A Hybrid Mixture Discriminant Analysis−Random Forest Computational Model for the Prediction of Volume of Distribution of Drugs in Human

Franco Lombardo,*,† R. Scott Obach,*,‖ Frank M. DiCapua,§ Gregory A. Bakken,§ Jing Lu,§ David M. Potter,‡ Feng Gao,‡ Michael D. Miller,§ and Yao Zhang‡

Molecular Properties Group, Pharmacokinetics, Dynamics and Metabolism, Computational Chemistry and Scientific Computing Groups and Groton Non-Clinical Statistics, Pfizer Global Research and Development, Groton Laboratories, Groton, Connecticut 06340

Received March 3, 2005

A computational approach is described that can predict the VD_{ss} of new compounds in humans, with an accuracy of within 2-fold of the actual value. A dataset of VD values for 384 drugs in humans was used to train a hybrid mixture discriminant analysis−random forest (MDA-RF) model using 31 computed descriptors. Descriptors included terms describing lipophilicity, ionization, molecular volume, and various molecular fragments. For a test set of 23 proprietary compounds not used in model construction, the geometric mean fold-error (GMFE) was 1.78-fold (±11.4%). The model was also tested using a leave-class out approach wherein subsets of drugs based on therapeutic class were removed from the training set of 384, the model was recast, and the VD_{ss} values for each of the subsets were predicted. GMFE values ranged from 1.46 to 2.94-fold, depending on the subset. Finally, for an additional set of 74 compounds, VD_{ss} predictions made using the computational model were compared to predictions made using previously described methods dependent on animal pharmacokinetic data. Computational VD_{ss} predictions were, on average, 2.13-fold different from the VD_{ss} predictions from animal data. The computational model described can predict human VD_{ss} with an accuracy comparable to predictions requiring substantially greater effort and can be applied in place of animal experimentation.

Introduction

The optimization of human pharmacokinetic behavior of new drugs is an important activity in the drug discovery process. Pharmacokinetics need to be appropriate for the target indication; appropriate pharmacokinetics lead to appropriate dosing regimens, which in turn yield greater patient compliance and ultimately improved efficacy. To ensure that new drugs offer convenience in dosing, many pharmaceutical companies will investigate the metabolism and pharmacokinetics of new chemical entities, in parallel to their efforts in medicinal chemistry and pharmacology. The half-life of a new drug will be a major factor in determining the dosing frequency. Compounds with short half-lives are more likely to require multiple administrations per day while those with longer half-lives tend to be more amenable to once-per-day dosing. The two pharmacokinetic parameters that determine the half-life are clearance (a measure of the rate at which the drug is removed from the body) and the volume of distribution (a measure of the extent of distribution from the plasma to the tissues).

The volume of distribution of a drug (VD at steady state or VD_{ss} throughout this work) is a function of the extent of drug partitioning into tissues vs that which remains within the plasma. Greater tissue partitioning, which requires that the drug can penetrate into tissues as well as bind reversibly to tissue components, will yield a greater VD_{ss}. With rare exceptions, the binding of drug to tissue components represents nonspecific binding to various macromolecular structures such as proteins, phospholipid membranes, etc. Nonspecific binding interactions are largely dictated by general physicochemical attributes of the drug, rather than specific pharmacophores. Thus, volume of distribution and physicochemical properties should be correlated in some manner, as has been demonstrated in previous reports by us and others.[1−3]

In this paper, we describe a computational approach to the prediction of human volume of distribution. Other methods to predict volume of distribution in humans have required the use of animal pharmacokinetic data.[4−7] Several models, reported in the literature, have taken either a fully computational approach or a hybrid approach[2,8,9] (i.e. partly experimental and partly computational) in order to predict VD_{ss}; these approaches do so either directly[9−12] or indirectly,[9] via the fraction unbound in tissues. Gathering such data can be resource intensive, requiring the synthesis of tens of milligrams of test compound, dosing and blood sampling from laboratory animals, the development of a sensitive and selective bioanalytical method (to measure the drug in plasma), and analysis of these plasma samples. A fully computational method does not require any synthesis or animal experimentation but only requires the time needed to draw the chemical structure, and the computer time for the required calculations. Furthermore, a computational approach can be applied to virtual chemical libraries. The method described in this paper approaches the accuracy of other previously described VD prediction methods and should prove to be a useful tool in drug discovery efforts while reducing the use of animals in research.

Results and Discussion

Volume of distribution is not a true physical volume, but rather a useful mathematical construct that describes the behavior of compounds in the body with regard to the degree of partitioning between the plasma compartment and the rest of the body. However, the VD value, determined from concentration−time curves for any given drug, is influenced by many physicochemical and physiological parameters, such as lipo-

* To whom correspondence should be addressed. Phone: (617) 871-4003; Fax: (617) 871-3078; e-mail: franco.lombardo@novartis.com.

† Molecular Properties Group. Present address: Novartis Institutes for BioMedical Research, Bldg. 600, 2C-357 250 Massachusetts Ave., Cambridge, MA 02139.

‖ Pharmacokinetics, Dynamics and Metabolism.

§ Computational Chemistry and Scientific Computing Groups.

‡ Groton Non-Clinical Statistics.

10.1021/jm050200r CCC: $33.50
Published on Web 03/14/2006

philicity, pK_a, tissue binding, and plasma protein binding.[1−3] Nevertheless, the question was posed as to whether such a complex construct, assuming a passive diffusion behavior, could be modeled directly and quantitatively using only computed parameters. The question of specific interactions of a drug with tissues or subcellular organelles, as well as with active transport systems, is not a trivial one, and in most cases the answer is not known. However, the assumption was made that in most cases a passive diffusion model would hold, and therefore a large data set of human VD data was gathered in an attempt to explore this approach.

Preliminary Models. Computed descriptors were identified and classified according to their covariance, and these were in turn subjected to multiple linear regression (MLR) forming self-consistent parameter sets (i.e. those that did not contain redundancies or were constant). The initial correlations obtained were modest (R^2 in the range of 0.5−0.6), but the diversity of the descriptors was significant and they were physicochemically and statistically meaningful. Examination of the results obtained with different models revealed that although the models exhibited comparable performance, the specific predictions for individual compounds were noticeably different, with the models predicting different subsets well. It was therefore reasoned that the complexity and diversity of effects contributing to VD_{ss} might not be amenable to reduction to a multiple linear approach with relatively few parameters.

The predictive performance of a model generated by the Cubist program[13] was assessed, in which piecewise multiple linear models (MLR) are fit in the nodes of a decision tree. A "committee" of multiple trees is built and averaged to produce a final prediction, using a boosting technique in which each subsequent tree attempts to improve the predictions from the previous trees. Following this approach, and utilizing a training set of approximately 300 compounds, several committee models were developed. These generally consisted of 18−20 trees and produced models that had R^2 of about 0.88 and Q^2 of about 0.80. This preliminary result was positive, and further testing of the prediction ability of these models confirmed the high likelihood that a reasonably good VD_{ss} model could be derived utilizing only computed parameters. However, in building these preliminary models, some of the parameters used may be termed "secondary", i.e., they were computed from other in-house models, and, more importantly, some of the VD data were from oral administration of the drug or obtained from secondary sources. Having obtained reasonable confidence that a computational model for VD was feasible, we decided to improve the approach. We therefore sought to expand and refine the data set, more thoroughly assess computed parameters (seeking to include only those parameters that can be deemed primary), and to assess multiple statistical modeling procedures in addition to the previously applied MLR and Cubist approaches.

Construction of the Hybrid Mixed Discriminant Analysis−Random Forest Model. A more extensive literature mining effort produced a set of 384 compounds containing exclusively iv clinical data, which were included only after careful scrutiny of the original literature (see Experimental Section). For a small number of compounds VD_β, i.e., the volume of distribution during the terminal elimination phase, rather than VD_{ss}, was used. The next step was to eliminate the secondary computed parameters mentioned above. However, it is not straightforward to decide whether a parameter can be defined as primary since the calculation can be derived from statistical analysis of thousands of fragmental values. Nevertheless we used the stability and availability of the parameters as one of the guiding factors, keeping the quality and breadth of VD_{ss} data as the primary factor.

Table 1. The 31 Descriptors Used in the Present Model

descriptor	description
frac_anion_7	ACDLabs fraction anionic at pH 7
frac_cation_7	ACDLabs fraction cationic at pH 7
rule12	*:,=[c,C][N;H0,H1][C;H1,H2,H3]
rule39	[!c;!C]~[!c;!C;H1]
rule40	[!c;!C]~N
rule64	[!H]~[!H](~[!H])(~[!H])~[!H]
rule85	[c,n]1o[c,n][c,n][c,n]1
rule91	[CH2,CH3]~[!C;!c]~[CH2,CH3]
rule98	[CH2](~*~[N,n])~*
rule108	[CH2]~[N,n]~*
rule114	[CH3][!H][CH2][!H]
rule155	[N,n]=,:[C,c;H1][N,n;H0,H1]
rule187	[O,o]~[C,c](~[C,c])~[C,c]
rule193	[O]~[S,P](~O)(~O)[!O]
rule205	[R;N,n,O,o,S,s]
rule211	[S,s]
rule288	I
rule308	N~C~O
rule347	S~[!H](~[!H])~[!H]
isis75	A!N$A
isis84	NH2
isis92	OC(N)C
isis96	5M RING
isis128	ACH2AAACH2A
INTHB	measure of internal hydrogen bonding ability
PEOE_PC_+	total positive charge (MOE v. 2004.3)
PEOE_RPC_+	relative positive charge (MOE, v. 2004.3)
ClogP	calculated logP (BioByte, ClogP v 4.1)
dXp10	simple difference chi index using order 10 paths (molconnz)
Gmin	smallest e-state value (molconnz)
nPag22	vertex alpha-gamma count (molconnz)

[a] The following descriptors were identified by simulated annealing: frac_anion_7, frac_cation_7, rule39, rule64, rule91, rule108, rule114, rule155, rule288, isis75, isis84, isis92, isis96, isis128, INTHB, ClogP, dXp10, gmin, nPag22.

The structures for the 384 compounds were generated as described in the Experimental Section, and 1149 parameters were calculated for each molecule. The matrix of 384 × 1149 data points was then analyzed using various statistical approaches. These included Random Forests (RF), Cubist, Partial Least Squares (PLS), Multiple Linear Regression (MLR) alone and in conjunction with Simulated Annealing (SA-MLR), and Principal Component Regression (PCR). Previous findings,[2−3] physicochemical intuition, and statistical measures were used to assess the likely influence of the chosen computed parameters on the MLR and PCR ability to predict VD_{ss}.

After elimination of descriptors with zero variance, 952 parameters remained. We further reduced the set by eliminating the descriptors with high pairwise correlations with other descriptors in the set, using a cutoff (R^2) of 0.8, and retained the member of each pair that was more easily computed. These steps left 550 descriptors that were used as input for the SA-MLR algorithm.

The SA-MLR algorithm (see Experimental Section) provided a subset of 20 descriptors representing the minimal fitness value encountered during the analysis and using a MLR approach as scoring function. This subset of 20 descriptors was combined with a subset of 16 descriptors selected independently based on their likely influence on VD_{ss}. The combined subset contained 31 unique descriptors. This subset of 31 unique descriptors was used in model generation, and they are reported in Table 1.

The MLR approach, with parameters derived from simulated annealing and physicochemical intuition, proved reasonably

successful in indicating subsets of relevant variables. While the correlation observed with 31 of the computed parameters was not very high ($R^2 = 0.69$, $Q^2 = 0.61$, log VD_{ss} as the dependent variable), randomization tests (as well as the independent test sets) demonstrated that this approach was fairly robust, satisfactorily predicting VD_{ss} values for "unknown" compounds. For example, we processed an independent test set of 23 proprietary compounds (all with clinical iv pharmacokinetic data), and obtained a GMFE of 2.01; this compares favorably with the GMFE for the training set ($N = 384$), which was 1.88.

It is also important to note that the continuous parameters described in Table 1 had MLR coefficients that made physicochemical sense (see Supporting Information). For example, it would be expected that ClogP would have a positive coefficient since a high lipophilicity was reported to directly correlate with VD_{ss}. Similarly the cationic fraction of the molecule at pH 7 would also be expected to correlate directly as well.[2,3] Furthermore, since the tight binding of anions to serum albumin tends to reduce VD_{ss} values, we expected that the anionic fraction of the molecule (at pH 7) to inversely correlate with VD_{ss}; this was indeed the case. Similarly, the intramolecular hydrogen-bond descriptor (INTHB) should be directly correlated with VD_{ss}, owing to the reduction of the polar surface area that would allow a better penetration through membranes, and the two partial charge descriptors (PEOE_PC_+ and PEOE_RPC_+) should be, on polarity arguments, inversely correlated with the dependent variable; this was observed as well. Another parameter, originally identified by the SA-MLR approach i.e., the E-state of quaternary N atoms, was eventually dropped upon further analysis, thus reducing the number of parameters to 19.

Considering that the "chemical intuition/MLR" approach overlapped with the SA results, we contend that the choice of the 31 parameters is appropriate, even though, in some instances, a particular chemotype may have had a large impact, as in the case, for example of rule 288 (count of I atoms). However, when 8 out of the 10 parameters that did not reach 95% significance in the MLR (see Supporting Information) were eliminated on the basis of their statistical significance in a stepwise-regression analysis, rule 288 was still retained as significant, and the very high significance of ClogP and fraction ionized at pH 7 (both anionic and cationic) was confirmed. The isis 75 and isis 92 fragments were retained because they acquired significance in the 23 parameter equation. The statistical parameters and the predictive tests performed on the latter equation were essentially identical to the equation using 31 parameters.

However, due in part to the paucity of VD_{ss} data in the upper end of the range of the values, and to the complexity of the target variable being modeled, a significant curvature in the plot of predicted vs experimental VD_{ss} was observed (data not shown), with the latter being underpredicted at high values. Similarly, efforts using PLS and Cubist-based approaches, despite a fairly broad exploration of the options offered by Cubist and the use of variables not included in the subset of 31, did not lead to a significant improvement over the MLR model.

A two-stage statistical approach, using mixture discriminant analysis (MDA) as the first step and random forests (RF) as the second was attempted using the 31 descriptors. The hybrid model was developed to overcome the bias toward lower VD_{ss} values in the predictions from the single RF model, which is due to the fact that the number of compounds with lower VD_{ss} values significantly outweighs the number of compounds with larger VD_{ss} values in the training set. The resulting hybrid model performs somewhat better than the pure RF model in predicting the independent clinical compounds (with GMFE 1.78 vs 2.03) especially in the case of compounds with relatively high VD values. The overall errors in cross validation of the training set are roughly the same (average GMFE $\sim$ 2.0). By first defining a low volume of distribution as <10 L·kg^{-1} and a high volume of distribution as ≥ 10 L·kg^{-1}, a two-category classification model was pursued. Mixture discriminant analysis (MDA) was applied at the first step and random forests (RF) at the second. Separate random forest regression models on the low and high VD_{ss} subsets of compounds were constructed. To improve the predictive performance on compounds near the 10 L·kg^{-1} boundary, the RF model for high VD_{ss} compounds was built on compounds with $VD_{ss} \geq 5$ L·kg^{-1}, and the RF model for low VD_{ss} compounds was built on all compounds. All random forest modeling was done on the log VD_{ss} scale. In application of this model for predicting VD of a new compound, it is first classified as a low or high VD_{ss} compound using the MDA classification model. The corresponding RF regression model is then used to provide the predicted VD_{ss} value. A 500-tree RF-model for each category was built resulting in a total of 1000 trees overall. The model yielded a R^2 value of 0.91 for the training set of 384 compounds. The observed geometric mean fold-error (GMFE) calculated for the (linear) VD_{ss} data used in the training set was 1.37 ($\pm$1.7%), i.e., well below a generally accepted threshold value of two.[4,7] The plot of the predicted vs observed VD_{ss} value for the training set is shown in Figure 1. An increase in the number of trees did not prove fruitful, and the above results prompted testing of the predictive ability of the model using several approaches. A comparison of summary statistics for the predictions of VD on a test set of 23 proprietary by MDA-RF, RF, and MLR models is listed in Table 2. These data support the notion that the hybrid MDA-RF model yields the highest performance.

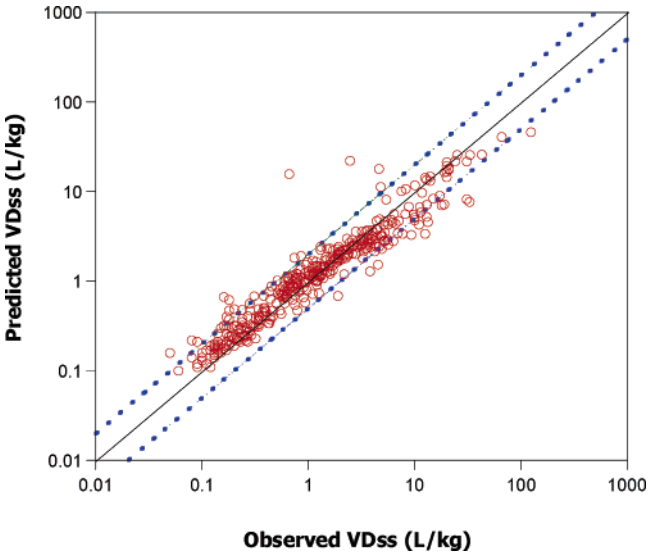

Figure 1. Plot of predicted VD_{ss} vs observed VD_{ss} for the 384 compounds in the training set. The dotted lines represent the 2-fold error limits.

Performance of the Hybrid MDA-RF Model in the Prediction of Human VD_{ss}. A 10-fold cross validation was conducted on the training data set with the MDA-RF model. The 384 compounds were randomly divided into 10 groups with roughly equal sizes. The model and predictions were then run 10 times, each time with a model built on 9 of the 10 groups combined and predictions made on the group that was not in the model. The process was iterated so that each group

Table 2. Summary Statistics for the Prediction of VD Values for 23 Proprietary Compounds Using the MDA-RF Model as Compared to RF and MLR Models

model	GMFE[a]	%CV[b]
Hybrid Mixed Discriminant Analysis−Random Forest	1.78	11.4
Random Forest only	2.03	15.0
multiple linear regression (all 31 combined descriptors)	2.01	11.4
multiple linear regression (19 descriptors from simulated annealing)	2.05	10.1
multiple linear regression (16 descriptors based on physicochemical intuition)	2.48	16.2

[a] GMFE: geometric mean-fold error [b] %CV: percent coefficient of variation of GMFE

Table 3. Summary Statistics of the Cross-Validation of the 384 Compound Training Set

	group									
	1	2	3	4	5	6	7	8	9	10
sample size	38	39	39	39	39	38	38	38	38	38
prediction GMFE	1.83	1.81	2.08	2.17	2.12	2.21	2.24	2.1	1.93	2.02
%coeff of variation error of GMFE	7.3%	7.5%	9.2%	12.2%	13.4%	14.0%	13.2%	12.5%	9.2%	10.9%

Table 4. Leave-Class-Out Analysis on the 384 Compounds Training Set

structural class	analogues in class	GMFE	%CV
steroids	14	1.71	13.1
β-blockers	16	1.70	9.3
fluoroquinolone antibiotics[a]	10	2.94	9.6
NSAIDs[a]	7	2.67	8.7
cephalosporines	17	1.46	5.7
benzodiazepines	15	1.61	7.2
tricyclic antidepressants	7	1.85	10.8
morphine-like[a]	10	2.26	21.5

[a] All compounds were predicted with a GMFE < 2 in the general model (N = 384).

Table 5. Prediction Accuracy vs Clinical iv Data

method	no. of proprietary compounds	GMFE
average of PK-based models[a]	20	1.61
experimental physicochemical model[b]	18	2.26
computational method (MDA-RF)[c]	23	1.78

[a] Reference 4. Average of two or three methods. [b] Reference 3. This prediction test used the same compounds as in the previous row with the exception of two acidic compounds not amenable to prediction via the method reported in ref 3. [c] This work. The set of 23 compounds includes all 20 (or 18) compounds used in the two studies reported above. The coefficient of variation on the GMFE for the MDA-RF model was 11.4%.

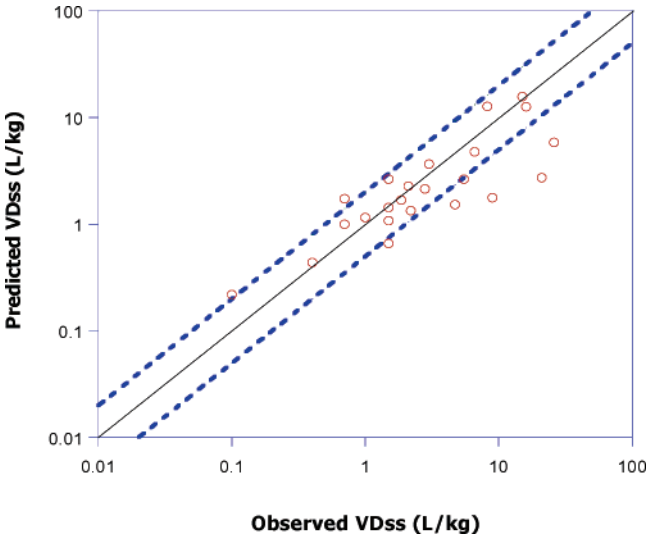

Figure 2. A. Plot of predicted VD_{ss} vs observed VD_{ss} for the 23 compounds in the clinical test set using the MDA-RF model. The dotted lines represent the 2-fold error limits.

has been independently predicted once. The results are summarized in Table 3.

A leave-class-out (LCO) approach, as described in the Experimental Section, was applied to test whether any class would be poorly predicted if it had not been included in the model, a priori. This can be thought of as a "simulation" of real-life prediction work when a new structural class is at hand and for which no experimental data is available. The results of LCO are shown in Table 4. While some classes, once excluded from the training set, did not yield a GMFE at or below 2, each member of any of the classes was predicted well within a factor of 2 using all training set data (N = 384) and the weighed average was 1.91, across all classes, in the LCO test.

A third, and perhaps more stringent, test was represented by the use of proprietary clinical iv VD_{ss} data encompassing a fairly wide range of structural classes and covering a range of VD_{ss} data equal to 96% of the range of data in the training set. This test sought to compare, among them, the animal PK-based approaches, described by Obach et al.,[4] to the physicochemical approach reported by Lombardo et al.[2,3] and the present computational model. The results, shown in Table 5, attest to an excellent predictive performance of the present method, yielding a GMFE value comparable to methods based on (the much more expensive and time-consuming) in vivo data. It also showed a better performance than the approach using the correlation between physicochemical data and fraction unbound in tissues as calculated using the Oie−Tozer equation.[14] The different number of compounds in each set is due to the lack of availability of the corresponding animal data for the PK-based approaches and to the applicability of the physicochemical model to basic and neutral compounds only. The 18 compounds used in the latter case and the 20 compounds used to test the animal PK-based methods were part of the set of 23 compounds used in the present work, and so there was a complete overlap of the latter set with the other two. The plot of the predicted vs observed VD_{ss} value for the test set is shown in Figure 2, and, in analogy with the plot for the training set, the dashed lines represent the factor of two thresholds.

In a fourth test of model performance, the predictions of human VD_{ss} made for over 70 compounds, using previously described animal-based prediction methods[4] vs predictions made using the present MDA-RF model were compared. The computational model was compared to each of the animal-based methods described as well as to the average of all three of them, as shown in Table 6. It can be seen that the predictions using the MDA-RF model, using nothing more than a structural input to calculate a VD_{ss} value, are on average at about a factor of 2 from the experimental PK-based predictions made for these compounds. Similar to the previous analysis on clinical data, the number of compounds reported for the allometric scaling

Table 6. Comparison with Animal PK-Based Predictions

comparison[a]	no. of proprietary compounds	GMFE
MDA-RF vs average animal fu(tissue) method	74	2.08
MDA-RF vs dog−human proportionality	74	2.16
MDA-RF vs allometry[b,c]	71	2.57
MDA-RF vs average of all methods	74	2.13[d]

[a] For a discussion of the accuracy of animal PK-based methods see ref 4. These three methods are referred to as V1, V2, and V3, respectively, in ref 4. [b] Corrected for fu differences. [c] Data for three compounds not available. [d] The coefficient of variation was 7.2%.

approach is slightly lower than the other methods due to lack of an allometric prediction for three of the compounds used. The average value reported is also based on only two methods for the same three compounds.

Conclusion

We have developed a computational VD_{ss} model based on carefully checked clinical pharmacokinetic data for 384 drugs and tested its predictive ability using several statistical approaches, including an external test set. The ability of this model to accurately predict human VD is essentially identical to previously described animal-based approaches, but the method offers a suitable way to spare costly synthetic and analytical resources as well as to reduce the use of animals, therefore positively impacting on its broader implications. Only a structural input is required, which makes the method amenable to virtual screening and certainly useful for a drastic reduction of in vivo experiments. Our work will continue with tests and modifications of the current model based on newer data, and we will also explore more in detail the impact of structural fragments on the prediction outcome.

Experimental Section

Volume of Distribution Data. The volume of distribution data for the 384 compounds used in the training set were gathered by examining more than 600 original references. All data are from reported studies in human, and in all cases the drug was administered intravenously. We considered with particular attention the bioanalytical techniques used for plasma drug concentration determination, e.g. whether the parent compound was identified and followed during the course of the study, by what means, whether total radioactivity was used, and other potential issues. If each of those concerns could not be satisfactorily resolved, the data would not be accepted. In about 10% of the cases the VD_{ss} values were calculated from concentration vs time plots reported by the original authors, that were digitized to yield concentration−time data (www.digitizeit.de) or from a data table, using WinNonLin v. 3.2 (Pharsight Co., Mountain View, CA). In all cases the obtained PK parameter values, other than VD_{ss}, e.g. Cl and/or VD_{β}, were compared with the reported values to ensure quality in the digitization and calculation steps. About 10% of the data used were VD_{β} data that were considered suitable, in the absence of a VD_{ss} value or of available mean concentration vs time plots. Finally, in a few cases, the VD_{ss} values were calculated from reported pharmacokinetic micro-constants using known PK relationships such as: $VD_{ss} = V_c(1 + k_{12}/k_{21})$, where V_c is the volume of the central (or plasma) compartment and k_{12} and k_{21} are the distribution rate constants for transfer from central to peripheral compartment and from peripheral to central compartment, respectively.

Computational Approach. The structure of each molecule was obtained from the Derwent World Drug Index (WDI) database and converted into 3D format using the Tripos implementation of Concord 5.1.1. For each molecule in the dataset, we calculated a number of physical properties that have been shown to be important in influencing the effectiveness of drug-like compounds. Three of the Lipinski properties (molecular weight and number of hydrogen bond donors and acceptors) were calculated via in-house software. CPSA terms were calculated using SAVOL2,[15] and the ClogP values were calculated using version 4.1 of the well-known package (BioByte, Claremont, CA). The LogD, LogP and pK_a values were calculated using the ACDLabs v.8.0 (ACD/Labs, Toronto, Canada). Additional information indices, describing the molecular connectivity, shape, and E-states of the molecules were calculated using the Tripos implementation of Molconn-Z (Sybyl, v 7.0, Tripos, St. Louis, MO). The public ISIS keys were also used to break down each of the molecules into their respective fragments, allowing us to probe for functionality that might influence the drugs volume of distribution values. These keys represent the presence or absence of the functionality described, while the "rules" based on SMARTS strings developed in-house by Dr. M. Tu (Pfizer, Groton) are counts of actual occurrence of the fragment represented. The PEOE charges were computed using MOE v. 2004.3,[16] and, finally, the calculation of the INTHB descriptor, representing the propensity for internal hydrogen bonding was performed via an algorithm developed in-house. In all, 1149 descriptors were generated for each molecule.

Statistical Methods. Simulated annealing[17] was used to search for subsets of descriptors useful in modeling VD_{ss}. The simulated annealing algorithm initially selected a random subset of 20 descriptors. The fitness measure of the subset was the training set root-mean-square error based on a linear regression (MLR). The descriptor set was perturbed, by replacing a single descriptor with another descriptor, and a new fitness value calculated. The fitness values were compared to determine if the fitness measure had improved. If the fitness measure improved, the new descriptor subset was accepted and the iterative process continued. If the fitness measure did not improve, the acceptance of the new descriptor subset was based on a Boltzmann probability distribution to allow escape from local minima. As the number of iterations increased, the probability of accepting a detrimental step decreased.

Mixture Discriminant Analysis[18] is an extension of Linear Discriminant Analysis (LDA). In LDA, a new compound is predicted to be a member of the "nearest" class, where the distance is based on assuming a normal distribution for the descriptors, and for which it is assumed that the variability and correlation among the descriptors is the same in each class. MDA is one of several extensions to LDA, in which we allow multiple normal distributions or "prototypes" within each class.

Random Forests[19,20] is a tree-based method, which comprises two key components. First, multiple trees are generated using bootstrap resampling of the data, and the predictions from the individual trees are averaged to obtain a single prediction for each compound. Within each tree, and for each individual node of the tree, random subsets of the predictors are chosen, from which the single best predictor is chosen on which to split that node. We constructed MDA and Random Forest models using R statistical packages[21] (v1.9.1) with their default model tuning parameters.

Leave-Class-Out Approach. To test the methods for its performance on particular classes of analogues, we identified several classes of structurally similar compounds, and for each leave-class-out analysis we generated two new 500-tree models, keeping, in turn, each class out of the training set. The predictive power of each model, on the class not included in its generation, was then assessed. We limited the approach to a few classes for which we had a reasonable number of analogues, and the results are presented and discussed in the Results and Discussion section.

Acknowledgment. We wish to acknowledge Dr. Michael Fisher, (PDM, PGRD Groton Laboratories) for interesting discussions and application of this computational model to project work. We would also like to thank L. Powers, H. Siemon, P. J. Smith, and D. Zyry (Library, PGRD Groton Laboratories) for their assistance with data searches and retrieval,

and Mr. Dong Li (Scientific Computing, PGRD Groton Laboratories) for his help with technical issues that arose during this effort.

Supporting Information Available: Example of RF tree, MLR coefficients for the parameters used and complete statistics. Complete list of VD data and computed parameters for the training set compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Poulin, P.; Thiel, F.-P. Prediction of Pharmacokinetics Prior to *In Vivo* Studies. 1. Mechanism-based Prediction of Volume of Distribution. *J. Pharm. Sci.* **2002**, *91*, 129−156.
(2) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of Volume of Distribution in Humans for Neutral and Basic Drugs Using Physicochemical Measurements and Plasma Protein Binding Data. *J. Med. Chem.* **2002**, *45*, 2867−2876.
(3) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F. Prediction of Volume of Distribution in Humans for Neutral and Basic Drugs Using Physicochemical Measurements and Plasma Protein Binding Data. *J. Med. Chem.* **2004**, *47*, 1242−1250.
(4) Obach, R. S.; Baxter, J. G.; Liston, T. E.; Silber, B. M.; Jones, B. C.; MacIntyre, F.; Rance, D. J.; Wastall, P. The Prediction of Human Pharmacokinetic Parameters from Preclinical and In Vitro Metabolism Data. *J. Pharm. Exp. Ther.* **1997**, *283*, 46−58.
(5) Ward, K. W.; Smith, B. R. A Comprehensive Quantitative and qualitative Evaluation of Extrapolation of Intravenous Pharmacokinetic Parameters from Rat, Dog and Monkey to Humans. II. Volume of Distribution and Mean Residence Time. *Drug Met. Dispos.* **2004**, *32*, 612−619.
(6) Mahmood, I. Interspecies scaling: predicting volumes, mean residence time and elimination half-life. Some suggestions. *J. Pharm. Pharmacol.* **1998**, *50*, 493−499.
(7) Caldwell, G. W.; Masucci, J. A.; Yan, Z.; Hageman, W. Allometric Scaling of Pharmacokinetic Parameters in Drug Discovery: Can human CL, V_{ss} and $t_{1/2}$ be predicted from In-vivo Rat Data? *Eur. J. Drug Met. Pharmacokin.* **2004**, *29*, 133−143.
(8) Wajima, T.; Fukumura, K.; Yano, Y.; Oguma, T. Prediction of human pharmacokinetics from animal data and molecular structural parameters using multivariate regression analysis: volume of distribution at steady state. *J. Pharm. Pharmacol*. **2003**, *55*, 939−949.
(9) Lombardo, F.; Obach, R. S.; Shalaeva, M. Y.; Gao, F.; Miller, M. D. Experimental and computational prediction of volume of distribution in humans for neutral and basic drugs II. Use of physicochemical and plasma protein-binding data or computed parameters. In *EUROQSAR 2002* Designing Drugs and Crop Protectants: processes, problems and solutions. Ford, M.; Livingstone, D.; Dearden, J.; van de Waterbeemd, H., Eds.; Blackwell Publishing Ltd., Oxford, UK 2003; pp 211−214.
(10) Hunt, P. SIMCA and its application to Human Clinical PK Data. Poster presented at the EUROQSAR 2002, Bournemouth, UK. September 8−13, 2002.
(11) A computational volume of distribution model is implemented in VolSurf, v.4.1.3. The program is commercially available from http://www.moldiscovery.com (accessed August 26, 2005).
(12) Ghafourian, T.; Barzegar-Jalali, M.; Hakimiha, N.; Cronin, M. T. D. Quantitative structure-pharmacokinetic relationship modelling: apparent volume of distribution. *J. Pharm. Pharmacol*. **2004**, *56*, 339−350.
(13) http://www.rulequest.com (accessed August 26, 2005).
(14) Øie, S.; Tozer, T. N. Effect of Altered Plasma Protein Binding on Apparent Volume of Distribution. *J. Pharm. Sci.* **1979**, *68*, 1203−1205.
(15) Skell, J. M.; Pearlman, R. S. *SAVOL2*. University of Texas, Austin, TX, 1988.
(16) http://www.compchem.com (accessed August 26, 2005).
(17) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(18) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning*; Springer-Verlag: New York, 2001; pp 399−405.
(19) Breiman, L. Random Forests. *Machine Learning* **2001,** *45(1)*, 5−32.
(20) 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*, 1947−1958.
(21) R Development Core Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2004; ISBN 3-900051-00-3; http://www.R-project.org (accessed August 26, 2005)

JM050200R