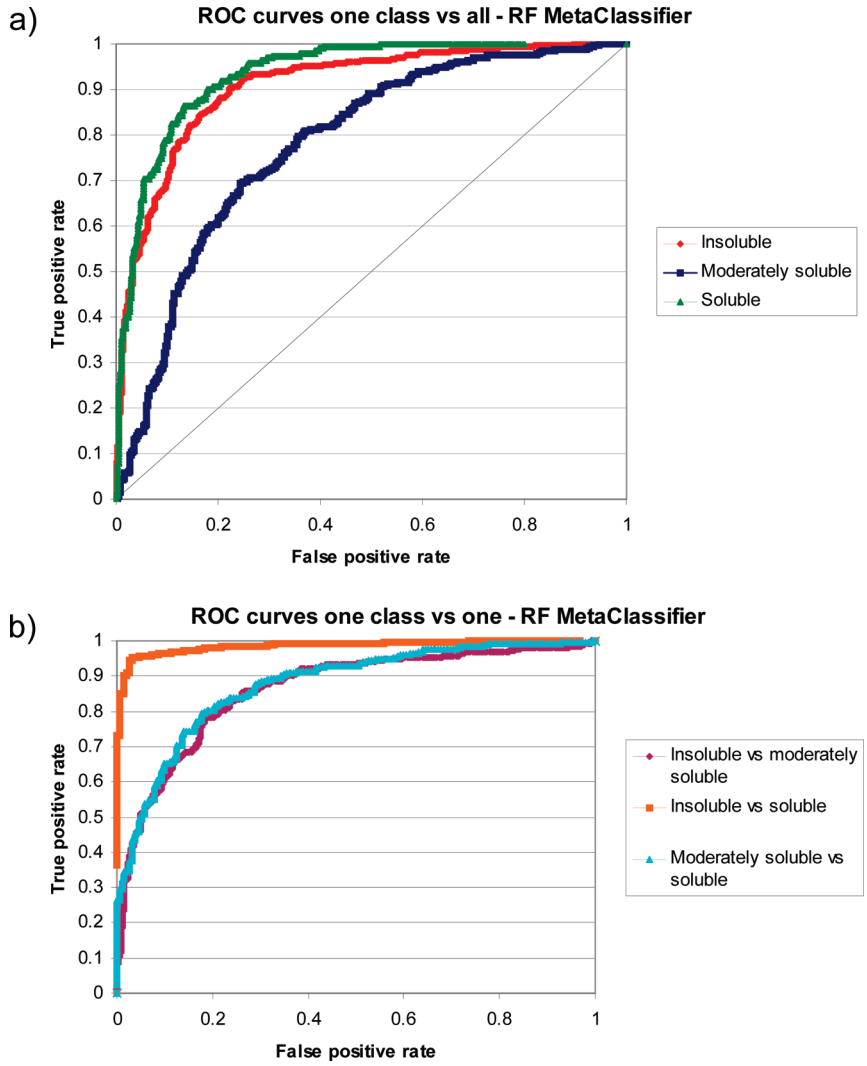

Figure 3. ROC curves for the RF metaclassifier (a) one vs all and (b) one vs one.

Table 5. ROC-AUC Table for the RF Metaclassifier

predicted → measured ↓	insoluble	moderately soluble	soluble
insoluble	0.916	0.873	0.990
moderately soluble	0.873	0.787	0.880
soluble	0.990	0.880	0.937

Probability Examination. All the models generated so far give CMPs. CMP thresholds could be used to "cherry-pick" compounds with safe predictions. The confusion matrices given by applying different probability thresholds are shown in Table S3 in the Supporting Information. These data can be used to test the reliability of the CMPs by comparing the mean CMPs with the observed accuracies for each class and each CMP threshold. These results are shown in Figure 4.

The CMPs are in excellent agreement with the observed accuracies of prediction. The deviations from the 1:1 line are explained by the statistical deviations within the small groups for higher probabilities. Figure 4 supports the use of CMPs as a measure of the reliability of predictions.

Validation. The RF metaclassifier model was validated with an in-house solubility data set from Boehringer-Ingelheim that consists of 747 project-related compounds that were measured using the same protocol as for the training data set. The measurements were made up to 2 years before those of the training data set. The initial class distribution of the validation set compounds is different from the training set; there are less insoluble compounds. The structural

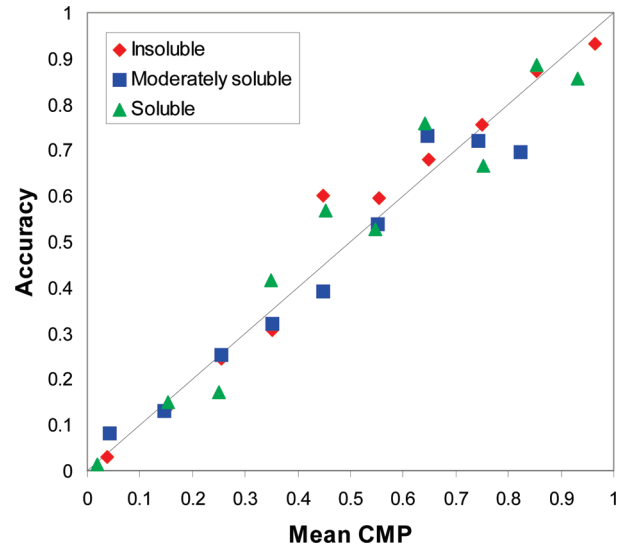

Figure 4. Comparison of the mean classification probability with the observed accuracy for the data shown in Table S3 in the Supporting Information.

Table 6. Results for the In-House Validation Data Set for the RF Metaclassifier

predicted → measured ↓	insoluble	moderately soluble	soluble
Observed Performance			
insoluble	53	42	4
moderately soluble	55	173	89
soluble	2	71	258
ROC-AUC			
insoluble	0.885	0.786	0.974
moderately soluble	0.786	0.704	0.824
soluble	0.974	0.824	0.861

Table 7. Results for the Metaclassifier for the Compounds Assigned to Bins with Finite Bounds Only, Validation Set

predicted → measured ↓	insoluble	moderately soluble	soluble
insoluble	12	8	2
moderately soluble	55	173	89
soluble	0	46	102
Predicted Distribution for $\sigma = 0.45$ logS units			
insoluble	14.7	7.2	0.1
moderately soluble	58.1	192.1	63.8
soluble	0.7	44.1	103.2

Table 8. Accuracies Achieved on the Training and Validation Sets[a]

model	accuracy (%)		robustness (%)
	training set	validation set	
RF metaclassifier	72	65	90
MWt, ClogP, ionization state	62	49	79
MWt, AP	49	39	80
ACDlabs logS	62	57	92
MOE logS	62	54	87

[a] Training set performance taken from ref 37.

distribution of the validation set from project related work also differs from that of the training set. This is shown in a plot of the first two principal components of the MOE descriptor set in Figure S1 in the Supporting Information. The confusion and ROC-AUC matrices for the validation data set are shown in Table 6. The overall accuracy is 64.7%.

The confusion matrix for all compounds assigned to bins with finite bounds for the validation set and the corresponding confusion matrix obtained by assuming a standard experimental error, $\sigma = 0.45$, are shown in Table 7.

The distribution of the predictions of the validation set is fairly consistent with that expected for $\sigma = 0.45$ logS units but not as convincing as for the training/test set. We also examined the predictive powers of simple rules-of-thumb and commercially available solubility predictors that were published in ref 37. The accuracies achieved are summarized in Table 8, which also shows what we have called "robustness", which is simply the percentage of the accuracy achieved for the training set found for the validation set (i.e., accuracy-(validation set)/accuracy(training set) expressed as a percentage).

Overall, the performance is similar to the training set, with the difference that insoluble compounds (of which there are less in the validation set) cannot be identified as well. The difference between the insoluble compounds from the training and validation data sets lies in their origin: The training set was assembled randomly. The insoluble compounds from the validation set originate directly from project work; they are less obviously insoluble and on average closer to the classification threshold than the training set. It does not contain simply discernible insoluble compounds. This is illustrated in plots of the first two principal components of the descriptor sets for the insoluble compounds shown in Figure S1 in the Supporting Information. One trend that is clearly visible from Table 8 is that the two very simple rule-of-thumb approaches only achieve approximately 80% of their performance for the training set when they are used for the validation set. The more complex metaclassifier and commercial models, on the other hand, all achieve approximately 90% of their training-set performance for the validation set. Thus, the metaclassifier not only performs best for its own training data but also retains its higher performance for the validation data, despite the obvious differences in the two data sets.

Table 9. Results for the Solubility Challenge Validation Set

predicted → measured ↓	insoluble	moderately soluble	soluble
Solubility Challenge Training Set			
insoluble	3	3	2
moderately soluble	1	5	8
soluble	0	8	69

Llinas, Glen, and Goodman recently published a data set of 101 solubility measurements for known drugs[17] and challenged the QSAR/QSPR community to predict a test set of 32 drugs with unpublished solubilities.[65] Although these solubilities are not strictly comparable to those that are the subject of our models, we used them as an additional validation set. The solubilities used for building our models were measured using a kinetic method at pH 7.4 whereas the measurements of Llinas, Glen, and Goodman correspond to intrinsic solubilities, the solubilities of the free acid or base in the absence of DMSO. Two of the measured compounds decomposed under the measuring conditions, so we only give predictions for 99 compounds, of which only 8 are insoluble within our classification scheme. We only show predictions for the training data set of the solubility challenge, since no kinetic solubilities were published for the test set.[65] The confusion matrix for the solubility challenge validation set is shown in Table 9.

The overall accuracy of prediction is 77.8%. Exactly this fraction of the solubility challenge compounds is soluble in our classification system. This is not surprising, since the challenge data set consists only of drugs, which are usually more soluble than the early phase development compounds targeted in our models. However, our model only identifies three of eight insoluble compounds with two strong outliers, mefenamic acid ($pK_a = 4.22$) and trimipramine ($pK_a = 9.34$). They are both ionized at pH 7.4, the pH of our measurements, so more soluble than measured at the pH of their intrinsic solubility and thus probably predicted correctly. Our model tends to predict compounds from the solubility challenge data set to be more soluble than their measured values. This is not surprising since our model is based on data measured at pH 7.4 and in the presence of some DMSO, which makes

compounds more soluble. Thus, the performance of our model is essentially as expected for the solubility challenge training data set. We expect a constant pH model for aqueous solubility in the presence of DMSO to predict higher solubilities than those contained in the solubility challenge data set. Nonetheless, the overall accuracy is similar to that found for the other data sets and would be higher without the model-inherent systematic overestimation of the solubility caused by physical differences in the measuring techniques used for the two different training data sets.

The performance of our model in identifying insoluble compounds (its proclaimed purpose) is poor when evaluated on the solubility challenge dataset. However, the three compounds insoluble predicted to be moderately soluble (chlorpromazine, chlorprothixene, and sertraline) are also all protonated at the pH used for the measurements, so that we would expect them to show different behavior under the conditions used to measure our training set.

DISCUSSION

The metaclassifier described above can classify 73% of compounds from the insolubility data set correctly and performs marginally better than the best single classifier. At first sight, however, the performance of nonlinear classifiers trained with complete descriptor sets is disappointingly similar to that of the simple models based on decision trees and basic descriptors introduced earlier (approximately 62%).[37] Our analysis of the best possible prediction performance based on the likely experimental errors, however, suggests that 73% is probably the maximum performance achievable while 40% is the performance of a random classifier. The improvement obtained using nonlinear high-end classifiers is therefore around $[(73 - 62)/(73 - 40)] = 33\%$, and the metaclassifier is probably performing as well as possible for this data set. It also retains its performance advantage over the commercial models for the in-house validation data set and is markedly superior to simple rules-of-thumb models for all data sets.

The known[20] narrow range of applicability of solubility models is neither a function of the descriptors nor the classification technique but rather of the training data set. This is also evident from the performance of the metaclassifier for compounds that are not druglike. The accuracy of the metaclassifier for the Huuskonen data set is only 48%, not much higher than random, whereas it performs significantly better (61%) for the druglike subset of the Huuskonen data set.

The RF-based metaclassifier described here gives a ROC-AUC of 0.990 for insoluble versus soluble classes. It is also able to differentiate between these two classes nearly perfectly for the independent validation set. The differentiation between moderately soluble and soluble compounds is also good. The poorer performance in differentiating between insoluble and moderately soluble compounds indicates some locality in the metaclassifier that limits its performance when confronted with new types of insoluble compounds. This can be seen from the distribution in the PCA plots, where most of the insoluble compounds from the validation set are located close to the origin of the principal components. There are fewer compounds in the obviously insoluble region, where all the other insoluble compounds from the training set are located.

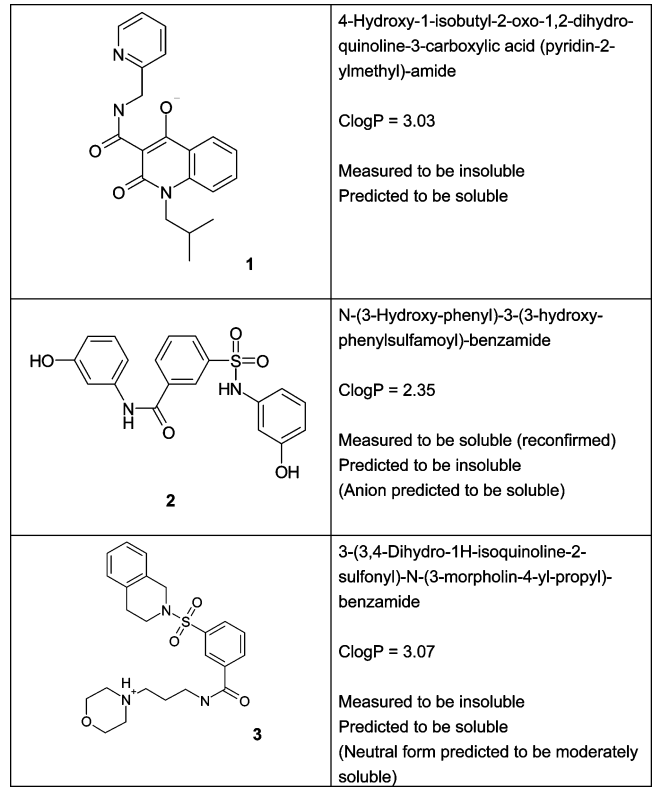

Figure 5. Strong outliers.

The RF metaclassifier model has four strong outliers. Three of them are from ref 37 and discussed below. They are shown in Figure 5.

1 could not be remeasured because of lack of substance. It is predicted to be soluble with a probability of 61%. We used the form shown in Figure 5 for the calculations. However JChem[38] predicts a pK_a of 7.0 for the oxygen position. Thus, the error for this compound is likely to be due to using the inappropriate tautomer or protonation state in the calculations. This is a common disadvantage of techniques based on 3D structures. To illustrate this point we list all possible and somehow realistic tautomers and protomers in Figure 6. Additionally we give the probabilities for the different solubility classes in the order insoluble/moderately soluble/soluble as calculated by the metaclassifier. We did not generate a new metaclassifier without compound **1**, since this is computationally very expensive. We rather point to the fact that the presence of microspecies **1.3** in the training set shifts all predictions toward worse solubility. However the probability for being insoluble for **1.3** is the lowest as compared to the other microspecies, although it was in the training set. Thus all other microspecies are predicted to be less soluble than **1.3**. Microspecies **1.3** is incorrectly predicted to be soluble being in the test set.

2 has been remeasured and confirmed. The probability of belonging to be insoluble for **2** is ~64%. However JChem[38] here predicts a pK_a of 7.5 for the sulfonamide nitrogen. The monodeprotonated compound is predicted to be soluble with a probability of 61%.

3 could not be remeasured because of lack of substance. It is predicted with ~62% probability, so this is also not a safe prediction. JChem calculates pK_a = 7.5 for the morpholino-nitrogen. We used the singly protonated form for

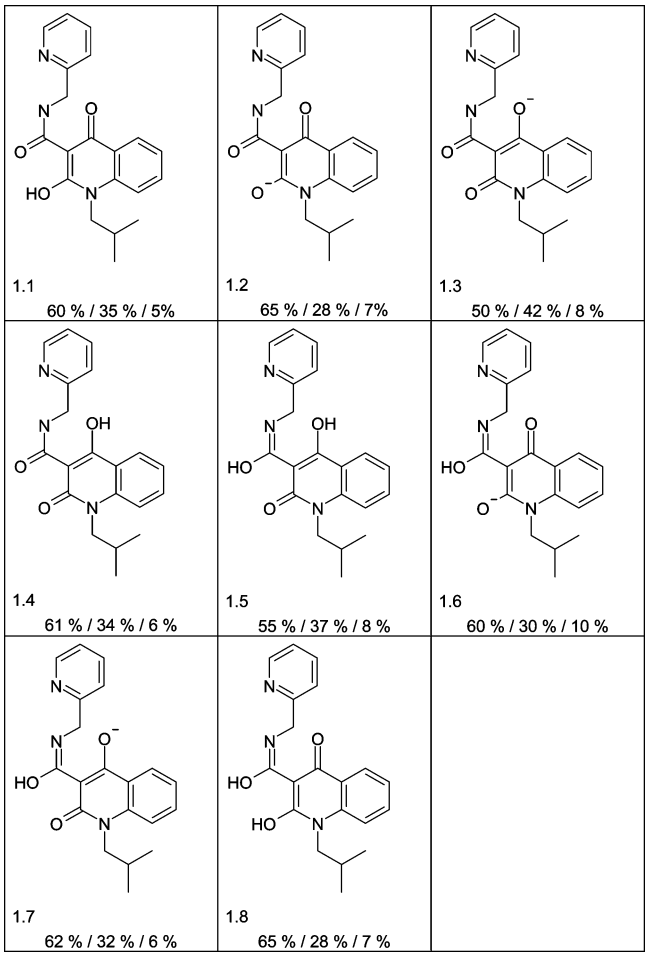

Figure 6. Possible tautomers and protomers for compound **1** at pH 7.4 and probabilities (insoluble/moderately soluble/soluble) predicted with **1.3** in the training set.

the calculations. The neutral form is predicted to be moderately soluble with a probability of 52%.

The strong outliers are thus most probably caused by incorrect protonation states. They all have pK_a's predicted to be around the pH value used for the measurements. If another protonation form is used, the prediction becomes correct or a minor error.

Finding only 4 of 525 possible strong outliers speaks for a robust and conservative model. The ROC curve for the predictions of insoluble versus soluble compounds shows that the errors are all located in the area where the model is not expected to be safe. Among the safe predictions, there are no strong outliers. Thus, we believe that this model is highly reliable and can be used for early phase drug design purposes.

The most significant descriptors from each set of descriptors were determined from the RF models using a combination of the gain in gini-purity and the occupation of the corresponding leaves. The results are shown in Table 10.

The top 3 descriptors of the MOE2D set and all but 2 of the top 10 from VolSurf are derived from logS, logP, or logD models. The important ParaSurf descriptors include the kurtosis of the distributions of the local electronegativity and the local electron affinity on the molecular surface, the dipole density, the dipole moment itself, and the minimum of the local electron affinity. The kurtosis of local electronegativity and local electron affinity are correlated with $R^2 = 0.52$. Some conceptual link between these descriptors and the solubility can be imagined. The two E-state descriptors relate to nitrogen atoms and may be functioning as pseudo atom counts. The variance in the Shannon entropy of the surface properties describes the ability of the molecule to interact with itself and is thus related to the crystal lattice energy.

Table 10. Most Significant Descriptors from Each Set in the Metaclassifier

position	MOE2D	ParaSurf	VolSurf
1	logS	ENEGkurt	$LogD_5$
2	SlogP	SHANEvar	$LogD_6$
3	logP_o/w	EALkurt	$LogD_7$
4	GCUT_PEOE_0	dipden	$LogD_{7,5}$
5	PEOE_VSA_NEG	MEPvartot	$LogD_8$
6	PEOE_VSA_FNEG	LocPol[a]	HL2
7	GCUT_SLOGP_0	EstateN[b]	$LogD_9$
8	PEOE_VSA_FPOS	Dipole	LogP
9	BCUT_SLOGP_2	Estate2N[c]	$LogD_{10}$
10	SlogP_VSA7	EALmin	CW6

[a] The local polarity.[66] [b] The sums of the standard Kier-Hall E-states.[67,68] for all nitrogen atoms. [c] The sums of E-states based on bond orders, rather than topological distances.[69]

We have deliberately not used data reduction to build the metaclassifier models. This is partly because random forests do not require prior data reduction[33,47,48] and also because the data reduction procedure is delicate and care must be taken in order not to overtrain. This risk exists, for example, in all approaches that use evaluation based on the cross-validated correlation coefficient.[70]

SUMMARY AND OUTLOOK

We have presented a solubility metaclassifier based on a wide variety of descriptors using the random forest classification approach. Two important aspects of this work are the estimation of the maximum possible performance that we can expect based on the assumed accuracy of the data and the extensive use of probabilities of correct classification (CMPs) both to combine the individual models to the metaclassifier and to judge the likely reliability of a given prediction.

We expect both these techniques to gain importance in QSPR modeling as the realization that most models are data-limited becomes widespread. In a way, this development is a reaction to overfitting and excessive locality in QSPR models. It should be clear from the above discussion that we can usually expect only approximately 65% classification accuracy for solubility of "foreign" validation sets using the three solubility classes above. The reasons lie not only in the obvious scatter in the experimental values for each compound but also in ambiguities as to the exact structure (usually tautomer or protonation state) of the compound in solution.

ACKNOWLEDGMENT

This work was supported by Boehringer-Ingelheim Pharma & Co. KG. We thank Thilo Fligge for his valuable discussions.

Supporting Information Available: The confusion matrix and ROC_AUC matrix scheme for a three-class problem and the impact of experimental uncertainty in the classification

scenario. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Xia, X.; Maliski, E.; Poppe, L.; Cheetham, J. Solubility Prediction by Recursive Partitioning. *Pharm. Res.* **2003**, *20*, 1634–1640.

(2) Balakin, K. V.; Savchuk, N. P.; Tetko, I. V. In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: trends, problems and solutions. *Curr. Med. Chem.* **2006**, *13*, 223–241.

(3) 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.

(4) Dearden, J. C. In silico prediction of aqueous solubility. *Exp. Opin. Drug Discovery* **2006**, *1*, 31–52.

(5) Delaney, J. S. ESOL: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.

(6) Duchowicz, P. R.; Talevi, A.; Bruno-Blanch, L. E.; Castro, E. A. New QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorg. Med. Chem.* **2008**, *16*, 7944–7955.

(7) Du-Cuny, L.; Huwyler, J.; Wiese, M.; Kansy, M. Computational aqueous solubility prediction for drug-like compounds in congeneric series. *Eur. J. Med. Chem.* **2008**, *43*, 501–512.

(8) Eros, D.; Keri, G.; Kovesdi, I.; Szantai-Kis, C.; Meszaros, G.; Orfi, L. Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods. *Mini Rev. Med. Chem.* **2004**, *4*, 167–177.

(9) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* **2002**, *19*, 497–503.

(10) 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.

(11) Huuskonen, J. Estimation of aqueous solubility in drug design. *Comb. Chem. High Throughput Screening* **2001**, *4*, 311–316.

(12) Huuskonen, J. Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 491–497.

(13) Huuskonen, J.; Livingstone, D. J.; Manallack, D. T. Prediction of drug solubility from molecular structure using a drug-like training set. *SAR QSAR Environ. Res.* **2008**, *19*, 191–212.

(14) Klamt, A.; Eckert, F.; Hornig, M. COSMO-RS: a novel view to physiological solvation and partition questions. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 355–365.

(15) Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process. *J. Med. Chem.* **2008**, *51*, 2891–2897.

(16) Liu, R.; So, S. S. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.

(17) Llinas, A.; Glen, R. C.; Goodman, J. M. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements. *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.

(18) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A consensus neural network-based technique for discriminating soluble and poorly soluble compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 674–679.

(19) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(20) Schroeter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 651–664.

(21) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sulzle, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Accurate solubility prediction with error bars for electrolytes: a machine learning approach. *J. Chem. Inf. Model.* **2007**, *47*, 407–424.

(22) Tantishaiyakul, V.; Worakul, N.; Wongpoowarak, W. Prediction of solubility parameters using partial least square regression. *Int. J. Pharm.* **2006**, *325*, 8–14.

(23) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodiversity* **2004**, *1*, 1829–1841.

(24) 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 Moleculaes. *J. Chem. Inf. Model.* **2004**, *44*, 1477–1488.

(25) Jain, N.; Yang, G.; Machatha, S. G.; Yalkowsky, S. H. Estimation of the aqueous solubility of weak electrolytes. *Int. J. Pharm.* **2006**, *319*, 169–171.

(26) Raevsky, O. A.; Raevskaja, O. E.; Schaper, K.-J. Analysis of water Solubility data on the basis of HYBOT descriptors. Part 3. Solubility of solid neutral chemicals and drugs. *QSAR Comb. Sci.* **2004**, *23*, 327–343.

(27) Clark, M. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.* **2005**, *45*, 30–38.

(28) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. W. F. Linear and nonlinear methods in modeling the aqueous solubility of organic compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170–176.

(29) Yan, A.; Gasteiger, J.; Krug, M.; Schaper, K.-J. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75–87.

(30) Lind, P.; Maltseva, T. Support vector machines for the estimation of aqueous solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.

(31) Todeschini, R.; Consonni, V.; Mauri, A.; Paven, M. *DRAGON for Windows and Linux*; Talete SRL: Milano, Italy, 2007, accessible via http://www.talete.mi.it.

(32) *Molecular Operating Environment 2008.10*; Chemical Computing Group: Montreal, Quebec, Canada, 2008.

(33) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.

(34) Tetko, I. V. VCCLAB, Virtual Computational Chemistry Laboratory (http://www.vcclab.org), 2005.

(35) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity; State-of-the-art and comparison of logP methods on more than 96,000 compounds. *J. Pharm. Sci.* **2008**, *98*, 861–893.

(36) Syracuse Research Corporation. The Physical Properties Database (PHYSPROP); SRC Environmental Science Center: Syracuse, NY.

(37) Kramer, C.; Beck, B.; Fligge, T. A.; Heinisch, T.; Clark, T. A Consistent Dataset of Kinetic Solubilities for Early-phase Drug Discovery. *ChemMedChem* **2009**, *4*, 1529–1536.

(38) *JChem*, version 5.1; ChemAxon: Budapest, Hungary, 2008.

(39) *ParaSurf08*; CEPOS Insilico Ltd.: Erlangen, Germany, 2008.

(40) *VolSurf*, version 4; Molecular Discovery Ltd.: Pinner, U.K., 2008.

(41) Chang, C.-C.; Lin, C.-J. *LIBSVM: A Library for Support Vector Machines*; 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin.

(42) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowledge Discovery* **1998**, *2*, 121–167.

(43) Norinder, U. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimizations and variable selection. *Neurocomputing* **2003**, *55*, 337–346.

(44) Neal, R. *Software for Flexible Bayesian Modeling and Markov Chain Sampling*, release 2004-11-10; University of Toronto: Toronto, Canada, 2004.

(45) Neal, R. M. *Bayesian Learning for Neural Networks*; Springer-Verlag: New York, 1996; Vol. 118.

(46) Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. Use of Automatic Relevance Dtermination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423–1430.

(47) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(48) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–58.

(49) Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees*; Wadsworth: New York, 1984.

(50) Hall, L. H.; Kier, L. B. *Molecular Connectivity in Structure-Activity Analysis*; RSP-Wiley: Chichester, U.K., 1986.

(51) Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.

(52) Gasteiger, J.; Engel, T. *Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2003.

(53) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular model. 76. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(54) *CORINA*, version 3.4; Molecular Networks Inc.: Erlangen, Germany, 2006.

(55) Clark, T.; Alex, A.; Beck, A.; Burkhardt, F.; Chandrasekhar, J.; Gedeck, P.; Horn, A. H. C.; Hutter, M.; Martin, B.; Rauhut, G.; Sauer, W.; Schindler, T.; Steinke, T. *VAMP*, version 8.2; Accelrys Inc.: San Diego, CA, 2002.

(56) Holder, A. J. AM1. In *Encyclopedia of Computational Chemistry*, Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds. Wiley: Chichester, U.K., 1998; pp 8−11.

(57) Cruciani, G.; Crivori, P.; Carrupt, R.-A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: tThe VolSurf Approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17–30.
(58) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849–857.
(59) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2*, 559–572.
(60) Palm, K.; Luthmann, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of Drug Absorption with Molecular Surface Properties. *J. Pharm. Sci.* **1996**, *85*, 32–39.
(61) Threlfall, T. Structural and Thermodynamic Explanations of Ostwald's Rule. *Org. Process Res. Dev.* **2003**, *7*, 1017–1027.
(62) Kramer, C.; Beck, B.; Clark, T. A Surface-Integral Model for $\log P_{OW}$, *J. Chem. Inf. Model*. **2010**, *50*, in press.
(63) 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.
(64) Clark, T. Modelling the Chemistry: Time to Break the Mould? In *EuroQSAR 2002: Designing Drugs and Crop Protectants: Processes, Problems and Solutions*; Ford, M. G., Livingstone, D. J., Dearden, J. C., van de Waterbeemd, H., Eds. Blackwell Publishing: Oxford, U.K., 2002; pp 111−121.
(65) Hopfinger, A. J.; Esposito, E. X.; Llinas, A.; Glen, R. C.; Goodman, J. M. Findings of the challenge to predict aqueous solubility. *J. Chem. Inf. Model.* **2009**, *49*, 1289–1303.
(66) Murray, J. S.; Politzer, P. A. 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.
(67) Hall, L. H.; Mohney, B.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76–82.
(68) Kier, L. B.; Hall, L. H. *Molecular Structure Descriptors: The Electrotopological State*; Academic Press: New York, 1999.
(69) Beck, B. unpublished results.
(70) Golbraikh, A.; Tropsha, A. Beware of Q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–277.

CI900377E