ADME Evaluation in Drug Discovery. 7. Prediction of Oral Absorption by Correlation and Classification

Tingjun Hou,*,† Junmei Wang,‡ Wei Zhang,‡ and Xiaojie Xu‡

Department of Chemistry and Biochemistry, Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, California 92093, and College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, P. R. China

Received August 7, 2006

A critically evaluated database of human intestinal absorption for 648 chemical compounds is reported in this study, among which 579 are believed to be transported by passive diffusion. The correlation analysis between the intestinal absorption and several important molecular properties demonstrated that no single molecular property could be used as a good discriminator to efficiently distinguish the poorly absorbed compounds from those that are well absorbed. The theoretical correlation models for a training set of 455 compounds were proposed by using the genetic function approximation technique. The best prediction model contains four molecular descriptors: topological polar surface area, the predicted distribution coefficient at pH = 6.5, the number of violations of the Lipinski's rule-of-five, and the square of the number of hydrogen-bond donors. The model was able to predict the fractional absorption with an $r = 0.84$ and a prediction error (absolute mean error) of 11.2% for the training set. Moreover, it achieves an $r = 0.90$ and a prediction error of 7.8% for a 98-compound test set. The recursive partitioning technique was applied to find the simple hierarchical rules to classify the compounds into poor (%FA ≤ 30%) and good (%FA > 30%) intestinal absorption classes. The high quality of the classification model was validated by the satisfactory predictions on the training set (correctly identifying 95.9% of the compounds in the poor-absorption class and 96.1% of the compounds in the good-absorption class) and on the test set (correctly identifying 100% of the compounds in the poor-absorption class and 96.8% of the compounds in the good-absorption class). We expect that, in the future, the rules for the prediction of carrier-mediated transporting and first pass metabolism can be integrated into the current hierarchical rules, and the classification model may become more powerful in the prediction of intestinal absorption or even human bioavailability. The databases of human intestinal absorption reported here are available for download from the supporting Web site: http://modem.ucsd.edu/adme.

INTRODUCTION

Because of the fact that the failure of many compounds in the development stage is caused by unfavorable absorption, distribution, metabolism, and excretion (ADME) properties, more and more efforts are put to the field of ADME predictions.[1,2] As an alternative to experimental measurements, the in silico prediction of ADME properties is very attractive, because it provides an inexpensive and high-throughput way to assess the ADME properties of a molecule prior to synthesis and biological testing. Among ADME properties, good oral bioavailability is one of the most desirable attributes of a new drug. The prediction of oral bioavailability is very challenging due to the fact that bioavailability is a complex function of many biologic and physicochemical factors, such as dissolution in the gastrointestinal tract, intestinal membrane permeation, intestinal and hepatic first-pass metabolism, and even the dosage form. On the current stage, major efforts are focused on the prediction of human intestinal absorption (HIA), because the first step for obtaining a high oral bioavailability is to achieve a good oral absorption.

In experiments, intestinal absorption is usually measured by fraction absorption, %FA, which is defined by the total mass absorbed divided by the given dose of the drug. The field on the prediction of oral absorption might have been pioneered by the rule of five proposed by Lipinski and coworkers.[3] The rule of five defined several rules for identifying compounds with possible poor absorption and permeability: (1) molecular weight > 500, (2) calculated log P > 5 (CLOGP) or > 4.15 (MLOGP), (3) number of hydrogen-bond donors (OH and NH groups) > 5, and (4) number of hydrogen-bond acceptors (N and O atoms) > 10. A disadvantage of the rule of five is that it can only give a quite rough classification of molecules, allowing the elimination of only a very limited set of molecules. Since then, numerous classification and regression prediction models for the predictions of HIA were reported by applying a variety of statistical and machine-leaning approaches, which include multiple linear regression,[4] nonlinear regression,[5] partial least squares regression,[6] linear discriminant analysis,[7] classification and regression trees,[8] artificial neural networks (ANNs),[9] genetic algorithms (GAs),[9] support vector machines (SVMs),[10] and so forth. Because of the fact that many factors are related to intestinal absorption, many physicochemical descriptors were introduced into the prediction of HIA, such as polar

* Corresponding author e-mail: tingjunhou@hotmail.com.
† University of California at San Diego.
‡ Peking University.

10.1021/ci600343x CCC: $37.00
Published on Web 11/09/2006

surface area (PSA), partition coefficients, molecular size, hydrogen-bonding descriptors, topological descriptors, and even quantum chemical descriptors. The detailed descriptions of the prediction models were reviewed in some recent articles.[11−13] Please note that all reported models could only deal with molecules transported by passive diffusion. Passive diffusion is the major route for drug molecules permeating through cell membranes in the intestine; however, other diffusion mechanisms may play an important role under some circumstances. For example, amino acids and glucoses are usually actively transported by specific transporters, including peptide transporters, organic cation transporters, and ABC transporters.[14] Adversely, some efflux proteins, especially P-gp, localized in the apical or basolateral cell membranes have the potential to pump drugs out from the cell into the apical or basolateral extracellular fluids.

Besides molecular descriptors and statistical methods, another important element for developing a reliable prediction model is the high-quality data set. Many of the previous models were generated on the basis of a small number of compounds (20−40), with the exceptions of Wessel et al.,[15] Deretey et al.,[16] Zhao et al.,[4] and Klopman et al.[17] For example, Zhao et al. used a data set of 169 drugs with reliable HIA data, and Klopman et al. utilized an even larger data set of 417 drugs to construct models. Unfortunately, not all of the data sets were released for the public scientific community, and the reliability and validity cannot be guaranteed for models based on the limited data sets. So our first objective is to construct a large database of human intestinal absorption by collecting data from the literature. On the basis of the extended data set, we expect to study the relationships between %FA with well-used molecular properties similarly to our previous work on Caco-2 permeability[18] and then construct reliable prediction models for HIA that can be used as rapid screening filters for candidate drugs.

METHODS AND MATERIALS

Human Intestinal Absorption Data. The data set reported here includes 648 drug and druglike molecules collected from various literature sources. The compound names and the corresponding experimental %FA values are included in the SDF file (the supporting web page: http://modem.ucsd.edu/adme). The data in this database were mainly based on three sources. The first important source is previously reported compilations, such as Palm et al.'s collection,[5] Wessel et al.'s collection,[15] Zhao et al.'s collection,[4] Deretey et al.'s collection,[16] and so forth. The second important source is the reported intestinal absorption data found in references, especially data listed in Therapeutic Drugs.[19] The third important source is based on the bioavailability data (%F). When %F is high, it can be assumed that the bioavailability of the drug can reflect absorption because the effect of first-pass metabolism is minimal and almost all of the absorbed drug can reach the systemic circulation. Here, %FA was defined to the same value of %F when %F is higher than 95%. The bioavailability data were obtained from our previous work.[20] For compounds with %FA values reported as being complete, the value considered was 100%. For compounds with %FA values reported as being poor, the value considered was 5%. When the %FA values were given as a range or when more than one value was reported, an average value was adopted. In general, the deviation observed for the experimental %FA could be as larger as 20%; therefore, the artificial treatment should not have a large influence on the overall reliability of the database and the developed models.[17]

The structures of the compounds were built with the Cerius2 molecular simulation package [21] in their neutral forms, and they were optimized by a molecular mechanism with the MMFF force field.[22] The molecules were then saved in the MACCS SDF and SMILES formats for further analysis. The primary focus of prediction in this work was the modeling of passive drug absorption. Nonetheless, some potential drugs might be subject to other transport mechanisms. Three classes of compounds, as listed in Table S1 in the Supporting Information, were eliminated from the initial collections. These consist of drugs transported by carrier proteins, drugs that show dose-limited absorption and dose-dependent absorption, and drugs that are structurally characterized as non-neural, especially molecules containing an ammonium group, as they bring ambiguity as to what the counteranion is and how the given salt formation may affect absorption. After the eliminating process, the remaining database only includes 553 molecules. We expect that the accuracy of modeling passive diffusion can be guaranteed by using our data set, despite the fact that some of the remaining molecules may also be identified in the future as being actively transported. The database of human intestinal absorption with 648 molecules and that with 553 molecules transported by passive diffusion can be downloaded from the supporting Web site: http://modem.ucsd.edu/adme. It is necessary to emphasize here that the 26 compounds with positively charged nitrogen listed in outliers are not included in correlation while included in classification.

It should be noted that even for the same compound the %FA values reported by different compilations are usually not consistent. Three possible reasons lead to this kind of inconsistency. First, people copied the data of others without verifying the accuracy of the data, and so errors were propagated. For example, the %FA value of sulfasalazine is 12 in Palm et al.'s set,[5] is 65 in Wessel et al.'s set,[15] and 59 in Zhao et al.'s set.[4] According to the reference, the percentage of cumulative drug and its metabolites in urine following oral administration is about 56∼61%.[23] So the %FA of 12 for sulfasalazine is obviously too low, and 65 or 59 seems more reasonable. Second, some compiled %FA data were based on indirect measurements, such as bioavailability, the excretion in urine and feces following oral administration, and the ratio of cumulative urinary excretion of drug-related material following oral and intravenous administration. The reliability of %FA based on these indirect measurements may be questionable in some cases. For example, according to the percentage of cumulative drug and its metabolites in urine following oral/intravenous administration, Zhao et al. defined the %FA of vigabatrin to be about 58.[4] But according to its high bioavailability (100%),[24] a %FA of 58 is obviously low. Third, the intestinal absorption for some compounds varies considerably among different preparations and dosages mainly because of their poor aqueous solubility, crystallinity, or purity. For example, in Wessel et al.'s data set, the %FA of methotrexate is 100,[15] while it is 70 in Zhao et al.'s data set.[4] Finally, cautions

should be taken to guarantee that the duplicates are eliminated from the data set. Many compounds may have several names. For example, phenazone in Palm et al.'s set [5] is as the same as antipyrine in Wessel et al.'s set.[15] So here the canonical SMILES string of each molecule was compared with those of the other molecules iteratively to eliminate all duplicates.

Molecular Descriptors. HIA is mostly a physicochemical process; therefore, physicochemical properties should be used in the prediction of HIA. In the current study, 45 molecular descriptors were used, including topological polar surface area (TPSA), molecular weight (MW), rotatable bond count (N_{rot}), H-bond donor count (N_{HBD}), H-bond acceptor count (N_{HBA}), octanol−water partitioning coefficient (log *P*), apparent partition coefficient (log *D*) at pH = 6.5, intrinsic solubility (log *S*), molecular molar volume, molecular molar refractivity (MR), number of violations of the rule of 5 ($N_{rule-of-5}$), radius of gyration, molecular area (*S*), molecular volume (*V*), principal moment of inertia, 10 shadow indices, six κ indices, 12 Kier and Hall molecular connectivity indices (χ), Wiener index (*W*), and Zagreb index (Zagreb).

TPSA was calculated using the parameters originally proposed by Ertl et al.,[25] which was developed to calculate the polar surface area of a molecule on the basis of its 2D molecular bonding information. Because the 3D structure is not needed to calculate TPSA, it allows van der Waals polar surface area calculations to be implemented in virtual screening approaches. The parameter, log *P*, which defines the hydrophobic feature of a molecule in the uncharged state, was calculated by adding up the well-characterized log *P* contributions of separate atoms, structural fragments, and intramolecular interactions between different fragments defined in ACDLABS 9.0.[26] The apparent coefficient, log *D*, giving a more appropriate description of complex partitioning equilibrium, was estimated on the basis of the predicted log *P* and pK_a calculated by ACDLABS 9.0. The intrinsic solubility log *S* is the solubility for the neutral form of compounds. The parameter, $N_{rule-of-5}$, is defined as the number of violations of the four rule-of-five rules proposed by Lipinski et al.[3] The 10 shadow indices that characterize the shape of the molecules were computed by projecting the molecular surface on three mutually perpendicular planes, XY, YZ, and XZ after the molecules were rotated to align the principal moments of inertia with the X, Y, and Z axes.[27] κ indices were used to quantify attributes of a molecular structure's shape.[28] The Kier and Hall molecular connectivity indices describe different aspects of atom connectivity within a molecule—the amount of branching ring structures, and flexibility, by using four subgraph types: Path, Cluster, Path/Cluster, and Chain.[29,30] The Wiener index defines that the sum of the chemical bonds exists between all pairs of heavy atoms in the molecule.[31] The Zagreb index is defined as the sum of the squares of vertex valencies.[32] TPSA, N_{rot}, N_{HBD}, N_{HBA}, log *P*, log *D* at pH = 7.4, MR, log *S*, and $N_{rule-of-5}$ were calculated using ACDLAB 9.0,[26] and the other descriptors were calculated using the Cerius2 molecular simulation package.[21]

Prediction Models of Intestinal Absorption. The correlation analysis between %FA and several important molecular properties was accomplished by using simple linear fitting. The prediction models of HIA were obtained by using the genetic function approximation (GFA) technique in Cerius2 [21] developed by Rogers and Hopfinger,[33] which combined two seemingly disparate algorithms together: GA[34] and the multivariate adaptive regression splines (MARS) algorithm.[35] The MARS algorithm is a statistical technique for modeling data, which provides an error measure, called the lack of fit (LOF) score, that automatically penalizes models with too many features. Nonlinear modeling can also be achieved by using splines in MARS. In GFA, GA was applied to identify the best prediction models by automatically selecting the most optimal combination of molecular descriptors and functional forms. Compared with other traditional statistical methods, quantitative structure−activity relationship (QSAR) or quantitative structure−property relationship based on GA uses a population of many models and tests only the final, fully constructed models. The details of QSAR analysis based on GA can be found in previous publications.[33,36,37] In this work, an initial population of 100 equations was generated randomly; then, pairs from the population of equations are chosen for "crossover" operations from this set of 100 equations randomly. The number of crossover operations was set to 10 000. The fitness function used to assess the equations is the Friedman's LOF score, which is described by the following equation:

$$\mathrm{LOF} = \mathrm{LSE}/[1 - (c + dp)m]^2 \quad (1)$$

where LSE is the least-squares error, *c* is the number of basis functions in the model, *d* is the smoothing parameter, *p* is the number of descriptors, and *m* is the number of observations in the training set. The smoothing parameter that controls the scoring bias between equations of different sizes was set to the default value of 1.0, and the new term was added with a probability of 50%. Here, the linear equation terms, the quadratic equation terms, and the linear spline equation terms were used for model building. The quadratic equation and the linear spline equation terms were applied to account for the nonlinear effect of some molecular descriptors. The best equation out of the 100 equations was taken according to the statistical parameters' LOF scores. Moreover, the regression coefficient (*r*), the regression coefficient of cross validation (*q*), the standard error of estimate (*s*), and the variance ratio (*F*) were reported. Cross-validated q^2 is defined as $q^2 = (\mathrm{SSY} - \mathrm{PRESS})/\mathrm{SSY}$, where SSY is the sum of squared deviations of the dependent variable values from their mean and PRESS is the sum of the squared prediction error between the actual and the predicted values for the independent variables.

Recursive Partitioning (RP). PR[38] in Cerius2 [21] was used to develop a decision tree to classify the compounds into poor (%FA ≤ 30%) and good (%FA > 30%) intestinal absorption classes. RP is a technique that builds a classification rule to predict the class membership on the basis of feature information. In general terms, RP is a data-analysis method for relating a "dependent" variable (*Y*) to a collection of independent variables (*X*) in order to uncover or simply understand the elusive relationship, $Y = f(X)$. The result of RP is a "decision tree" or "graph", which is constructed through a recursive partitioning process that divides the study sample into smaller and smaller samples (every subsample is called a node) according to whether a particular selected predictor is above a chosen cutoff value or not. At each step of RP, all of the molecular descriptors are sequentially

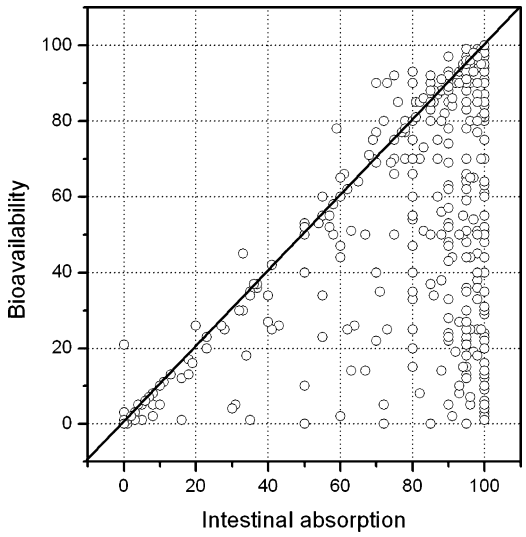

Figure 1. Plot of intestinal absorption vs bioavailability for 470 compounds.

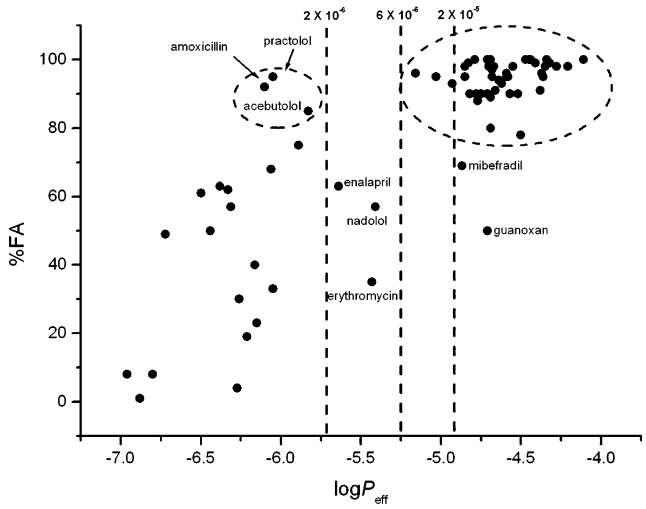

Figure 2. Correlation between Caco-2 permeability and human intestinal absorption.

analyzed to find the best criterion for subdividing compounds into the "good" or "poor" class. Once the best criterion is found, the procedure is repeated for each of the obtained classes of compounds. RP does not try to stop splitting at the right moment; instead, it is designed to "over split" and then prune the tree backwards. In this study, moderate pruning options were set to control the amount of pruning. The minimum number of samples at each node was set to four; the maximum tree depth was set to five, and the number of cross-validation groups was set to 10.

RESULTS AND DISCUSSIONS

1. Relationship between Oral Bioavailability and Intestinal Absorption. The scatter plot of oral bioavailability versus intestinal absorption for 470 common compounds is shown in Figure 1. The %F values for these 470 compounds were obtained from the human bioavailability database reported in our previous work.[20] It is clear that nearly all compounds are distributed in the triangle area below the diagonal. It is not a surprise because the oral bioavailability is a complex function of both absorption and clearance. For a drug to be orally bioavailable, it should reach the general circulation by passing not only through the intestine but also through the liver where it is subject to first-pass metabolism (hepatic clearance). In Figure 1, for those compounds far from the diagonal, they should be metabolized by the liver significantly. It is interesting to give a rough estimation on how many compounds are strongly metabolized through the liver by considering the difference between %F and %FA. Here, if the difference (%FA − %F) is larger than 20%, the metabolized effect was considered to be significant. According to this criterion, 171 compounds (36%) were identified as highly metabolized molecules, while the others were not highly involved in metabolism in the liver. That is to say, the bioavailabilities of most compounds (64%) were mainly controlled by the intestinal absorption process. So the prediction of intestinal absorption is the first step toward the prediction of human oral bioavailability.

2. Relationship between Intestinal Absorption and Caco-2 Permeability. For an accurate and effective prediction of intestinal absorption, several in vitro methods have been developed. Among them, the most popular human-cell-based model for intestinal permeability is the Caco-2 cell system.[39] Caco-2 cells, derived from colorectal carcinoma cells, display many of the morphological and functional properties of the in vivo intestinal epithelial cell barrier. The previous studies have shown that oral drug absorption and the Caco-2 permeability coefficient have a sigmoidal relationship,[40] suggesting that human absorption may be well-predicted by this in vitro model. While, in several studies, this kind of sigmoidal relationship was not very clear.[41,42] Here, the %FA data and the Caco-2 permeability data (log P_{eff}) for a large number of compounds (69 compounds) were collected and compared. The Caco-2 permeability data were obtained from our previous collection.[18] The relationship between log P_{eff} and %FA does not follow a good sigmoidal pattern as shown in Figure 2 because the samples with lowest permeabilities do not form a distinct plateau region. In Figure 2, we could roughly determine an important parameter, the threshold P_c value, allowing the anticipation of high absorption in humans, to be -5.25 ($\approx 6 \times 10^{-6}$ cm/s). For all compounds with a log P_{eff} higher than -5.25, only two compounds (miberadil and guanoxan) are misclassified. It should be noted that the reported P_c values are not completely consistent. For example, Rubas et al. reported the P_c value associated with an apparent permeability of approximately 7×10^{-5} cm/s,[43] Grès et al. reported the P_c value to be approximately 2×10^{-6} cm/s,[41] while Stewart et al. reported the P_c value to be about 2×10^{-5} cm/s.[44] As shown in Figure 2, the performance of the P_c value of 6×10^{-6} cm/s is better than the other reported P_c values. For example, if we used the P_c value proposed by Grès et al., besides miberadil and guanoxan, three other misclassified compounds (enalapril, nadolol, and erythromycin) were included. On the other hand, if we used a larger P_c value, such as the P_c value proposed by Rubas et al. or Stewart et al., some compounds with high absorption were omitted. Why are the P_c values proposed by different groups different? There are two reasons. First, the reported data used in previous analyses are very limited. For example, in Stewart's work, the compounds used for developing the rule were six. In Rubas's work, the data set only included seven compounds. Second, the reported Caco-2 permeability values have large interlaboratory differences,

Table 1. Multivariate Prediction Models for Intestinal Absorption

(1) $\%FA = 109.12 - 0.34TPSA$
$n = 553$, $r = 0.70$, SD = 20.03, $F = 544.51$
(2) $\%FA = 68.54 + 6.07 \log P$
$n = 553$, $r = 0.48$, SD = 24.75, $F = 165.12$
(3) $\%FA = 78.17 + 6.32 \log D_{6.5}$
$n = 553$, $r = 0.63$, SD = 22.03, $F = 354.21$
(4) $\%FA = 89.40 - 25.11N_{rule-of-5}$
$n = 553$, $r = 0.61$, SD = 22.36, $F = 327.08$
(5) $\%FA = 70.64 + 0.14n_{HBD}^2 + 11.74<2 - N_{rule-of-5}> - 9.59<0.05 - \log D_{6.5}> - 0.23<TPSA - 71.00> + 0.30 \log D_{6.5}^2$
$n = 455$, LOF = 284.69, $r = 0.83$, SD = 15.50, $F = 192.58$
(6) $\%FA = 103.87 + 0.35n_{HBD}^2 - 3.98n_{HBD} - 7.78<0.09 - \log D_{6.5}> - 0.26<TPSA - 71.43> - 0.02shadow - X^2$
$n = 455$, LOF = 285.65, $r = 0.82$, SD = 15.67, $F = 190.36$
(7) $\%FA = 93.87 + 0.16n_{HBD}^2 - 9.70<0.05 - \log D_{6.5}> - 0.23<TPSA - 71.43> - 10.73N_{rule-of-5}$
$n = 455$, LOF = 286.19, $r = 0.82$, SD = 15.71, $F = 189.83$
(8) $\%FA = 97.12 - 11.48N_{rule-of-5} - 8.99<0.05 - \log D_{6.5}> - 0.15<TPSA - 49.41> + 0.17 \log D_{6.5}^2 + 3.76<n_{HBD} - 7>$
$n = 435$, LOF = 171.81, $r = 0.87$, SD = 12.70, $F = 277.59$
(9) $\%FA = 98.00 - 9.80N_{rule-of-5} - 8.02<0.08 - \log D_{6.5}> - 0.17<TPSA - 49.69> + 4.27<n_{HBD} - 7>$
$n = 435$, LOF = 173.22, $r = 0.87$, SD = 12.68, $F = 339.84$
(10) $\%FA = 96.99 - 11.86N_{rule-of-5} - 9.19<0.05 - \log D_{6.5}> - 0.15<TPSA - 49.41> + 2.89<n_{HBD} - 5> + 0.19 \log D_{6.5}^2$
$n = 435$, LOF = 173.68, $r = 0.87$, SD = 12.77, $F = 273.69$
(11) $\%FA = 92.84 - 11.28N_{rule-of-5} - 9.25<0.05 - \log D_{6.5}> - 0.26<TPSA - 49.33> + 3.02<n_{HBD} - 5> + 0.19 \log D_{6.5}^2 + 0.10TPSA$
$n = 435$, LOF = 174.58, $r = 0.87$, SD = 12.64, $F = 228.82$

which may lead to the inconsistent P_c values proposed by different works.

In Figure 2, three compounds with %FA values larger than 80% show relatively poor permeability across Caco-2 monolayers. For amoxicillin, it is absorbed by carrier-mediated transport, and the discrepancy can be partially explained either by the saturation of the carrier or, more likely, by the fact that Caco-2 cells displayed a variable and generally lower expression of carrier-mediated transport than that seen in vivo.[41] Although as shown in Figure 2, the intestinal absorption has a good linear correlation with Caco-2 permeability ($r = 0.82$ and SD = 16.2), it should be pointed out that the good correlation was primarily caused by the high density of compounds with high intestinal absorption. When these molecules (44 compounds) in the larger circle in Figure 2 were eliminated from the data set, the correlation between %FA and log P_{eff} of the other 25 molecules was only $r = 0.41$. For these compounds with low or medium intestinal absorption, the predictions of oral absorption based on Caco-2 permeability are not very reliable. For compounds with high intestinal absorption, permeability of Caco-2 monolayers can be used as a predictive tool to estimate oral absorption, while for compounds with low or medium absorption, the permeability of Caco-2 monolayers may not give a very good rank for estimating oral absorption. So even if we have experimental Caco-2 permeability data or have good prediction models for Caco-2 permeabilities, the development of prediction models of HIA is also demanding.

3. Correlation between Important Molecular Properties and Intestinal Absorption. The correlation analysis was conducted between each molecular descriptor and intestinal absorption. In all molecular descriptors, several of them have a high correlation with intestinal absorption ($|r| \geq 0.6$), including TPSA ($r = -0.70$), N_{HBD} ($r = -0.68$), log $D_{6.5}$ ($r = 0.63$), N_{HBA} ($r = -0.63$), and $N_{rule-of-5}$ ($r = -0.61$). The contributions of these descriptors are more important than those of the other descriptors.

3.a. Topological Polar Surface Area (TPSA). In 1992, van de Waterbeemd and Kansy first correlated the PSA of a series of central nervous system drugs to log BB.[45] Thenceforward, PSA has become the most popular parameter for the prediction of molecular transport properties. Here, the correlation ($r = -0.70$) is better than those of the fittings between %FA and other important molecular descriptors (eq 2 in Table 1 and Figure 3a). Clark even found that an excellent sigmoidal relationship could be established between

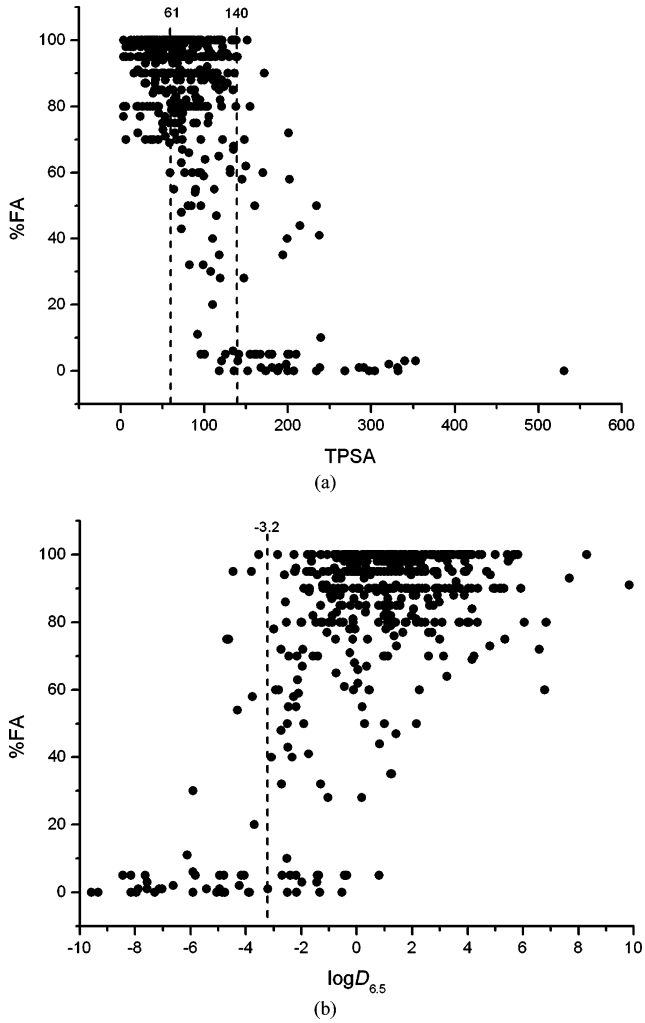

Figure 3. Correlation between (a) TPSA and (b) log $D_{6.5}$ with intestinal absorption.

%FA and PSA ($r^2 = 0.94$) for a set of 20 drugs covering a wide range of %FA values in humans and claimed that drugs that are completely absorbed (FA > 90%) had a $PSA_d \leq 61$ Å^2 while drugs that are less than 10% absorbed had a $PSA_d \geq 140$ Å.[46] It is interesting to validate the performance of the rules proposed by Clark on our data set. In the data set, there are 48 molecules with a FA% equal to or smaller than 10% (the compounds with positively charged nitrogen are not included). In these 48 molecules, seven of them have a polar surface area smaller than 140 Å^2. Furthermore, 14 molecules are apparent false positives using the rule of TPSA larger than 140 Å^2. Because all possible drugs transported by carriers were eliminated in our data set, it is likely that many compounds are real false positives. When we applied the value of 61 Å^2 to this set, we picked out 230 compounds as possibly being well-absorbed. In these 230 compounds, 47 have an intestinal absorption smaller than 90% and 17 smaller than 80%. For the 266 compounds with a TPSA larger than 61 and smaller 140 Å^2, 165 compounds have an intestinal absorption larger than 90% and five compounds smaller than 10%. It is clear that the performance of the TPSA criterion is not very reliable to identify poor absorption or good absorption, and HIA is certainly not only determined by polar surface area.

PSA or TPSA is usually considered as a parameter to define the hydrogen-bonding potential because it is closely correlated with the number of hydrogen-bond donors or acceptors. For example, TPSA is highly correlated with the number of hydrogen-bond acceptors ($r = 0.93$) and the total number of hydrogen-bond donors ($r = 0.82$). Put simply, TPSA can account for the possible negative electrostatic or hydrogen-bonding contribution of the polar atoms. But PSA or TPSA can only account for the hydrogen-bonding or electrostatic contribution of these atoms on the molecular surface. In some cases, the highly charged atoms located in the interior of the molecule may have great impact on the interactions between the transported molecule and membrane. In our data set, there are 26 compounds with at least one charged nitrogen. For these 26 compounds, most of them have very low TPSA values, but all of them are poorly absorbed. The reason may be that the electrostatic contribution of the high-charged atoms cannot be effectively accounted for by the polar surface area because the positive-charged nitrogen atoms are usually shielded by the connected atoms. Another possible explanation is that the charged molecule might diffuse across the membrane boundary as ion pairs. It is likely that ion-pair formation with cations derived from typical drug bases will be much more favored than ion-pair formation with tetraalkyammonium cations. So the isolated positive charge may have a great negative effect on the diffusion of drugs.[47]

3.b. Hydrophobicity. In our previous work,[18] it has been proven that the distribution coefficient, log *D*, was a very important indicator of Caco-2 permeability. Actually, the hydrophobic parameters (log *P* or log *D*) have long been known to be important for membrane permeation. First, a direct fitting of %FA with the partition coefficient (log *P*) was conducted, which produced an *r* of approximately 0.48 (eq 2 in Table 1). Then, we correlated %FA with the predicted log *D* values at pH = 2, 5.5, 6.5, 7.4, and 10, and the correlation coefficients are 0.48, 0.60, 0.63, 0.62, and 0.54, respectively. The best correlation was obtained at pH = 6.5 (eq 3 in Table 1). Compared with eq 2, the correlation coefficient and the variance ratio of eq 3 were improved greatly. The plot of correlation of log $D_{6.5}$ versus %FA is shown in Figure 3b, indicating that hydrophobic molecules with high log $D_{6.5}$ values are favorable to diffuse across the biological membrane, and the hydrophilic molecules with low log $D_{6.5}$ values usually have a low percentage of intestinal absorption. In most works, researchers like to use log *P* instead of log *D* because log *P* is easier to compute. But indubitably, log *D* is more effective in the prediction of membrane permeability than log *P*. In Figure 3b, log $D_{6.5} = -3.2$ may be identified as a rough bound to identify the compounds with a %FA smaller than 10% from the others.

We should emphasize that in our analysis the predicted log *D* values were used. In our previous work,[18] we correlated Caco-2 permeabilities with experimental and predicted log *D* values and found that the experimental log *D* performed better than the predicted values. Here, we performed a correlation between the experimental log *D* values and the predicted log *D* values at pH = 7.4 for 68 compounds collected in our previous work[18] (Figure S1 in the Supporting Information). For these 68 compounds, 48 compounds show prediction errors smaller than 1.0 log unit, and seven compounds show prediction errors larger than 2.0 log units. So, although the log *D* values of most compounds can be satisfactorily predicted by ACDLABS, the available prediction methods for log *D* still have great room for improvement both on log *P* prediction and on pK_a prediction.

3.c. The Parameter Related to Rule of Five. Here, the parameter, $N_{\text{rule-of-5}}$, defined the number of violations of the rule of five proposed by Lipinski et al.[3] The rule of five was widely applied to identify compounds with possible poor absorption and permeability. According to the correlation coefficient ($r = -0.61$), the prediction capability of $N_{\text{rule-of-5}}$ is not satisfactory. Certainly, the rule of five is not used for the accurate prediction of intestinal absorption but, rather, for a rough classification of compounds. We can give an estimation of the performance of $N_{\text{rule-of-5}}$ on the classification of intestinal absorption. Here, this compounds was considered to be poorly absorbed if $N_{\text{rule-of-5}}$ is equal to or larger than 2; otherwise, it was considered to be moderately or highly absorbed. For all 48 molecules with a FA% equal to or smaller than 10%, 28 compounds had an $N_{\text{rule-of-5}} \geq 2$. That is to say, 20 poorly absorbed compounds were misclassified. Moreover, 12 compounds among the other 505 compounds with a %FA larger than 10% were misclassified. Obviously, the rule of five is not a good predictor to estimate HIA. Compared with the performance of TPSA, the criterion of ≥ 2 is less reliable for identifying poorly absorbed molecules from the others.

4. Prediction Models Proposed by GFA. According to the above discussion, we know that HIA is controlled by many molecular properties rather than by a single one. The prediction based on these important molecular properties should give more reliable output than that based on a single molecular property. To automatically select the most crucial descriptors determining HIA, the GFA technique was applied here to search the combination space of molecular properties. We used 455 compounds to create the prediction models, and the remaining 98, randomly selected from the entire database, were used as an external test set. In GFA calculations, besides the linear equation terms, the quadratic

equation terms and the linear spline equation terms were also introduced. The splines used here were denoted with angled brackets. For example, $<f(x) - a>$ was equal to zero if the value of $f(x) - a$ was negative; otherwise, it was equal to $f(x) - a$. The regression with splines allows the incorporation of features that do not have a linear effect over their entire range.

After the GFA calculations, the 100 best models were obtained, and the top 10 were picked out for further analysis. Actually, the top 10 best-scored equations shared very similar information in terms of statistical parameters and types of descriptors. Some equations can be classified into one group, and the final representative two equations are shown in Table 1 (eqs 5, 6, and 7). Using eqs 5−7, the absolute mean errors are only 11.6%, 11.8%, and 11.8%, respectively, but there are still some compounds with large prediction errors. If we used eq 5 for prediction, there are 21 compounds with a prediction error larger than 35%, which include meropenem, pentamidine, streptozocin, nedocromil, imipenem, phthalylsulfathiazole, succinylsulfathiazole, sulbactam, amygdalin, chlorhexidine, diatrizoate, mitoxantrone, moexipril diacid, netivudine, nadolol, trandolapril, ceftizoxime, telithromycin, acipimox, cyclopenthiazide, and ergotamine. Among these 21 compounds, 16 compounds with experimental %FA values smaller than 50% were highly overestimated, while the other five compounds were underestimated. Now, we cannot find solid evidence to explain why 16 compounds were highly overestimated. The first possible reason is that some highly charged groups could not be precisely described by polar surface area. For example, succinylsulfathiazole, phthalylsulfathiazole, and sulbactam have the −S═O(═O) group. The influence of this highly charged group may not be fully described by polar surface area. Another important reason is that some compounds may be the strong substrates of g-pg, and their diffusions are greatly effected by the efflux effect of p-gp. A very important feature of the prediction of the training set using eq 5 is that only one compound (ceftizoxime) moderately or highly absorbed was predicted to be poorly absorbed. This feature is very appealing because, from a practical point of view, we were most concerned about the false prediction of compounds highly absorbed. When the compounds are discarded on the basis of the prediction, there is a slim chance that those compounds were tested by experiments.

If these 21 possible outliers were not included in the training set, the correlations of the model could be greatly improved. The top four independent prediction models are listed in Table 1 (eqs 8−11). In eqs 8−11, it is interesting to find that the spline models were applied for three important descriptors: TPSA, n_{HBD}, and log $D_{6.5}$, indicating that these three descriptors are not linearly correlated with %FA in the whole property space. In fact, the relationships between molecular descriptors and %FA may not always be well-described by linear correlation. For example, Palm et al. have reported a good sigmoidal relationship between the intestinal absorption and PSA. But according to Figure 3a, it is obvious that the relationship between %FA and TPSA cannot be simply described by a linear regression, while it also cannot be described effectively by a sigmoidal fitting because a sigmoidal curve should possess two plateau regions at the low and high values of the variable. In eqs 8−11, the threshold value of TPSA is about 50 Å, demonstrating that higher TPSA values produce low permeation while the effect takes effect only when the polar surface area is larger than 50 Å^2. A spline model for log $D_{6.5}$ is also included in the prediction models. A threshold of 0.05 was found for log $D_{6.5}$, which means that lower log $D_{6.5}$ values produce low permeation when it is smaller than 0.05. The interpretation of the n_{HBD} term is not very straightforward. This term indicates that n_{HDB} is unfavorable for HIA when it is larger than 5 or 7. This term may be used for the neutralization of the strong effect of TPSA and $N_{rule-of-5}$.

Validation is crucial in any QSAR modeling. The calculated q (0.87) shows that eq 8 is reliable. Certainly, the high value of q appears to be the necessary but not the sufficient condition for the models to have a good predictive power. Golbraikh and Tropsha even emphasized that the actual predictive ability of a QSAR model can only be estimated using an external test set of compounds that were not used for building the model.[48] The selection of the definitive model was carried out on the basis of prediction for the compounds comprising the validation test. Here, the actual prediction powers of eqs 5−11 were validated by an external test set of 98 compounds. Actually, these seven models do not show large differences on the prediction capability of the test set, and the absolute mean error for eqs 5−11 are 7.78%, 8.52%, 7.86%, 7.33%, 7.32%, 7.42%, and 7.47%, respectively. It is thus difficult for us to give a decisive conclusion as to which model is the best on the basis of the fitness score and the predictions for the limited test compounds. According to the prediction on the external test set, eqs 8 and 9 may be a little better than the other models. It is interesting that the prediction models with and without the 21 possible outliers do not affect the actual prediction of the test set obviously.

Usually, that selection of a single model and the discarding of the remaining models may not be the most advantageous choice, and the average based on the outputs of the multiple models is more reliable.[49] The absolute mean error of 7.32% of the average prediction based on the outputs of eqs 8−11 is quite similar to that of eq 9, and it is better than those of the other equations. So using multiple models may not be as risky as using a single prediction model. According to the actual prediction on the external test set, eq 9 was the best model. The plot of the experimental %FA data versus the predicted values is shown in Figure 4. The observed %FA values, calculated and residuals, are shown in Table 2. For this test set, the prediction model gives very effective prediction [$r = 0.89$, SD $= 10.28$, AME (absolute mean error) $= 7.32\%$], which is even better than predictions of the training set. In the test set of 98 compounds, 73 have the prediction error smaller than 10%, 19 have the prediction error between 10% and 20%, two have it between 20 and 25%, and four have it larger than 25%. These four compounds with the worst predictions are tiludronic acid, metolazone, cinoxacin, and betahistine. In these four compounds, tiludronic acid and metolazone were highly overestimated. The overestimation may be caused by the intrinsic limitation of polar surface area. Tiludronic acid contains two $-PO_3$ groups, and metolazone contains a $-SO_2(NH_2)$ group. In the definition of polar surface area, sulfur and phosphor atoms are not considered. While in these two groups, P and S atoms are highly positive charged. So the polar surface area may not properly describe these highly charged groups. Actually, in the training set, some compounds containing

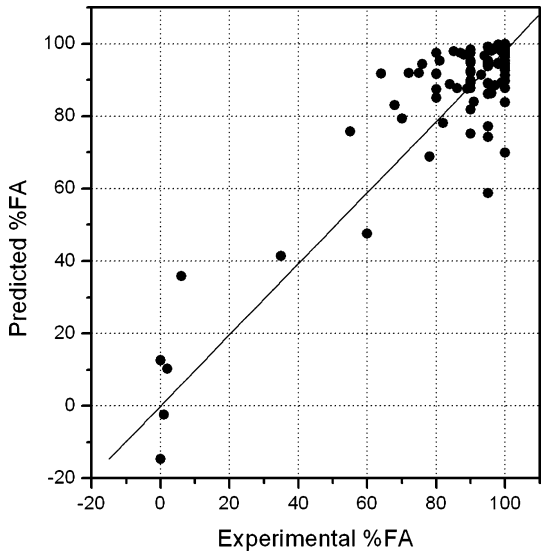

Figure 4. Correlation between experimental %FA and the predicted values for 98 compounds in the test set using eq 9 in Table 1.

the $-SO_2$ group were also overestimated, such as succinylsulfathiazole, phthalylsulfathiazole, and sulbactam.

5. Classification Based on Recursive Partitioning. In the above section, a group of regression models with good predictive power was developed. In the practical process of drug discovery, the accurate prediction for %FA is not always necessary, and we only want to classify the compounds into good or poor absorption. The classification model has some advantages over the linear correlation models. First, the nonlinear effects that cannot be effectively considered by the regression models are implicitly accounted for in classification. Second, the statistical classification methods can discriminate different biological mechanisms that regression models cannot. Third, the precise experimental values are not usually necessary for classification, because classification deals with binary data, accepting any variability of %FA above 30%. Here, RP was applied for classification, which can find decision trees to classify molecules into different categories. Compared with "the blind operations" of ANNs and SVMs, the results of RP can be easily converted to simple hierarchical rules, which are clearly interpreted. According to the criteria used by Kansy et al.,[50] in the training set of 481 compounds, 74 compounds that have low %FA values of less than 30% were grouped into class 1, and 407 compounds with a moderate or high percentage of intestinal absorption of more than 30% were group into class 2. It should be noted that the 26 compounds with at least one positively charged nitrogen atom were also included in the training set. The test set of 98 compounds was used to test the actual performance of the obtained classification model. Then, a RP analysis was conducted to see whether these models correctly predict the compounds in their respective groups. All 45 descriptors were applied in RP analysis.

It is encouraging to find that the obtained model has very good classification performance on the training set, and it can correctly identify 95.9% (71/74) of the compounds in class 1 and 96.1% (391/407) of the compounds in class 2. For the 74 compounds with a %FA equal to or smaller than 30%, only three compounds were misclassified, including sulbactam, moexipril diacid, and netivudine. In fact, these three compounds were also highly overestimated by the regression models. In the above analysis, we already knew that, for 48 compounds with an intestinal absorption smaller than or equal to 10%, 13 of them have large a prediction error. Among these 13 compounds, 11 of them can be correctly classified by RP. The possible reason of the better performance of the classification model is that the nonlinear effects of some molecular descriptors may not be well-explained by the regression models. Furthermore, the actual prediction of the classification model on the test set was verified. The test set included five compounds in class 1 and 93 compounds in class 2. The performance on the test set is also very satisfactory. All five compounds in class 1 were correctly classified, and only three compounds in class 2, which are erythromycin, reproterol, and telmisartan, were not correctly identified. Among these three compounds, the %FA value for erythromycin is not high, which is 35%. So it is understandable that this compound is easily misclassified. Although a single property cannot be used as an effective rule for classification, the proposed hierarchical rules combining several properties together is very effective.

The decision tree is shown in Figure 5, which can be easily

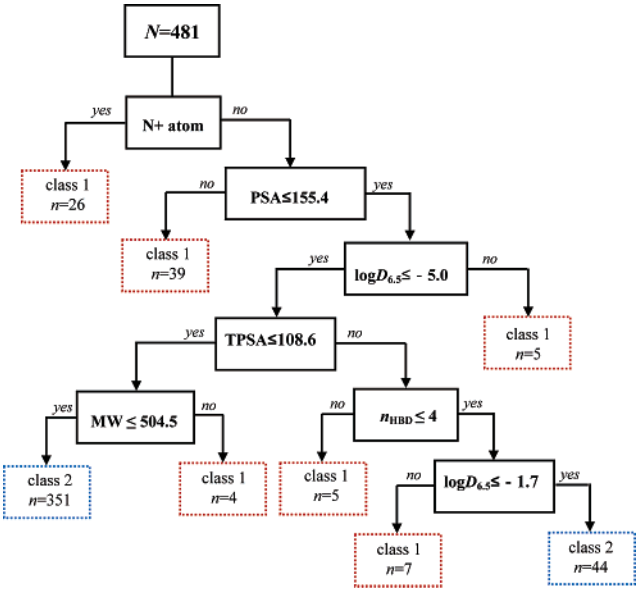

Figure 5. Decision tree to classify compounds into good and poor absorption by using recursive partitioning.

converted to a group of hierarchical rules. Besides the property of N+ atom (the positively charged nitrogen atom), the several other molecular parameters used for the hierarchical rule include log $D_{6.5}$, TPSA, N_{HBD}, and MW. Actually, log $D_{6.5}$, TPSA, and N_{HBD} are also used in the above correlation models. From the hierarchical level in Figure 4, the properties of TPSA, log $D_{6.5}$, and N_{HBD} are more important than MW. So HIA is a complicated function of several properties including hydrophobicity, hydrogen-bonding potential, and molecular weight, and the reliable rules can only be obtained by considering all of these important properties. Here, the decision tree is only used for the prediction of HIA related to passive diffusion. We certainly can extend the decision tree by integrating other rules for predicting active transport or first pass metabolism. For example, we can define "substructure-specific rules" related to specific metabolism processes.

Table 2. Calculated Molecular Properties and the Experimental and Calculated %FA Values for Compounds in the Test Set

number	name	N_{HBD}	log $D_{6.5}$	TPSA	$N_{rule-of-5}$	$\%FA_{exp}$	$\%FA_{pred}$	$class_{exp}$	$class_{pred}$
1	amikacin	17	−9.3	331.9	3	0	−12.9	1	1
2	moxalactam	4	−3.9	234.4	2	0	14.9	1	1
3	kanamycin	14	−7.0	285.7	3	1	0.7	1	1
4	zanamivir	9	−6.6	198.2	2	2	7.5	1	1
5	tiludronic acid	4	−5.9	134.7	0	6	35.3	1	1
6	erythromycin	5	1.3	193.9	3	35	43.7	2	1
7	tranexamic acid	3	−2.2	63.3	0	55	77.5	2	2
8	reproterol	4	−2.9	131.2	1	60	50.0	2	1
9	metolazone	3	3.3	100.9	0	64	89.2	2	2
10	hydrochlorothiazide	4	−0.1	135.1	0	68	82.1	2	2
11	naratriptan	2	−1.6	73.6	0	70	80.6	2	2
12	desogestrel	1	6.6	20.2	1	72	88.2	2	2
13	estramustine	1	5.3	49.8	1	75	88.2	2	2
14	propylthiouracil	2	1.4	73.2	0	76	93.9	2	2
15	ethambutol	4	−3.0	64.5	0	78	70.8	2	2
16	cyproheptadine	0	4.0	3.2	1	80	88.2	2	2
17	flunisolide	2	2.2	93.1	0	80	90.5	2	2
18	losartan	2	1.8	92.5	0	80	90.6	2	2
19	metyrapone	0	1.2	42.9	0	80	98.0	2	2
20	pizotyline	0	3.7	31.5	1	80	88.2	2	2
21	piroximone	2	2.1	71.1	0	81	94.3	2	2
22	sorivudine	4	−0.9	119.3	0	82	78.2	2	2
23	propiverine	0	4.2	38.8	1	84	88.2	2	2
24	fenoprofen	1	1.6	46.5	0	85	98.0	2	2
25	topiramate	2	3.0	123.9	0	86	85.2	2	2
26	clobazam	0	1.6	40.6	0	87	98.0	2	2
27	moclobemide	1	0.3	41.6	0	88	98.0	2	2
28	chloramphenicol	3	1.0	115.4	0	89	86.7	2	2
29	alprazolam	0	2.5	43.1	0	90	98.0	2	2
30	bicalutamide	2	4.9	115.6	0	90	86.6	2	2
31	diazoxide	1	1.1	66.9	0	90	95.0	2	2
32	ethionamide	2	1.2	71.0	0	90	94.3	2	2
33	hydroxychloroquine	2	0	48.4	0	90	97.6	2	2
34	levosimendan	2	0.1	113.4	0	90	87.0	2	2
35	mestranol	1	5.2	29.5	1	90	88.2	2	2
36	nifedipine	1	2.3	110.5	0	90	87.5	2	2
37	pindolol	3	−0.7	57.3	0	90	90.7	2	2
38	rizatriptan	1	−1.7	49.7	0	90	83.5	2	2
39	telmisartan	1	4.1	70.7	2	90	74.8	2	1
40	tolbutamide	2	0.6	83.7	0	90	92.1	2	2
41	saccharin	1	−1.1	71.6	0	91	84.9	2	2
42	codeine	1	−0.6	41.9	0	93	92.8	2	2
43	dienogest	1	2.0	61.1	0	94	96.0	2	2
44	acitretin	1	3.9	46.5	1	95	88.2	2	2
45	bifemelane	1	1.1	21.3	0	95	98.0	2	2
46	cinoxacin	1	−3.8	88.4	0	95	60.2	2	2
47	delmopinol	1	4.0	32.7	0	95	98.0	2	2
48	fenfluramine	1	0.1	12.0	0	95	97.8	2	2
49	gliquidone	2	1.6	130.3	1	95	74.3	2	2
50	labetalol	5	−0.2	95.6	1	95	78.4	2	2
51	naltrexone	2	0.6	70.0	0	95	94.5	2	2
52	oxprenolol	2	−0.3	50.7	0	95	94.7	2	2
53	phenprocoumon	1	2.8	46.5	0	95	98.0	2	2
54	propoxyphene	0	2.9	29.5	1	95	88.2	2	2
55	sulfamethazine	3	0.8	106.3	0	95	88.2	2	2
56	tramadol	1	−0.2	32.7	0	95	95.7	2	2
57	capecitabine	3	0.0	120.7	0	96	85.4	2	2
58	praziquantel	0	2.4	40.6	0	96	98.0	2	2
59	diclofenac	2	1.9	49.3	0	97	98.0	2	2
60	trimethoprim	4	0.0	105.5	0	97	87.9	2	2
61	desipramine	1	1.1	15.3	0	98	98.0	2	2
62	imipramine	0	2.1	6.5	0	98	98.0	2	2
63	maprotiline	1	1.5	12.0	0	98	98.0	2	2
64	phenylbutazone	0	3.9	40.6	0	98	98.0	2	2
65	ximoprofen	2	0.2	69.9	0	98	94.5	2	2
66	desmethyldiazepam	1	3.1	41.5	0	99	98.0	2	2
67	naproxen	1	1.3	46.5	0	99	98.0	2	2
68	tolmetin	1	−0.7	59.3	0	99	90.0	2	2

Table 2. Continued

number	name	N_{HBD}	log $D_{6.5}$	TPSA	$N_{rule-of-5}$	$\%FA_{exp}$	$\%FA_{pred}$	$class_{exp}$	$class_{pred}$
69	aminoglutethimide	3	1.4	72.2	0	100	94.1	2	2
70	azelastine	1	1.0	35.6	0	100	98.0	2	2
71	betahistine	1	−2.9	24.9	0	100	74.5	2	2
72	buspirone	0	3.1	69.6	0	100	94.6	2	2
73	chlorambucil	1	1.4	40.5	0	100	98.0	2	2
74	cinchonine	1	0.7	36.4	0	100	98.0	2	2
75	dextromoramide	0	3.1	32.8	0	100	98.0	2	2
76	doxepin	0	0.3	12.5	0	100	98.0	2	2
77	etoricoxib	0	−0.3	80.9	0	100	89.7	2	2
78	flurazepam	0	1.1	35.9	0	100	98.0	2	2
79	gestodene	1	3.7	37.3	0	100	98.0	2	2
80	guanfacine	4	1.6	81.5	0	100	92.5	2	2
81	indoprofen	1	0.7	57.6	0	100	96.6	2	2
82	ketazolam	0	3.4	49.9	0	100	98.0	2	2
83	linezolid	1	0.4	71.1	0	100	94.3	2	2
84	mebendazole	2	2.8	84.1	0	100	92.1	2	2
85	methocarbamol	3	0.6	91.0	0	100	90.9	2	2
86	nafronyl	0	2.4	38.8	0	100	98.0	2	2
87	nilutamide	1	3.3	95.2	0	100	90.1	2	2
88	norgestrel	1	3.7	37.3	0	100	98.0	2	2
89	oxatomide	1	3.0	38.8	0	100	98.0	2	2
90	penbutolol	2	1.6	41.5	0	100	98.0	2	2
91	phenobarbital	2	1.6	75.3	0	100	93.6	2	2
92	procyclidine	1	0.9	23.5	0	100	98.0	2	2
93	quinagolide	2	0.8	81.3	0	100	92.5	2	2
94	stavudine	3	−0.8	88.2	0	100	84.3	2	2
95	tamsulosin	3	0.0	108.3	0	100	87.0	2	2
96	tetrabenazine	0	3.2	38.8	0	100	98.0	2	2
97	trazodone	0	1.4	42.4	0	100	98.0	2	2
98	zaleplon	0	0.9	74.3	0	100	93.7	2	2

CONCLUSION

In the current work, a large carefully validated database of intestinal absorption was reported. On the basis of the large set of drug or druglike molecules, the correlation and classification models were proposed by genetic function approximation and recursive partitioning techniques, respectively. The high qualities of those models were validated by the satisfactory predictions on the training and test sets. Overall, our regression and classification models have good performance on the training and test sets; however, there is still room to further improve the models. First of all, the quality and quantity of our data set should be improved further. In the current work, our data analysis and models were purely based on the passive diffusion mechanism. Although most of the compounds diffused by carrier-mediated transport were excluded, we are not sure all of the remaining compounds for model construction are transported by the passive diffusion mechanism. We expect that our database can be further improved by checking more references and by introducing more experimental data with high quality. Another important problem of the current data set is that the %FA values are not "balanced", because the data set is heavily skewed toward well-absorbed compounds. In our data set, only 79 compounds have %FA values equal to or less than 30% (including 26 compounds with at least one charged nitrogen). The skewed distribution is also found for the other data sets. For example, Wessel et al.'s data set only includes 10 compounds with %FA values equal to or less than 30%,[15] and the Zhao et al.'s data set only includes 15 compounds having %FA values equal to or less than 30%.[4] This bias may tend to cause the prediction models to be less accurate in predicting %FA values for poorly absorbed compounds. Second, considering that drug absorption is a very complicated process arising from multiple physiological processes, approaches combining models considering both passive diffusion and active transport should be considered, especially when more high-quality data become available in the future. In addition, we should introduce more structure-based rules related to first-pass metabolism. Yoshida and Topliss reported a classification model for bioavailability.[51] This classification model includes three molecular descriptors and 15 other structural descriptors relating primarily to well-known metabolic processes. We expect that more meaningful structural descriptors on first-pass metabolism can be identified through the analysis of our bioavailability database.[20] The newly introduced descriptors can improve not only the prediction models but also the decision tree for classification.

ACKNOWLEDGMENT

T.H. is supported by a CTBP postdoctoral scholarship. We thank Prof. J. Andrew McCammon for providing access to the Cerius2 molecular simulation package.

Supporting Information Available: The scatter plot of experimental log *D* versus predicted values for 68 compounds is shown in Figure S1. The compounds excluded from the model development are listed in Table S1. This material is available free of charge via the Internet at http://pubs.acs.org. The human intestinal absorption database of 648 compounds, the human intestinal absorption data set of 455 compounds in the training set, the human intestinal absorption data set of 98 compounds in the test set, the combined database of 470 compounds with both intestinal absorption data and oral bioavailability data, and the combined database of 69 compounds with both intestinal absorption data and Caco-2 permeability data can be downloaded from the supporting Web site: http://modem.ucsd.edu/adme.

REFERENCES AND NOTES

(1) Hou, T. J.; Xu, X. J. Recent Development and Application of Virtual Screening in Drug Discovery: An Overview. *Curr. Pharm. Des.* **2004**, *10* (9), 1011−1033.

(2) Kennedy, T. Managing the Drug Discovery/Development Interface. *Drug Discovery Today* **1997**, *2* (10), 436−444.

(3) 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* (1−3), 3−25.

(4) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of Human Intestinal Absorption Data and Subsequent Derivation of a Quantitative Structure−Activity Relationship (QSAR) with the Abraham Descriptors. *J. Pharm. Sci.* **2001**, *90* (6), 749−784.

(5) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar Molecular Surface Properties Predict the Intestinal Absorption of Drugs in Humans. *Pharm. Res.* **1997**, *14* (5), 568−571.

(6) Osterberg, T.; Norinder, U. Prediction of Polar Surface Area and Drug Transport Processes Using Simple Parameters and PLS Statistics. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1408−1411.

(7) Perez, P. A. C.; Sanz, M. B.; Torres, L. R.; Avalos, R. C.; Gonzalez, M. P.; Diaz, H. G. A Topological Sub-Structural Approach for Predicting Human Intestinal Absorption of Drugs. *Eur. J. Med. Chem.* **2004**, *39* (11), 905−916.

(8) Deconinck, E.; Hancock, T.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. Classification of Drugs in Absorption Classes Using the Classification and Regression Trees (CART) Methodology. *J. Pharm. Biomed. Anal.* **2005**, *39* (1−2), 91−103.

(9) Agatonovic-Kustrin, S.; Beresford, R.; Yusof, A. P. M. Theoretically Derived Molecular Descriptors Important in Human Intestinal Absorption. *J. Pharm. Biomed. Anal.* **2001**, *25* (2), 227−237.

(10) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1630−1638.

(11) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2* (3), 192−204.

(12) Stenberg, P.; Bergstrom, C. A. S.; Luthman, K.; Artursson, P. Theoretical Predictions of Drug Absorption in Drug Discovery and Development. *Clin. Pharmacokinet.* **2002**, *41* (11), 877−899.

(13) Hou, T. J.; Wang, J. M.; Zhang, W.; Wang, W.; Xu, X. J. Recent Advances in Computational Prediction of Drug Absorption and Permeability in Drug Discovery. *Curr. Med. Chem.* **2006**, *13* (22), 2653−2667.

(14) Sai, Y.; Tsuji, A. Transporter-Mediated Drug Delivery: Recent Progress and Experimental Approaches. *Drug Discovery Today* **2004**, *9* (16), 712−720.

(15) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of Human Intestinal Absorption of Drug Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (4), 726−735.

(16) Deretey, E.; Feher, M.; Schmidt, J. M. Rapid Prediction of Human Intestinal Absorption. *Quant. Struct.-Act. Relat.* **2002**, *21* (5), 493−506.

(17) Klopman, G.; Stefan, L. R.; Saiakhov, R. D. ADME Evaluation 2. A Computer Model for the Prediction of Intestinal Absorption in Humans. *Eur. J. Pharmacol. Sci.* **2002**, *17* (4−5), 253−263.

(18) Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1585−1600.

(19) Dollery, C. T. *Therapeutic Drugs*, 2nd ed.; Churchill Livingstone: Edinburgh, U. K., 1999; p 2.

(20) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME Evaluation in Drug Discovery. 6. If the Oral Bioavailability in Human Can be Effectively Predicted by Simple Molecular Properties? *J. Chem. Inf. Comput. Sci.*, in revision.

(21) *Cerius2*, version 4.10. http://www.accelrys.com (accessed Oct 2006).

(22) Halgren, T. A. Merck Molecular Force field.1. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490−519.

(23) Pieniaszek, H. J.; Resetarits, D. E.; Wilferth, W. W.; Blumenthal, H. P.; Bates, T. R. Relative Systemic Availability of Sulfapyridine from Commercial Enteric-Coated and Uncoated Sulfasalazine Tablets. *J. Clin. Pharmacol.* **1979**, *19* (1), 39−45.

(24) Faught, E. Pharmacokinetic Considerations in Prescribing Antiepileptic Drugs. *Epilepsia* **2001**, *42*, 19−23.

(25) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714−3717.

(26) *ACDLABS*, version 9.0. http://www.acdlabs.com (accessed Oct 2006).

(27) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.

(28) Kier, L. B.; Hall, L. H. The Kappa Indices for Modeling Molecular Shape and Flexibility. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon and Breach: London, U. K., 1999.

(29) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular Connectivity. 1. Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64* (12), 1971−1974.

(30) Hall, L. H.; Kier, L. B. Structure−Activity Studies Using Valence Molecular Connectivity. *J. Pharm. Sci.* **1977**, *66* (5), 642−644.

(31) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69* (1), 17−20.

(32) Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological Indexes for Structure−Activity Correlations. *Top. Curr. Chem.* **1983**, *114*, 21−55.

(33) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity-Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 854−866.

(34) Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intellegence*; University of Michigan Press: Ann Arbor, MI, 1975; pp viii, 183.

(35) Friedman, J. *Multivariate Adaptive Regression Splines*; Laboratory for Computational Statistics: Stanford, CA, 1988.

(36) Kubinyi, H. Variable Selection in Qsar Studies. 1. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13* (3), 285−294.

(37) Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. Applications of Genetic Algorithms on the Structure−Activity Relationship Analysis of Some Cinnamamides. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 775−781.

(38) Hawkins, D. M.; Young, S. S.; Rusinko, A. Analysis of a Large Structure−Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16* (4), 296−302.

(39) Artursson, P. Cell-Cultures as Models for Drug Absorption across the Intestinal-Mucosa. *Crit. Rev. Ther. Drug Carrier Syst.* **1991**, *8* (4), 305−330.

(40) Artursson, P.; Karlsson, J. Correlation between Oral-Drug Absorption in Humans and Apparent Drug Permeability Coefficients in Human Intestinal Epithelial (Caco-2) Cells. *Biochem. Biophys. Res. Commun.* **1991**, *175* (3), 880−885.

(41) Gres, M. C.; Julian, B.; Bourrie, M.; Meunier, V.; Roques, C.; Berger, M.; Boulenc, X.; Berger, Y.; Fabre, G. Correlation between Oral Drug Absorption in Humans, and Apparent Drug Permeability in TC-7 Cells, a Human Epithelial Intestinal Cell Line: Comparison with the Parental Caco-2 Cell Line. *Pharm. Res.* **1998**, *15* (5), 726−733.

(42) Pade, V.; Stavchansky, S. Link between Drug Absorption Solubility and Permeability Measurements in Caco-2 Cells. *J. Pharm. Sci.* **1998**, *87* (12), 1604−1607.

(43) Rubas, W.; Jezyk, N.; Grass, G. M. Comparison of the Permeability Characteristics of a Human Colonic Epithelial (Caco-2) Cell-Line to Colon of Rabbit, Monkey, and Dog Intestine and Human Drug Absorption. *Pharm. Res.* **1993**, *10* (1), 113−118.

(44) Stewart, B. H.; Chan, O. H.; Lu, R. H.; Reyner, E. L.; Schmid, H. L.; Hamilton, H. W.; Steinbaugh, B. A.; Taylor, M. D. Comparison of Intestinal Permeabilities Determined in Multiple in-Vitro and in-Situ Models − Relationship to Absorption in Humans. *Pharm. Res.* **1995**, *12* (5), 693−699.

(45) Vandewaterbeemd, H.; Kansy, M. Hydrogen-Bonding Capacity and Brain Penetration. *Chimia* **1992**, *46* (7−8), 299−303.

(46) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption. *J. Pharm. Sci.* **1999**, *88* (8), 807−814.

(47) Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. On the Mechanism of Human Intestinal Absorption. *Eur. J. Med. Chem.* **2002**, *37* (7), 595−605.

(48) Golbraikh, A.; Tropsha, A. Beware of q(2)! *J. Mol. Graphics Modell.* **2002**, *20* (4), 269−276.

(49) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery − 1. Applications of Genetic Algorithms to the Prediction of Blood-Brain Partitioning of a Large Set of Drugs. *J. Mol. Model.* **2002**, *8* (12), 337−349.

(50) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J. Med. Chem.* **1998**, *41* (7), 1007−1010.

(51) Yoshida, F.; Topliss, J. G. QSAR Model for Drug Human Oral Bioavailability. *J. Med. Chem.* **2000**, *43* (13), 2575−2585.

CI600343X