Articles

Straightforward Recursive Partitioning Model for Discarding Insoluble Compounds in the Drug Discovery Process

Claudia Lamanna, Marta Bellini, Alessandro Padova, Goran Westerberg, and Laura Maccari*

Siena Biotech S.p.A., Via Fiorentina 1, 53100, Siena, Italy

Received November 8, 2007

Poor aqueous solubility is one of the major issues in drug discovery and development, impacting negatively on all aspects of the research and development process. The pharmaceutical industry has realized that solubility issues need to be resolved at the discovery stage. We here present an innovative way to address this problem via a model designed to address the simple question, "Is the compound likely to be sufficiently soluble to provide interpretable data in biological screening assays?" A recursive partitioning (RP) method was applied to a set of 3563 molecules, with in house determined aqueous solubility values. Five models were generated on the basis of a small number of descriptors affording intuitive information regarding structural features influencing solubility. The final model was based on only two descriptors: the molecular weight (MW) and the aromatic proportion (AP). This model provided satisfactory values of accuracy (81%) and precision (75%) for a test set of 1200 compounds, suggesting that the model may add value in compound selection and library design during early drug discovery.

Introduction

Modern drug discovery is a complex process that requires problem solving in multiple dimensions, often using tools with insufficient dimensionality. Prediction and modeling of physicochemical properties represent one of these challenges. In view of the number of parameters involved in determining the aqueous solubility and the plethora of molecular descriptors available, overfitting of models should always be treated with care.

Drug distribution, delivery, and transport depend at least in part on solubility. Solubility is also central to in vitro screening, since low solubility frequently results in poor reproducibility and unreliable results. If a drug precipitates before reaching its cellular target, the target will be exposed to a concentration of drug lower than the nominal and could therefore yield a response that is diminished, undetectable, or independent of the input concentration.[1]

Solubility issues can best be addressed at the discovery stage; however, experimental testing for solubility can be time and resource consuming, in particular when dealing with extremely large compound collections in early phases of drug discovery. Hence, methods are required to help the definition of acceptable solubility at the compound design or acquisition stage. Computational approaches capable of identifying compounds with a high probability of being soluble (or insoluble, as the case may be) would therefore be of value.

From our experience, solubility seems to be a particularly pertinent issue when purchasing compounds from commercial sources. From earlier work,[2] 66% of commercially available compounds had solubility lower than 80 μM at pH 7.4 (see Methods for assay conditions). Considering an average cost per compound of $18, poor solubility can contribute a significant cost when performing screening of tens or hundreds of thousands of compounds in the initial phase of the drug discovery process.

A large number of computational approaches for prediction of aqueous solubility have been published; models predicting aqueous solubility based on molecular surface area, lipophilicity, hydrophilic measures and electronic and topological descriptors have all been reported.[3] Several regression methods have been presented, and among them neural networks represent a particularly used approach[4–13] while linear methods provide results comparable to support vector regression.[14,15] In the literature, it is also possible to find less mathematically complex methodologies, including linear regression, recursive partitioning (RP),[16] and partial least-squares analysis (PLS).[a,17,18] However to date, complex models based on methods such as neural networks or continuum regression[19] often provide better models than simple methodologies, such as linear regression analysis.[8]

A common feature of the above models is that they are based on heterogeneous data sets[4–13] containing molecules with structural features not commonly seen in drugs; for example, solubility data of chemicals from pesticide lists have been used for this purpose.[20] Finally, these models also employ complex descriptors of limited use in understanding structure–property relationships, thus limiting usefulness when designing libraries.

We were interested in developing a model for predicting the solubility of organic compounds as a decision aid to whether or not to synthesize or purchase compounds rather than assigning a predicted numerical solubility value. In line with this goal,

* To whom correspondence should be addressed. Phone: +39 0577 381434. Fax: +39 0577 381410. E-mail: lmaccari@sienabiotech.it.

[a] Abbreviations: RP, recursive partitioning; MW, molecular weight; AP, aromatic proportion; PLS, partial least-squares; UPLC, ultraperformance liquid chromatography; UV, ultraviolet; TOF, time-of-flight; MS, mass spectrometry; 2D, bidimensional; 3D, tridimensional; PCA, principal component analysis; log *P*, logarithm of the partition coefficient of octanol/water; PSA, polar surface area; RTB, rotatable bonds; HBD, hydrogen bond donors; HBA, hydrogen bond acceptors; SMARTS, SMILES arbitrary target specification; SMILES, simplified molecular input line entry specification; QSAR, quantitative structure−activity relationship; SAR, structure–activity relationship; QSPR, quantitative structure−property relationship.

10.1021/jm701407x CCC: $40.75
Published on Web 04/18/2008

Table 1. Distribution of "Soluble" and "Insoluble" Compounds Corresponding to Each Solubility Threshold Value Used in the Study

solubility cutoff value (μM)	no. of "soluble" compds	no. of "insoluble" compds
20	1630	1933
30	1702	1861
40	1760	1803
50	1822	1741
60	1895	1668

we sought to distill the minimum requirements of a predictive model to help classify virtual structures into "soluble" and "insoluble", without attempting a numerical or rank-ordering approach. We were also attracted by the use of a limited number of readily interpretable descriptors as already described in the paper of Xia[16] and Delaney.[20] Thus, we chose the RP approach,[21] which is a valuable tool in the simple handling of nonlinear stepwise variable and complexity reduction. Moreover, simple descriptors (including the druglikeness descriptors developed by Lipinski[22] and Veber[23]) and a set of homogeneous experimental solubility data were used.

This "return to simplicity" seems to be a general trend in current literature; Zhao et al.[24] attempted to model brain permeability data from selected sources, using a RP approach for data classification. Another example used RP and reduced descriptor sets in classifying compounds as possible aggregants.[25]

The ultimate goal of the current work is to enrich our compound collection with more soluble compounds in order to achieve improved cost-effectiveness in compound purchasing and synthesis efforts. The model described herein thus has two main deliverables: the generation of a simple and efficient tool to identify compounds with a high probability of being insoluble (as defined by a predetermined cutoff value) to guide compound purchasing and prioritization of libraries for synthesis and, second, to use easily interpretable descriptors to represent factors contributing to solubility within or across series. The latter objective is subject to availability of a simple model, with few parameters, directly interpretable in terms of structural features.

Methods

Solubility Data. In order to obtain a homogeneous data set with a sufficient number of observations, with broad structural diversity, and with high quality experimental data, we measured aqueous solubility of 3563 compounds from our internal compound collection. The data set included both compounds synthesized internally and those purchased from commercial sources, spanning about 10 diverse major chemical series selected to be tested in 6 drug discovery projects plus many other series (roughly 50) coming from diverse selections performed on commercial libraries. Standard and sample solutions were prepared from a 10 mM DMSO stock solution using an automated dilution procedure to provide a nominal concentration of 250 μM, diluted in acetonitrile/ammonium acetate buffer at pH 7.4 (60/40 v/v), with a final DMSO content of 2.5% (v/v). Solubility was measured at pseudothermodynamic equilibrium conditions by incubating 5 μL aliquots of test article in 10 mM DMSO solution in 50 mM ammonium acetate buffer, pH 7.4, with a final content of DMSO of 2.5% (v/v) after 24 h of equilibration in a Millipore MultiScreen HTS 96-well filtration plate (0.4 μm) at room temperature with shaking. All liquid handling operations were conducted by an automated procedure running on a PerkinElmer Multiprobe II EX liquid handler. Following incubation, solutions were filtered on the MultiScreen plate and the resulting concentration in the filtrate and the standard solution measured using a Waters Acquity UPLC/UV/TOF-MS system. Calculation was based on the UV signal using UV detection at 254 nm, although both UV and MS data were collected for data interpretation. Results were presented as solubility in the range 1–250 μM; results falling outside this range were denoted <1 μM or >250 μM.[26,27]

It is noted that although the final assay solution contained 2.5% DMSO, which to some extent may bias the exact value of solubility vs the thermodynamic equilibrium methods, the scope of the model was to classify compounds into "soluble" and "insoluble". We believe that the impact of this bias is small and acceptable for the purpose of the current work.

The Response Set. Solubility data for the 3563 compounds were collected to be used as the dependent variable to build the statistical model, using a data range of $1-250$ μM. For partition tree purposes, the continuous values for aqueous solubility were transformed into classes. Thus, a solubility threshold value had to be identified to classify the set into "soluble" and "insoluble" subsets. However, because the definition of the cutoff threshold is not straightforward, the response variable was therefore classified into two groups on the basis of threshold values of 20, 30, 40, 50, and 60 μM, respectively, where in each case, compounds below these thresholds were denoted as "insoluble" and compounds with values higher than the cutoff were denoted as "soluble". The number of soluble and insoluble compounds in the data set for each threshold is reported in Table 1.

Each response variable classification was then used to generate RP models, resulting in five different models, hereafter referred to as model 20, model 30, model 40, model 50, and model 60.

The 2D structures for the 3563 compounds constituting the data set, further split into training and test sets (details reported in the next paragraph), were downloaded from our data warehouse using the proprietary software Nucleo.[2] Structures were stored after counterion removal, addition of hydrogen atoms, and adjustment of formal charges. In particular, the ionization state of basic and acidic groups was adjusted to physiological pH, using the LigPrep software.[28] No 3D information was used, as our main goal was to develop a model based on easily interpretable, low-dimensional descriptors.

In order to assess compound diversity within this set of molecules, a simple 3D chemical space, defined by molecular weight (MW), log P (calculated as ALogP, using the Accord SDK from Accelrys),[29] and aromatic proportion of the molecule (AP) as derived by Yan and Gasteiger,[8] was explored.

Figure 1 shows a 3D representation of the results obtained from this analysis suggesting that the data set covers a quite large chemical space ($100 < MW < 790$; $-2 < ALogP < 8$; $0 < AP < 1$). The vast majority of the data sits within a smaller region corresponding to the "druglike zone" as identified by Lipinski properties, since most of the compounds belonged to relatively advanced drug discovery projects.

Training and Test Set Selection. The global data set of 3563 compounds was split into training and test sets by computing fingerprints using OpenBabel.[30] The resulting data set was loaded into a Cerius2[29] study table, along with the experimentally determined solubility values. A subset of 1200 molecules was chosen via simultaneous optimization of structural diversity as encoded by the fingerprints and the solubility data profile. This subset was then used as the test set, and the remaining 2363 compounds were regarded as the training set for building the RP model. In particular, the Monte Carlo algorithm was

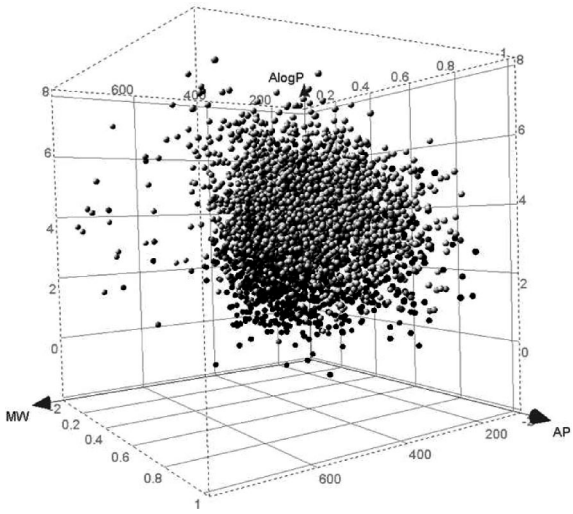

Figure 1. Distribution of aqueous solubility across the 3563 compounds in a 3D chemical space defined by MW, AP, and ALogP. Dots are coloured in continuum by experimental solubility at pH 7.4. Gray refers to 1 μM and black to 250 μM.

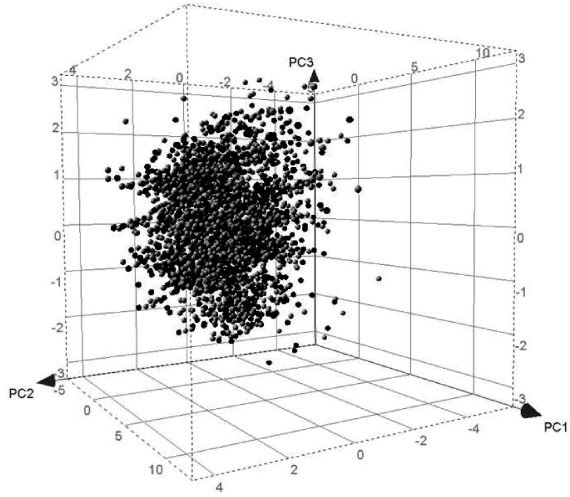

Figure 2. Training and test set compounds, as visualized in a PCA score plot for the first three components. The training set is shown as gray spheres, and the test set is shown as black spheres. The seven descriptors (AP, MW, AlogP, AP, PSA, HBA, HBD) calculated to derive the models were used to perform the PCA analysis.

applied to optimize an objective function consisting of two terms encoding for structural diversity and the solubility data profile. The former was evaluated by the MaxMin diversity function, whereas the latter was used as a penalty function. This was intended to afford a test set that was structurally representative and with similar solubility profile to the original set, aiming to reduce bias deriving from similar compounds in terms of both structure and response variable. A principal component analysis (PCA) was also performed by using a wider set of 2D descriptors within the Cerius2 software for visualizing this structural diversity in terms of physicochemical properties (Figure 2).

It is worth highlighting that the test set was not used in any model building but served as an independent test set used to evaluate the generality of the predictive power of each model.

The Descriptor Set. Seven descriptors were calculated and used as independent variables in the RP analysis: MW, ALogP, polar surface area (PSA), number of rotatable bonds (RBT), number of hydrogen bond donors (HBD), and number of hydrogen bond acceptors (HBA). An additional key descriptor, aromatic proportion (AP), was derived according to the procedure of Yan and Gasteiger.[8] The AP indicates the aromatic degree of a molecule and is defined as the ratio of the number of aromatic atoms to the total number of heavy atoms in the molecule. In order to calculate AP, the Daylight SMARTS definition of aromaticity was used, as derived from the unique SMILES code of the molecules in Accord.

RP Model Generation. Five RP models were generated corresponding to the threshold values used to split the data set into "insoluble" and "soluble" compounds (Table 1). Recursive partitioning decision trees were constructed using the quantitative structure–activity relationships (QSAR) module in Cerius2. In particular, the following settings were used. The Twoing metric was used to score the splitting of the tree. This is reported to give more balanced trees by splitting into nodes with roughly the same number of examples.[21] The tree was pruned backward through a moderate pruning process. Moreover, nodes had to contain a minimum of 1% of the samples to qualify for further splits. Finally, knot limits (i.e., the number of threshold values tested to split the range of each descriptor) per variable were set to 20 and the maximum depth of the tree was limited to 10, thus lowering the complexity of the resulting model.

During the five RP runs, each based on a different cutoff value, the algorithm identified the optimal number of descriptors out of the original seven that discriminated soluble and insoluble compounds. Thus, a variable number of descriptors may be involved in each RP output.

Results and Discussion

Inspired by the work of Delaney and co-workers,[20] we initially applied a linear regression to our proprietary data set, using the same basic simple descriptors. Results were not encouraging. While Delaney and co-workers managed to classify 72% of the compounds of their data set, in our hands, the same

Table 2. Results for the Five Generated RP Models over the Training and Test Sets

	cutoff value (μM)[a]									
	20		30		40		50		60	
	training set	test set	training set	test set	training set	test set	training set	test set	training set	test set
accuracy[b]	85.65	80.25	85.32	81.08	85.06	79.67	85.02	79.00	84.81	78.00
precision[b]	82.73	72.97	82.41	75.09	81.34	73.10	81.56	72.25	80.64	71.08
sensitivity[b]	85.19	79.58	83.86	81.56	82.33	81.78	81.57	82.91	79.79	83.81
specificity[b]	86.22	80.69	86.99	80.74	88.10	78.13	88.77	76.06	90.18	73.51
descriptors[c]	AP, MW, AlogP		AP, MW		AP, MW, AlogP, RTB		AP, MW, AlogP		AP, MW, AlogP, RTB	

[a] Threshold value used to split compounds of the data set into soluble and insoluble classes. See Methods for details. [b] Accuracy = (TP + TN)/(TP + TN + FP + FN). Precision = TP/(TP + FP). Sensitivity = TP/(TP + FN). Specificity = TN/(TN + FP). All are expressed as %. TP, TN, FP, and FN are true positive, true negative, false positive, false negative, respectively. [c] The reported descriptors were selected by the RP algorithm out of the original seven for building the model.

procedure and the same molecular properties (descriptors) afforded only about 40% of correctly predicted compounds, probably because of the different nature of the data set.

We thus changed to an approach more suited to our main goal, i.e., developing a model for predicting solubility of organic compounds as an aid to the decision of whether or not to synthesize or purchase a compound. Rather than assigning a predicted numerical solubility value to organic molecules, our main objective was to find the answer to the simple question, "Is this compound likely to be soluble?" RP appeared to be suitable technique to address this yes/no question.

RP is a statistical technique that can be used to find patterns in large data sets by building a decision tree. A data set is recursively divided into subsets depending on the value of a splitting criterion, such as the presence or absence of a feature. The feature is chosen so that the subsets are as diverse as possible according to a criterion such as, in this case, aqueous solubility.

In the initial stages of drug discovery, little or no information regarding pharmacological activity or structure–activity relationships (SAR) is usually available to drive compound prioritization, and computational tools for druglikeness, including solubility, should be applied to improve quality of the screening set. An additional objective of this work was to construct a model based on a small number of readily interpretable descriptors with the aim of capturing the molecular properties driving solubility.

As a result, seven descriptors including MW, RTB, HBA, HBD, PSA, ALogP, and AP were computed and five RP models were generated. During model development, the RP algorithm selected those descriptors out of the original seven able to give the best performance in discriminating the two classes.

The goodness of the models was assessed by calculating accuracy, precision, sensitivity, and specificity. Accuracy was defined as the proportion of all true results (correctly classified compounds) in the entire population. Precision reflects the ability of a decision tree to predict a specific class ("insoluble" in the current analysis). Sensitivity to the "insoluble" class is the probability that if a compound is insoluble (i.e., with a measured value below the cutoff threshold), the model will return it as a true positive. Sensitivity to the "soluble" class is instead equivalent to specificity, which is a measure of how well a model identifies the negative cases, or "soluble" compounds in this study. The average precision, sensitivity, specificity, and accuracy of all the models for the training set were 81.7%, 82.5%, 88.1%, and 85.2%, respectively. For the test set they were 72.9%, 81.9%, 77.8%, and 79.6%, respectively.

All models showed good performance, both in estimating and in predicting compounds of the training and test sets, respectively. Table 2 shows the calculated parameters for the five RP models.

Overall goodness of classification was constant across the five models, with only a slight decrement in the test set performance for the models based on 40−60 μM cutoff values. The best performing model was therefore able to correctly classify 85% of the compounds in the training set and, more importantly, 81% of the compounds in the test set. As our interest was the development of a model for identifying insoluble compounds, attention was focused more on precision and sensitivity.

In the training set, the precision increases when going from the lowest to the highest cutoff value. In the test set on the other hand, a clear maximum (75%) is discerned with the 30 μM cutoff, implying that model 30 is affected by 25% of false positives, i.e., soluble compounds erroneously assigned to the "insoluble" class.

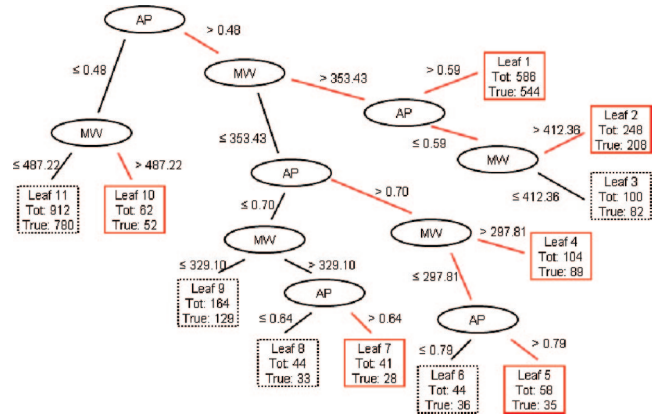

Figure 3. Selected classification tree for predicting insoluble compounds. AP and MW descriptors were used to classify compounds. Numbers are the splitting criteria. Leaves classifying "insoluble" compounds and the branches leading to their identification are highlighted in red, whereas "soluble" compounds are shown in black. Numbers in square boxes refer to the training set. The total number of training set compounds and the number of true positives (true insoluble in red squares and true soluble compounds in black squares) identified by each leaf are reported.

The sensitivity measure shows an opposite trend with respect to the precision within the training set, with the highest value associated with model 20 and the lowest value with model 60. Nevertheless, in the test set, model 20 is characterized by the lowest sensitivity value, increasing significantly for model 30 (increment of about 3%) and then further increasing in model 50 and model 60. In general, the percentage correctly predicted of insoluble compounds ranged from 79% to 84%. As a consequence, the percentage of false negatives ranged from 16% to 21%.

Our final model selection criterion was based on the best values of precision and sensitivity related to the test set. It is noted that all the models showed similar performances, independent of the solubility cutoff values and descriptors involved in the model generation. However, the model based on the threshold value of 30 μM appeared to be the most efficient in terms of false negatives (cost reduction in avoiding synthesis or purchase of insoluble compounds) and false positives (rejection of a potentially interesting compound) and was thus chosen as the insolubility prediction tool to be applied for the filtering purpose in our virtual screening cascade.

The model (Figure 3) is characterized by 5 branches and 11 terminal nodes (or leaves). Each leaf identifies a subset of compounds classified as "insoluble" (red squares, leaves numbers 1, 2, 4, 5, 7, 10) or "soluble" (black squares, leaves numbers 3, 6, 8, 9, 11).

The analysis of false positives (92 compounds) for the selected model (Figure 4) revealed that about 26% (28 compounds) of soluble compounds (i.e., solubility >30 μM) erroneously predicted to be insoluble (solubility <30 μM) had in fact solubility values lower than 60 μM, and 50% (46 compounds) had values lower than 100 μM. Thus, a substantial portion of errors were derived from compounds that nevertheless had limited solubility.

Conversely, about 25% (23 compounds) of false positives were indeed relatively soluble compounds, i.e., with an experimental solubility greater than 150 μM.

Figure 5 summarizes the results obtained for model 30. In particular, the match between experimental estimated/predicted

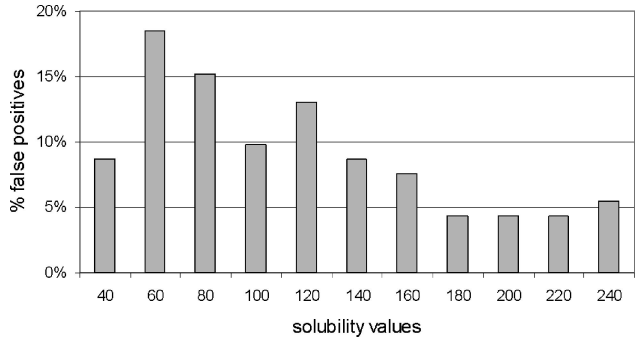

Figure 4. Distribution of false positives resulting from the model based on a cutoff value of 30 μM.

values is shown for each compound of the training and test sets, respectively. Because the test set was chosen to be representative of the entire data set and unbiased with respect to the training set, the leaf profile for the former shows a similar trend to the one for the latter, and the best performing leaves in the training set are the same as in the test set. This is particularly true for the leaves identifying insoluble compounds, which are those of key interest. It appears that leaves numbers 1, 2, and 4 showed a high percentage of correctly classified insoluble compounds in both the training and the test sets. Leaves numbers 8 and 9 performed relatively better in finding soluble compounds.

RP trees allow for identification of those leaves and relative branches that are characterized by the lowest error, and a useful application of this feature is to rely only on the prediction of such specific branches of the tree. In the selected model, most of the leaves correctly classify a high percentage of insoluble compounds with the lowest value of about 60%.

It should be emphasized that the selected model is based on only two descriptors: MW and AP. The classification tree finds the rules allowing for discriminating soluble and insoluble compounds by defining a set of criteria (ranges) depending on the MW and the aromaticity of the molecules.

The RP algorithm also explores three- (model 20 and model 50) and four-descriptor (model 40 and model 60) models in function of the cutoff values. Notably, adding descriptors does not appear to contribute significantly to the identification of insoluble compounds.

The correlation (Table 3) among the descriptors originally chosen to build the models was generally low. As expected, PSA shows a good correlation with HBD and RTB correlates well with MW. In the final models, PSA, HBA, and HBD were never involved as splitting criteria between soluble and insoluble compounds even though they are often associated with solubility of organic molecules. The number of RTB was investigated together with MW, despite their cross-correlation, in model 40 and model 60. These showed similar statistical values to model 20 and model 50 where RTB was not used. Hence, RTB seems to make no significant contribution to the description of insolubility with respect to MW.

Only four (MW, AP, AlogP, RTB) out of the original seven descriptors were selected by the algorithm to build the five models. Notably, they are the same descriptors used by Delaney et al. in their linear regression to model solubility. This underlines the central role of such descriptors in modeling this parameter. Moreover, MW and AP are present in every model, confirming their significant contribution to describe insolubility. The aromatic proportion of a molecule has previously been proposed[20] as a key property related to flexibility and melting point, while the molecular size is related to bulk molecular properties and hence their relation with solubility. The influence of MW on improving the estimation of water solubility has been also shown.[31]

Hypothesis Testing vs Validation. Model Follow-Up. In order to test the model predictivity using an additional external test set, 49 randomly chosen compounds that were not part of the original 3563 were selected.

The prediction values (Table 4) from the five models showed a good performance. The percentage of insoluble compounds was correctly predicted in a percentage ranging from 68% (model 60) to 93% (model 40) of cases, while their relative accuracy values in prediction ranged from 67% to 73%.

Assuming this test set was representative of a real project case, the method was able to discard up to 93% of the insoluble compounds. We consider this result to be an interesting illustration of the potential of this approach.

In this particular case, model 30 did not show the best performance (80% vs 93% in precision) as for the test set. This might have been due to the physicochemical properties of this set, which covered a restricted area of the wide descriptors space of the original set (Figure 6).

Conclusions

Out of the five models generated, the best performing RP tree was based on only two descriptors: AP containing information about the aromaticity of molecules; the molecular weight. The model was shown to have a high predictive power on the test set and suggests the utility of this approach in the early phase of the drug discovery process.

The results of the RP study should also be discussed in terms of the properties of the input data set. For example, (i) the performance of the two-descriptor model could be due to peculiarities of the data set, as seen in the case of the validation set of 49 compounds. (ii) Any generalization outside the boundary of the data set is speculative. While these are common limitations of any QSAR/QSPR method, in this case, a data set of over 3500 structures was used, and even though major structural classes within the data set may be identified as belonging to a specific project, the set encompasses a wide structural diversity, which will increase with the development of the compound collection. Rather then being an end point, the selected model is a working hypothesis that captures the solubility trends in an internal data warehouse. Thus, updating the data set will ensure a constant enrichment of the model in terms of generality and broader application.

In contrast to more complex models the current study was based on three distinct concepts. First, the property under scrutiny is insolubility. With the aim of saving money and time, insoluble compounds should be identified and rejected at the selection stage. Second, the choice of a basic model and simple descriptors; the Lipinski and Veber criteria owe their success to the intuitive quality of their descriptors, making classification trees easily accessible and interpretable. Third, we reasoned that when dealing with large numbers of compounds, the need for predicting numerical values for solubility would be less. The challenge is to roughly determine if a compound may be insoluble or not, or rephrasing the concept, if it is worth further investigation.

As such, this limited descriptor based model affords intuitive information regarding structural features influencing solubility. It is derived from simple 2D information, is computationally trivial, and provides a clear-cut output with a measurable accuracy and precision.

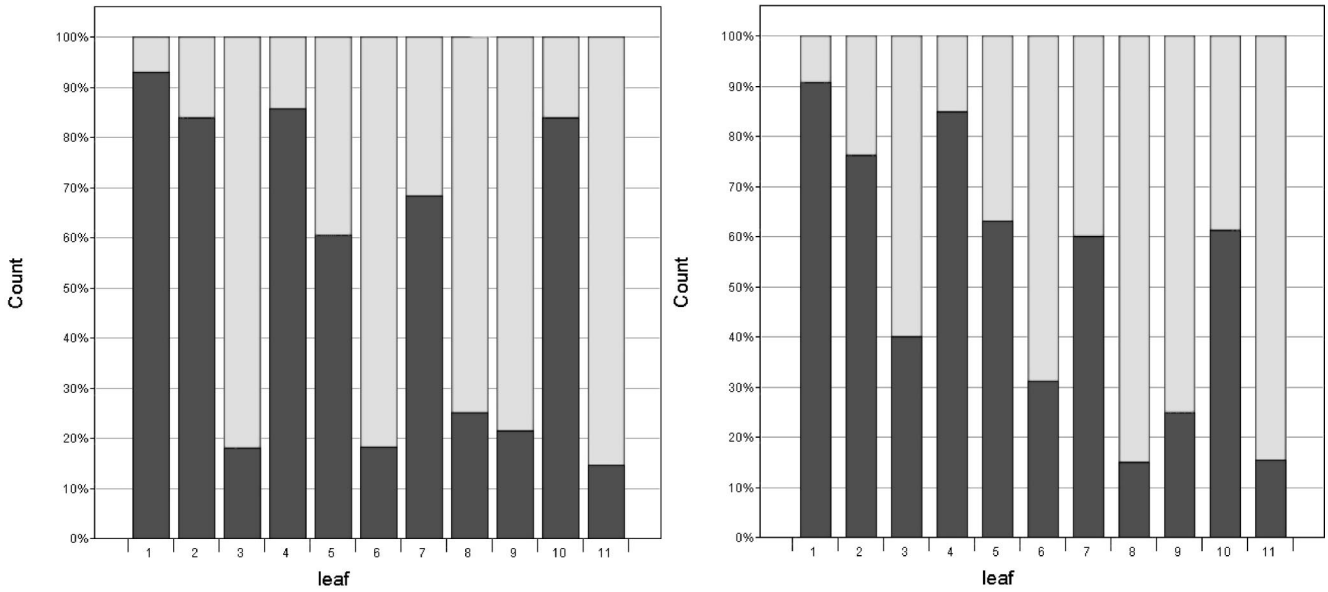

Figure 5. 100% stacked bars showing the percentage of correctly classified compounds per leaf (dark-gray for insoluble compounds, leaves numbers 1, 2, 4, 5, 7, 10; light-gray for soluble compounds, leaves numbers 3, 6, 8, 9, 11) for training (left) and test set (right).

Table 3. Correlation Matrix[a] between Descriptors Used for RP Modeling

	PSA	MW	RTB	HBA	HBD	AlogP	AP
PSA	1.00	0.39	0.49	0.53	0.72	−0.32	−0.17
MW	0.39	1.00	0.80	0.43	0.23	0.47	−0.18
RTB	0.49	0.80	1.00	0.32	0.44	0.17	−0.44
HBA	0.53	0.43	0.32	1.00	0.08	−0.03	−0.19
HBD	0.72	0.23	0.44	0.08	1.00	−0.25	−0.14
AlogP	−0.32	0.47	0.17	−0.03	−0.25	1.00	0.38
AP	−0.17	−0.18	−0.44	−0.19	−0.14	0.38	1.00

[a] All values are correlation coefficients calculated for the whole data set of 3563 compounds.

Table 4. Results for a Test Set of 49 Randomly Chosen Compounds

	accuracy, %	precision, %	sensitivity, %
model 20	67.35	80.65	57.14
model 30	67.35	80.65	57.14
model 40	73.47	93.94	72.43
model 50	69.39	79.41	50.00
model 60	67.35	68.57	45.00

The model is in routine use in our laboratories, and evaluation of its performance will be monitored in correlation to reducing the number of false negatives encountered in biological screening and overall savings in the iterations of compound design.

Acknowledgment. We thank Graeme Robertson for helpful discussions and editing and Russell Thomas and Chiara Ghiron for their invaluable feedback and suggestions. We also thank Matteo Andreini and Claus Andersen for technical support.

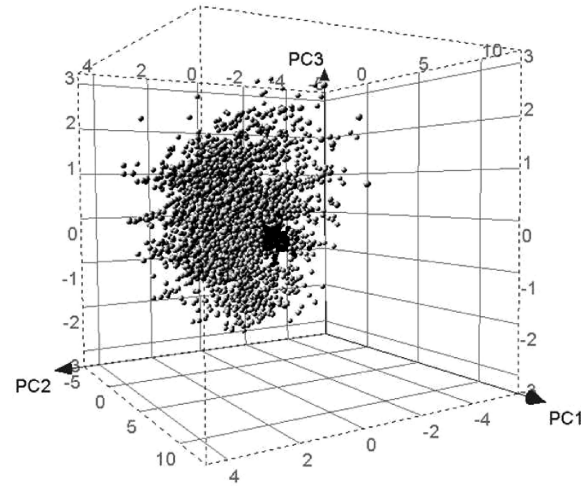

Figure 6. Variance of the original set structures (gray spheres) and the 49 compounds (black spheres) belonging to the small test set, visualized in a PCA score plot for the first three components.

References

(1) Bhattachar, S. N.; Deschenes, L.; Wesley, J. A. Solubility: It's not for Physical Chemists. *Drug Discovery Today* **2006**, *11*, 1012–1018.
(2) Maccari, L.; Andreini, M.; Benn, A.; Cesari, L.; Coniglio, S.; Fruscoloni, D.; Paoli, F.; Padova, A. Rational Approach to the Selection of a Diverse Set of Compounds for in Vitro Screening against CNS Therapeutic Targets Utilizing Nucleo, a Cheminformatic and Modelling Platform. Manuscript submitted.
(3) Bergstrom, 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*, 1477–1488.
(4) Schwaighofer, A.; Schroeter, T.; Mika, S.; Laub, J.; ter Laak, A.; Sulze, 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.
(5) 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.
(6) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds by Topological Descriptors. *QSAR Comb. Sci* **2003**, *22*, 821–829.
(7) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient by a Genetic Algorithm Based Descriptor Selection. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077–1084.
(8) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic

Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429–434.
(9) Liu, R.; Sun, H.; So, S. S. Development of Quantitative Structure−Property Relationship Models for Early ADME Evaluation on Drug Discovery. 1. Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1633–1639.
(10) Bruneau, P. Search for Predictive Generic Model of Aqueous Solubiltiy Using Bayesian Neural Nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.
(11) Yaffe, D.; Cohen, Y.; Espinosa, G.; Arenas, A.; Giralt, F. A Fuzzy ARTMAP Based on Quantitative Structure−Property Relationships (QSPRs) for Predicting Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1177–1207.
(12) Engkvist, O.; Wrede, P. High-Throughput, in Silico Prediction of Aqueous Solubility Based on One- and Two-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1247–1249.
(13) Yan, A.; Gasteiger, J.; Krug, M.; Anzali, S. Linear and Non Linear Functions on Modeling of Aqueous Solubility of Organic Compounds by Two Structure Representation Methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75–87.
(14) Frohlich, H.; Wegner, J. K.; Zell, A. Towards Optimal Descriptor Subset Selection with Support Vector Machines in Classification and regression. *QSAR Comb. Sci.* **2004**, *23*, 311–318.
(15) Lind, P.; Maltseva, T. Support Vector Machines for the Estimation of Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1855–1859.
(16) Xia, X.; Maliski, E.; Cheetham, J.; Poppe, L. Solubility Prediction by Recursive Partitioning. *Pharm. Res.* **2003**, *20*, 1634–1640.
(17) Eriksson, L.; Johansson, E.; Kettameg-Wold, N.; Wold, S. *PLS In Multi and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*; Umetrics AB.: Umea, Sweden, 1999; pp 69−112.
(18) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A New Tool for the Pharmacokinetic Optimization of Lead Compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29–S39.
(19) Catana, C.; Gao, H.; Orrenius, C.; Stouten, P. F. W. Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Model.* **2005**, *45*, 170–176.
(20) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
(21) Breiman, L.; Friedman, J. H.; Olshen, R. A. *Stone, Classification and Regression Trees*; Chapman & Hall: New York, 1984.
(22) 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.
(23) Veber, D. F.; Johnson, S. R.; Cheng, H-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.
(24) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. P. Predicting Penetration across the Blood−Brain Barrier from Simple Descriptors and Fragmentation Schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.
(25) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477–4486.
(26) Bellini, M.; Caradonna, N.; Coniglio, S.; Grava, C.; Marcucci, K.; Turlizzi, E.; Zanelli, U.; Westerberg, G. Compact Living. Combining High Throughput and High Content ADMET Profiling in a 30m^2 Laboratory. Presented at the Pharmaceutical Sciences World Congress, 2007.
(27) Kerns, E. H.; Di, L. Physicochemical Profiling: Overview of the Screens. *Drug Discovery Today: Technol.* **2004**, *4*, 343–348.
(28) Schrodinger, Quatro House, Frimley Road, Camberley GU16 7ER, U.K. (http://www.schrodinger.com).
(29) Software and documentation from Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752 (http://www.accelrys.com/).
(30) http://openbabel.sourceforge.net.
(31) Meylan, W. M.; Howard, P. H.; Boethling, R. S. Improved Method for Estimating Water Solubility from Octanol/Water Partition Coefficient. *Environ. Toxicol. Chem.* **1996**, *15*, 100–106.

JM701407X