molecular pharmaceutics

ARTICLE

pubs.acs.org/molecularpharmaceutics

ADME Evaluation in Drug Discovery. 9. Prediction of Oral Bioavailability in Humans Based on Molecular Properties and Structural Fingerprints

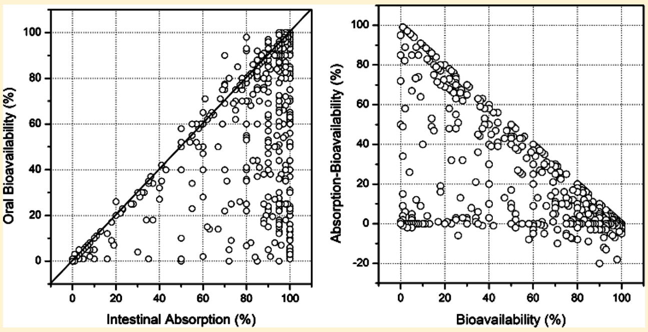
Sheng Tian,[†] Youyong Li,[†] Junmei Wang,[§] Jian Zhang,[*,‡] and Tingjun Hou[*,†]

[†]Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China

[‡]The Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiao Tong University, School of Medicine, Shanghai 200025, China

[§]Department of Pharmacology, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas 75390, United States

S *Supporting Information*



ABSTRACT: Oral bioavailability is an essential parameter in drug screening cascades and a good indicator of the capability of the delivery of a given compound to the systemic circulation by oral administration. In the present work, we report a database of oral bioavailability of 1014 molecules determined in humans. A systematic examination of the relationships between various physicochemical properties and oral bioavailability were carried out to investigate the influence of these properties on oral bioavailability. A number of property-based rules for bioavailability classification were generated and evaluated. We found that no rule was an effective predictor for oral bioavailability because these simple rules cannot characterize the influence of important metabolic processes on bioavailability. Finally, the genetic function approximation (GFA) technique was employed to construct the multiple linear regression models for oral bioavailability using structural fingerprints as the basic parameters, together with several important molecular properties. The best model is able to predict human oral bioavailability with an r of 0.79, a q of 0.72, and a RMSE (root-mean-square error) of 22.30% of the compounds from the training set. The analysis of the descriptors chosen by GFA shows that the important structural fingerprints are primarily related to important intestinal absorption and well-known metabolic processes. The predictive power of the models was further evaluated using a separate test set of 80 compounds, and the consensus model can predict the oral bioavailability with $r_{\text{test}} = 0.71$ and RMSE = 23.55% for the tested compounds. Since the necessary molecular properties and structural fingerprints can be calculated easily and quickly, the models we proposed here may help speed up the process of finding or designing compounds with improved oral bioavailability.

KEYWORDS: oral bioavailability, ADME/T, fingerprint, genetic algorithm, genetic function approximation, intestinal absorption

INTRODUCTION

The oral route is the most convenient way of administration for patients. Thus, oral administration is likely to be the desired route for any drug candidate. For any drug administrated by the oral route, oral bioavailability is undoubtedly one of the most important pharmacokinetic parameters because it is the indicator of the efficiency of the drug delivery to the systemic circulation. Unfortunately, oral bioavailability can only be measured in vivo; therefore, it is critical to develop in silico methods that can predict oral bioavailability reliably and quickly.[1−3]

To date, several computational approaches have been reported to predict oral bioavailability. The simple and popular computational approaches to estimate oral bioavailability are the property-based rules, such as Veber's rules and Martin's scoring scheme.[4,5] By analyzing a rat bioavailability database of 1100 drug or drug-like molecules, Veber proposed a simple rule with two molecular properties: molecules with oral bioavailability in rats exceeding 20% usually possess fewer than 10 rotatable bonds and a polar surface area (PSA) less than 140 Å^2 (or H-bond count less than 12). In 2005, Martin and co-workers proposed a simple scoring scheme to predict rat bioavailability based on several molecular properties, PSA, Lipinski's rule-of-five, and the molecular charged state.[5] The advantage of the simple rules based on several molecular properties is that they can be easily understood and utilized by scientists besides the computational chemists. However, the drawback is that they cannot characterize some

Received: December 21, 2010
Accepted: May 6, 2011
Revised: April 12, 2011
Published: May 06, 2011

complicated subprocesses for oral bioavailability, such as the metabolisms mediated by cytochromes P450 family, and therefore the prediction accuracy of those models is questionable. Previously,[6] we systematically evaluated some property-based rules in the prediction of intestinal absorption and oral bioavailability, and we found that no simple rule based on molecular properties could predict oral bioavailability with high confidence. Therefore, the application of the property-based rules on the prediction of human bioavailability should be carried out cautiously.

To build more general models for oral bioavailability, a larger number of molecular properties and more complicated multivariate data analysis and machine learning techniques, such as artificial neural networks (ANN), genetic algorithm (GA), support vector machine (SVM), and so forth have been applied.[7−11] However, most of these models do not show satisfactory predictions for oral bioavailability. For example, in 2000, Andrews and co-workers developed a regression model to predict bioavailability.[7] Compared with Lipinski's rule-of-five, the false negative rate was reduced from 5% to 3%, and the false positive rate decreased from 78% to 53%.[7] This model is not satisfactory because of the high rate of false positives. In 2000, Yoshida and Topliss proposed a classification model for bioavailability based on three descriptors related to distribution coefficient and 15 structural descriptors. The authors concluded that 15 structural descriptors used in the model were closely related to well-known metabolic processes. However, the classification model can only give a correct rate of 60% for the tested compounds.[8] In 2008, Ma and co-workers developed a prediction model of oral bioavailability using SVM combined with GA. The prediction model achieved acceptable overall classification accuracy (∼80%). However, this model could not give reasonable prediction accuracy for the low-bioavailability class (∼25%). Considering the high unbalanced nature of the oral bioavailability database, a prediction accuracy of ∼80% is meaningless if the model cannot derives good predictions for the low-bioavailability class.[9] Overall, no model can give reliable predictions for oral bioavailability. The poor predictions for oral bioavailability are understandable because oral bioavailability is a complicated parameter that involves a number of chemical and physical processes such as permeability, solubility, chemical stability, first-pass metabolism, and so forth. Each subprocess may lead to poor oral bioavailability, and that is to say, poor bioavailability may be caused by several different reasons.[1,2]

Another thing we need to emphasize is that the availability of experimental data for oral bioavailability is limited for both quantity and quality. The large in-house databases collected by pharmaceutical companies are usually not available for the public.[4,5,7,11] Recently, we have reported the largest data set of oral bioavailability in human of 768 compounds. Certainly the data set still need to be improved. New data are continuously added to our data set, and now we have 1014 compounds in the data set. We believe that the analysis based on the new data is more reliable. Here we will analyze the relationships between oral bioavailability and important molecular properties, propose, and evaluate the property-based rules for bioavailability classification. We expect to develop generally applicable prediction models using more complicated multivariate data analysis methods and machine learning techniques. We introduced structural fragments as descriptors in model development because we believe that some structural descriptors are necessary to characterize well-known metabolic processes. We utilized the GFA technique to automatically determine the most crucial structural descriptors for predictions from the huge pool of structural fingerprints. Compared with ANN and SVM, GFA can find a group of prediction models from a large numbers of samples efficiently, rather than one. Moreover, the models given by GFA are easily interpreted, not just "black boxes" given by ANN and SVM.

METHODS AND MATERIALS

1. Database of Oral Bioavailability in Humans. The oral bioavailability (%F) database in human includes 1014 structurally diverse drug and drug-like molecules. The experimental data in the database were collected from three sources. The first and the most important source is the previous version of the oral bioavailability database for 768 compounds in human reported by us;[6] the second source is the oral bioavailability data reported in the *Pharmacological Basis of Therapeutics*;[12] the third source is the data of oral bioavailability data reported in other publications. The experimental data reported in the *Pharmacological Basis of Therapeutics* were set to the highest priority for the duplicated entries. If two or more values for a drug are not from the *Pharmacological Basis of Therapeutics*, the average value was employed. Furthermore, the entries in the previous version of the oral bioavailability database[6] were checked carefully, and some errors have been corrected.

The structures of the new compounds were built using the SYBYL8.1 molecular simulation package.[13] Each molecule in the database was optimized by using molecular mechanics (MM) with the MMFF94 force field.[14] All molecules were saved to a MACCS sdf file and a SMILES database for further analysis. The 3-D structures and the corresponding oral bioavailability data for all compounds are available online (http://modem.ucsd.edu/adme).

2. Molecular Descriptors. We analyzed the relationships between oral bioavailability in humans and 17 molecular descriptors commonly used in ADME analysis. These descriptors include molecular weight (MW), topological polar surface area (TPSA), flexible rotatable bond count (N_{rot}), number of violations of the rule-of-five ($N_{\mathrm{rule\text{-}of\text{-}5}}$), H-bond donor count ($N_{\mathrm{HBD}}$), H-bond acceptor count (N_{HBA}), total H-bond count (N_{HB}), octanol−water partitioning coefficient (log P), apparent partition coefficients (log D_2, log $D_{5.5}$, log $D_{6.5}$, log $D_{7.4}$, and log D_{10}) at pH = 2, 5.5, 6.5, 7.4, and 10, intrinsic solubility (log S), radius of gyration (RadOfGyration), molecular surface area (MSA), and molecular volume (V). The descriptors RadOfGyration, MSA, and V were calculated using Discovery Studio molecular simulation package,[15] and the others were calculated using ACD/Laboratories (version 9.0).[16] Due to the "missing fragment" problem, log D or log S for nine molecules could not be calculated by ACD/Laboratories. Moreover, in ACD/Laboratories, the calculation scheme is only based on 2D structure, so the isomers for the same molecule are considered as duplicates, for example, esomeprazole and omeprazole or (R)-albuterol and (S)-albuterol. After eliminating the molecules without log D or log S values and duplicated isomers, the data set has 1004 molecules. The values for these descriptors are included in the sdf database file. The distributions of eight molecular properties are shown in Figure 1.

3. Developing Simple Rules to Predict Oral Bioavailability. We checked the performance of 35 rules with two molecular properties selected from MW, TPSA, N_{rot}, $N_{\mathrm{rule\text{-}of\text{-}5}}$, log P, log $D_{5.5}$, N_{HBD}, N_{HBA}, and N_{HB} (see Table S1 in the Supporting Information).

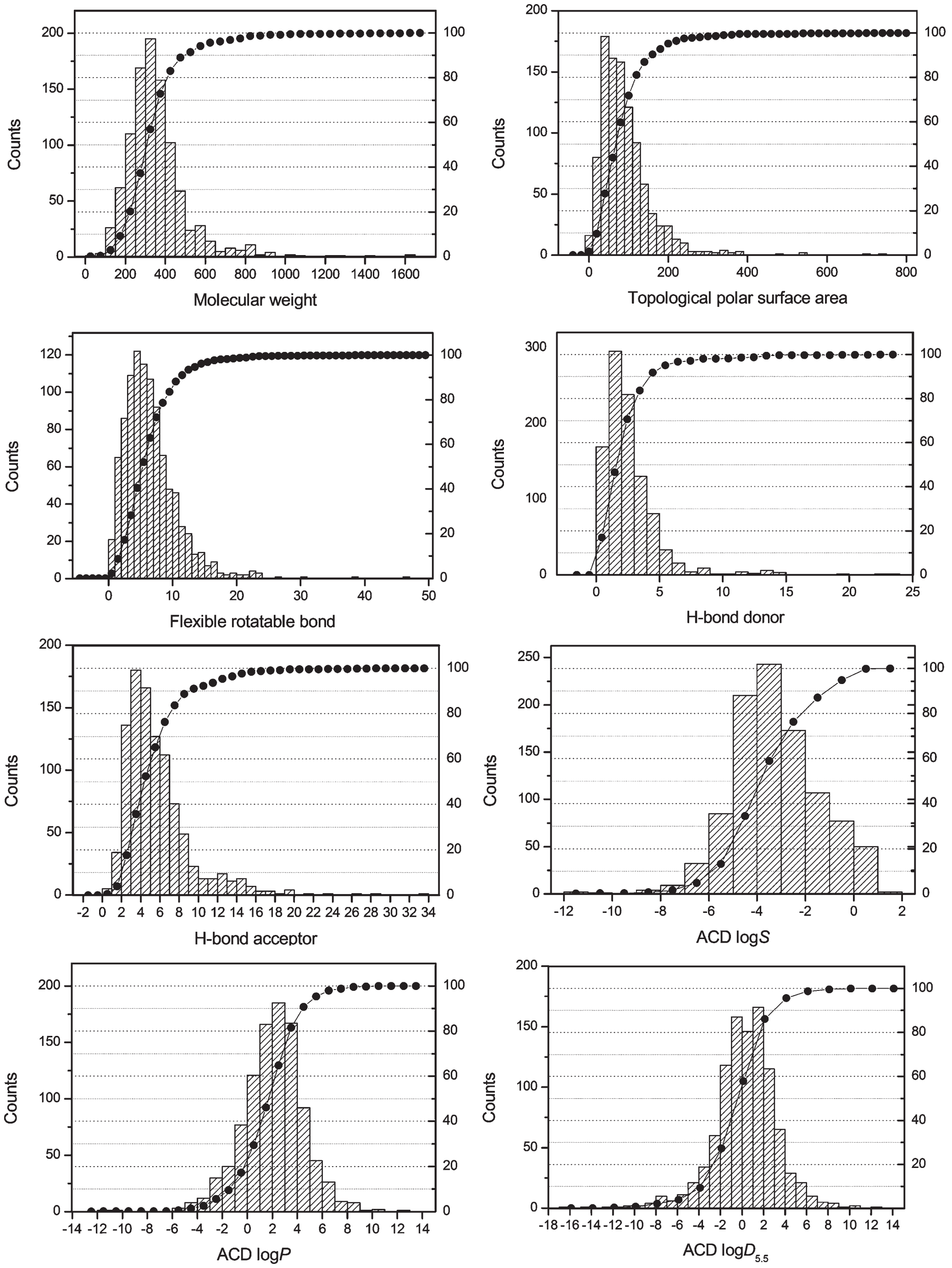


Figure 1. Distributions of eight studied properties.

For each rule, the best values for the two properties in one rule were determined by a grid search. For MW and TPSA the step of change is 10, for log *P* and log $D_{5.5}$ the step of change is 0.1, and for the others the step of change is 1. For each move in grid search, TP (true positive), FP (false positive), TN (true negative), and FN (false negative) were counted. The predictive performance was evaluated by sensitivity, SE = TP/(TP + FN); specificity, SP = TN/(TN + FP); prediction accuracy for the compounds with low bioavailability, Q_+ = TP/(TP + FP); prediction accuracy for the compounds with high bioavailability, Q_- = TN/(TN + FN); and Matthews correlation coefficient, $C = (TP \times TN - FN \times FP)/[((TP + FN)(TP + FP)(TN + FN)(TN + FP))^{1/2}]$. The Matthews correlation coefficient *C* ranges from 0 to 1 (*C* = 1 means perfect prediction).

4. Developing Quantitative Prediction Models of Oral Bioavailability Using Genetic Function Approximation (GFA). The ultimate goal of our study is to develop quantitative prediction models for oral bioavailability. Considering the huge number of molecular descriptors we used, a genetic function approximation (GFA) is applied to choose the best combination of descriptors automatically and construct the multiple regression models with the highest statistical significance. We eliminate eight molecules with molecular weight larger than 1000, and the final data set for the GFA calculations has 996 molecules. The

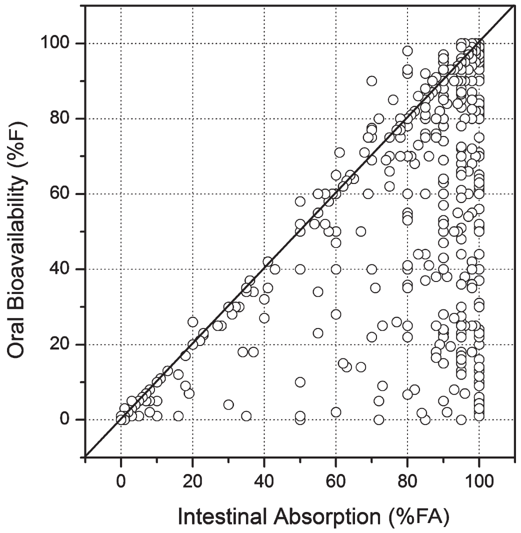


Figure 2. Scatter plot of oral bioavailability versus intestinal absorption for 513 compounds.

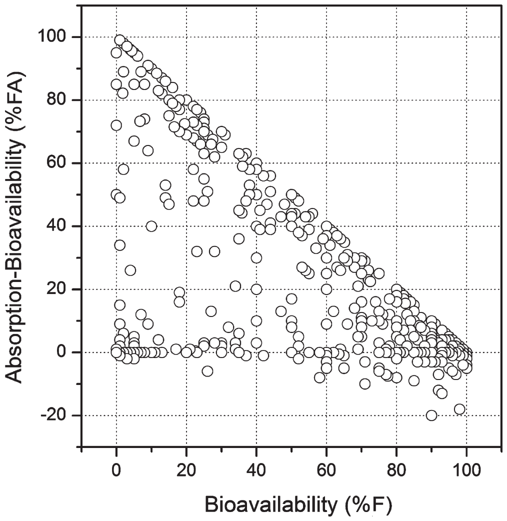


Figure 3. Scatter plot of oral bioavailability versus the difference between intestinal absorption and oral bioavailability for 513 compounds.

whole data set is split into a training set of 916 molecules and a test set of 80 molecules randomly selected from the data set.

In our analysis, we use two types of molecular descriptors: molecular properties and structural fingerprints. In total 17 molecular properties are used, including MW, TPSA, N_{rot}, $N_{\text{rule-of-5}}$, log P, log $D_{5.5}$, log $D_{6.5}$, N_{HBD}, N_{HBA}, N_{HB}, log S, RadOfGyration, MSA, V, log $D_{5.5}{}^2$, log $D_{6.5}{}^2$, and MW^2. The properties include the square values of three molecular properties. The reason for using log $D_{5.5}{}^2$ and log $D_{6.5}{}^2$ is that absorption from the intestinal tract of rats generally shows a parabolic relationship with log D.[17] The reason for using MW^2 is that MW^2 can improve the prediction of solubility.[18] Oral bioavailability is related with solubility, and therefore MW^2 may be a useful descriptor for predicting oral bioavailability.

Here, we use two sets of SciTegic extended-connectivity fingerprints (FCFP_6 and ECFP_6) and two sets of Daylight-style path fingerprints (FPFP_6 and EPFP_6). The FCFP_6 and FPFP_6 fingerprints are generated by using the functional-role code, which is a combination of a hydrogen-bond acceptor, hydrogen-bond donor, positively ionized or positively ionizable, negatively ionized or negatively ionizable, aromatic, and halogen. The ECFP_6 and EPFP_6 fingerprints are generated by using the atom-role code derived from the number of connections to an atom, the element type, the charge, and the atomic mass. The detailed descriptions of these fingerprints are reported in the literature.[19,20] The fingerprints we use here are different from the substructures in the prediction of oral bioavailability used previously.[8,11] The fingerprints used here represent a much larger set of features than a set of predefined substructures. Furthermore, these fingerprints do not need to be preselected or predefined because they are generated directly from the molecules. Therefore, novel molecular classes are as easily handled as the common classes. Here we generate the structural fingerprints by using Discovery Studio molecular simulation package.[15]

Considering the huge number of descriptors used in analysis, we cannot simply apply multiple linear regressions to develop prediction models. Here, the regression prediction models of F% with the most important molecular descriptors are constructed using the GFA technique[21] in Discovery Studio, which combined two different algorithms together; genetic algorithm (GA)[22] and the multivariate adaptive regression splines (MARS) algorithm.[23] MARS is a statistical technique for modeling data, which provides an error measure, called the lack of fit (LOF) score, to automatically penalize models with too many features. In MARS, nonlinear modeling can be achieved by using splines. In GFA, GA is applied to identify the best prediction models by automatically optimizing the combination of molecular descriptors and functional forms. The details of quantitative structure—activity relationship (QSAR) analysis based on GA or GFA are reported previously.[24,25] In this work, an initial population of 200 equations is generated randomly; then, pairs from the population of equations are chosen for "crossover" operations from this set of 200 equations randomly. The size of crossover operations is set to 5000. The leave-one-out (LOO) cross-validation coefficient (q) is used to evaluate the fitness of the equations. Cross-validated q^2 is defined as q^2 = (SSY − PRESS)/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. The regression coefficient (r), the adjusted regression coefficient (r_{adj}), the root-mean-square error (RMSE), and the variance ratio (F) are also reported. The adjusted r_{adj} is calculated by reducing the variance in proportion to the size of the estimated model, and the adjusted r^2 is defined as $r_{\mathrm{adj}}{}^2 = 1 - [(\mathrm{SSE}/(n-p)]/[(\mathrm{SST}/(n-1)]$, where SSE is the sum of squares of errors, SST is the total sum of squares, and p is the number of parameters in the regression equation. Considering that a large number of descriptors are used in regressions, we only include the linear terms to simplify the complexity of the GA optimization. The number of the descriptors is fixed in each cycle of the GFA calculations. We perform 13 separated GFA calculations with 10—130 descriptors with a step of 10.

■ RESULTS AND DISCUSSION

1. Relationship between Oral Bioavailability and Intestinal Absorption. For the 513 compounds common to the oral bioavailability and intestinal absorption data sets the correlation coefficient between oral bioavailability (%F) and intestinal

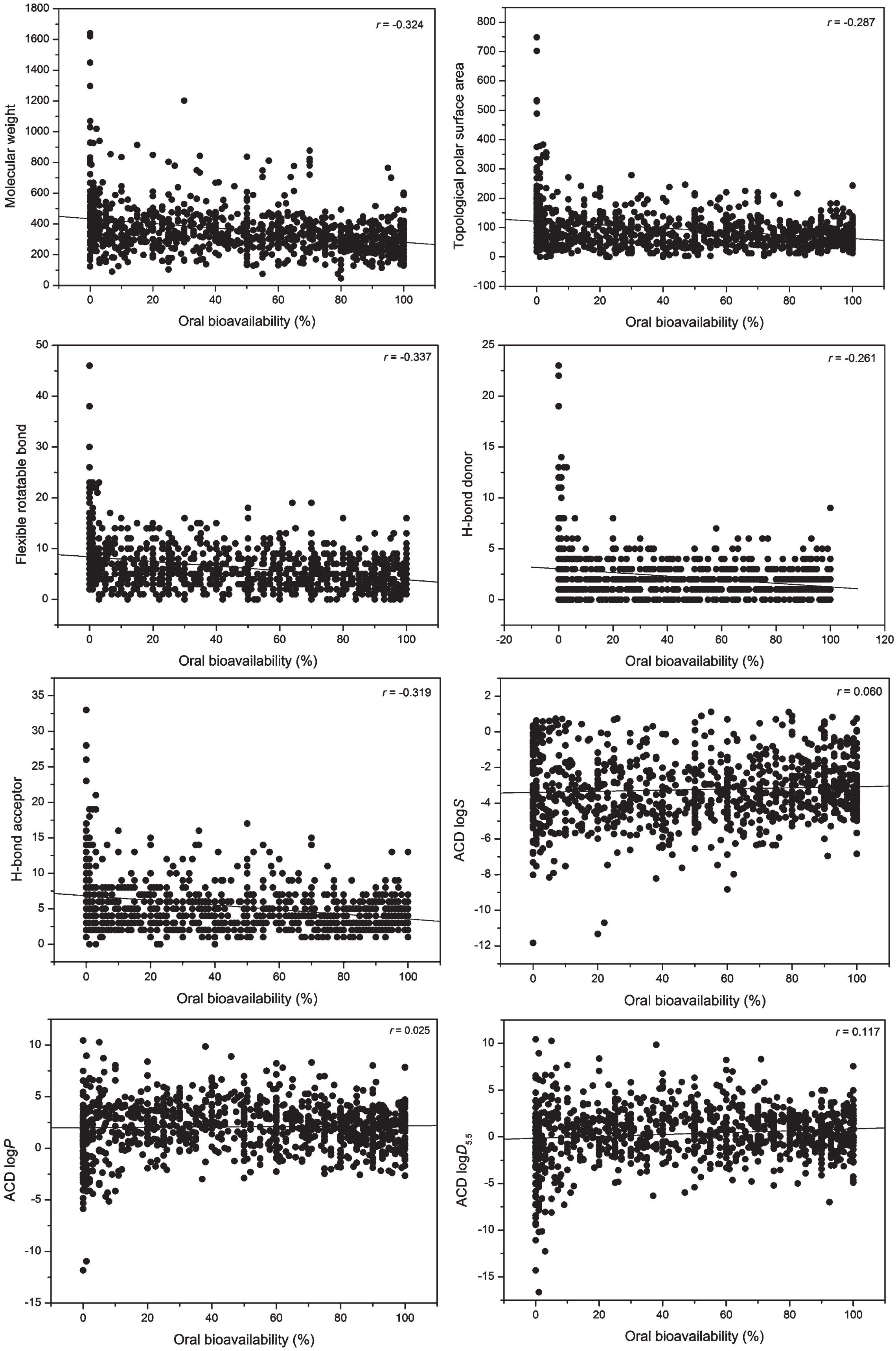


Figure 4. Correlations between oral bioavailability and eight important molecular properties.

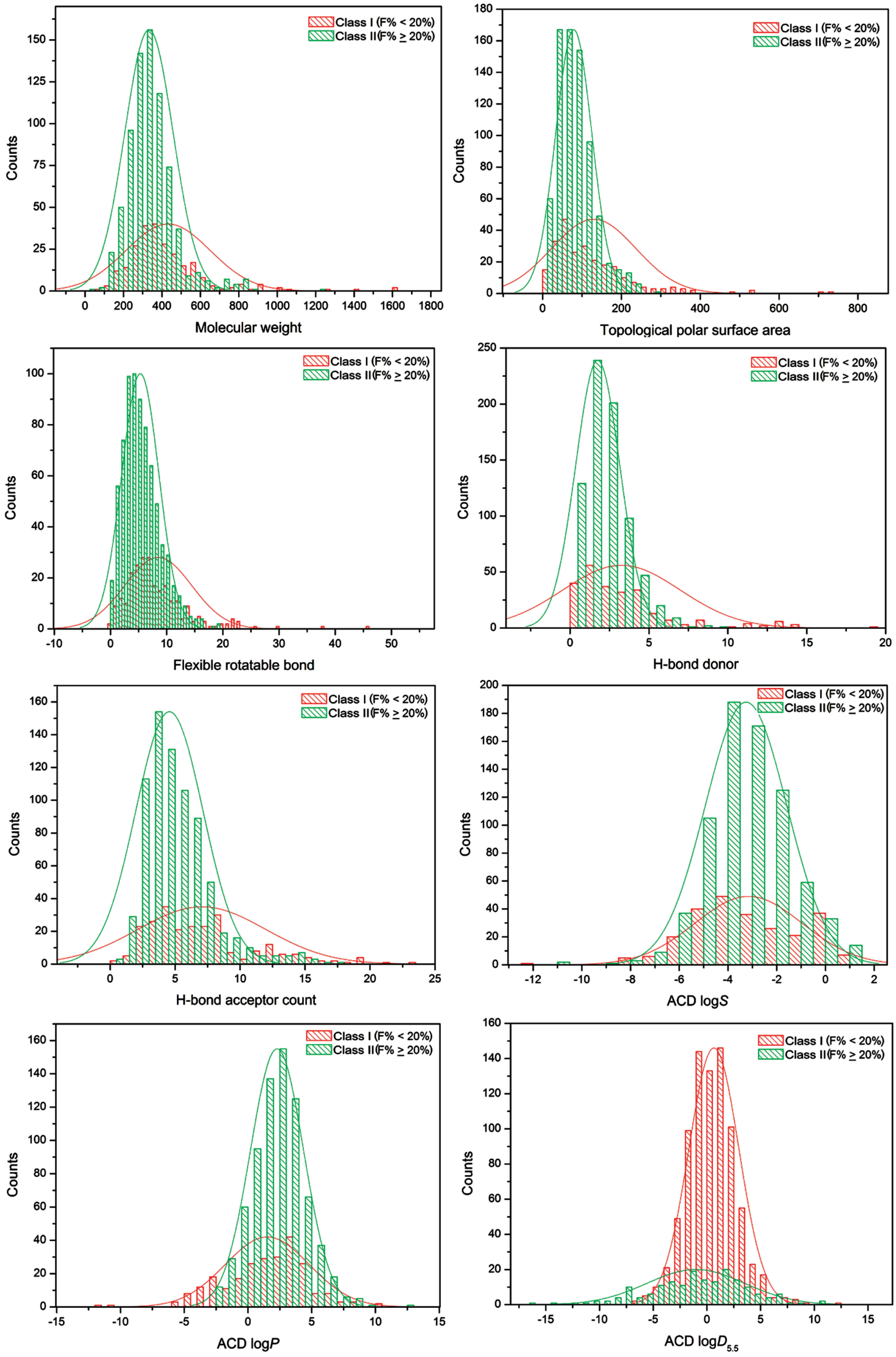


Figure 5. Distributions of eight molecular properties for Class I (%F < 20%) and Class II (%F ≥ 20%).

absorption (%FA) is 0.64. The values of %FA are obtained from the updated version of intestinal absorption database reported by us.[26] The scatter plot of oral bioavailability versus intestinal absorption is shown in Figure 2. In these 513 compounds, 183 compounds (35.7%) show a significant difference between %FA and %*F* ([%FA − %*F*] ≥ 20). These 183 compounds far from the diagonal in Figure 2 are considered strongly metabolized through liver while the others are not highly involved in metabolism. This conclusion is almost identical to our previous result based on 470 compounds.[26] We then conduct correlation analysis between [%FA − %*F*] and %*F*, and the correlation is −0.52 (Figure 3); that is to say, metabolism is a very important contributor for oral bioavailability. Moreover, according to our analysis, [%FA − %*F*] does not correlate well with any single molecular property. It explains the difficulty to predict oral bioavailability. The combined database of oral bioavailability and intestinal absorption for 513 compounds is available online (http://modem.ucsd.edu/adme).

2. Relationships between Oral Bioavailability and Molecular Properties. The correlation analysis between oral bioavailability and each molecular descriptor is conducted, and the scatter plots of eight molecular properties versus %*F* are shown in Figure 4. The compounds are split into two classes: Class I (%*F* < 20%) and Class II (%*F* ≥ 20%). The distributions of eight molecular properties for the two classes are shown in Figure 5.

In these molecular properties, the descriptor, the number of violations of the rule-of-five ($N_{\text{rule-of-5}}$), is the only one calculated based on the other several descriptors. As shown in Figure 4, the correlation coefficient between $N_{\text{rule-of-5}}$ and %*F* is 0.345, which is marginally better than that between N_{rot} and %*F* ($r = 0.337$) and those between the other properties and %*F*. As the most famous drug-likeness filter, the rule-of-five has been widely used to estimate the oral absorption and permeability,[27] so it is not surprising that it shows better correlation with %*F* than the other molecular properties.

For the properties we used here, seven of them are related to hydrophobicity, including log *P*, log D_2, log $D_{5.5}$, log $D_{6.5}$, log $D_{7.4}$, log D_{10}, and log *S*. Traditionally, log *P* or log *D* is considered to be associated with many ADME properties.[1,2,26,28−30] The distributions of log *P* and log $D_{5.5}$ are shown in Figure 4. log *P* is distributed between −11.82 and 12.77, with a mean of 2.09, and log $D_{5.5}$ between −16.64 and 12.77, with a mean of 0.35. According to the distributions of log *P* and log $D_{5.5}$ (Figure 4), the log *P* and log $D_{5.5}$ values for 90% of compounds in the data set are less than ∼4.4 and ∼2.8, respectively. The log *P* does not show any correlation with %*F* ($r = 0.025$). Then, the apparent partition coefficient at pH = 2.0, 5.5, 6.5, 7.4, or 10.0 is fitted with %*F*, respectively, and the correlations (r = 0.084, 0.117, 0.096, 0.071, and 0.041) are improved although not significant. The property, log $D_{5.5}$, shows better correlation than the other four log *D* values. The distributions of log $D_{5.5}$ of the compounds with %*F* ≥ 20% and those of the compounds with %*F* < 20% are shown in Figure 5. The mean values of log $D_{5.5}$ for the 177 compounds with %*F* < 20% versus the 818 compounds with %*F* ≥ 20% are −1.03 and 0.65, respectively. This result is reasonable since compounds with higher hydrophilicity usually have worse capability to diffuse through intestinal membrane. The Student's *t*-test is employed to evaluate the significance of the difference between the means. The *p*-value associated with the difference in the mean log $D_{5.5}$ values of the compounds with %*F* < 20% versus those with %*F* ≥ 20% was 2.35×10^{-12} at the 95% confidence level, indicating that the two distributions are significantly different. However, these two distributions are strongly overlapped, and there is no good boundary can be defined to distinguish these two classes.

Table 1. The *p*-Values of the Two Distributions for All Compounds, the 183 Compounds with Strong Metabolism and the 330 Compounds with Weak Metabolism

descriptor	*p*-value[a]	*p*-value[b]	*p*-value[c]
MW	3.93×10^{-10}	0.015	1.59×10^{-9}
TPSA	2.84×10^{-18}	2.89×10^{-4}	5.08×10^{-22}
N_{FRB}	6.35×10^{-16}	0.0116	5.13×10^{-17}
N_{HBD}	1.51×10^{-15}	0.00273	2.01×10^{-18}
N_{HBA}	1.43×10^{-16}	7.68×10^{-4}	1.02×10^{-18}
ACD log *S*	0.0168	0.984	3.56×10^{-5}
ACD log *P*	6.48×10^{-8}	0.349	9.11×10^{-17}
ACD log $D_{5.5}$	1.51×10^{-11}	0.511	7.69×10^{-23}

[a] *p*-values of the two distributions for all compounds. [b] *p*-values of the two distributions for the 183 compounds (class-HM) with strong metabolism. [c] *p*-values of the two distributions for the 330 compounds with weak metabolism (class-LM).

As shown in Figure 1, 90% of compounds in the database have a MW smaller than 500, and 95% of compounds have a MW smaller than 590. The correlation analysis shows that MW has a relatively high impact on oral bioavailability, indicated by the obvious anticorrelation between MW and oral bioavailability ($r = -0.32$). The other two properties, MSA and MV, have high correlations (r = 0.98 and 0.98) with MW. Similarly, these two properties also have similar correlations with oral bioavailability as that between MW and oral bioavailability. The distributions of MW for Class I and Class II are shown in Figure 5. The mean values for Class I and Class II are 432 and 332, respectively. The *p*-value associated with the difference in the mean MW values of Class I and Class II is 5.14×10^{-18} at the 95% confidence level, indicating that MW is better to discriminate Class I from Class II than log $D_{5.5}$. The impact of MW on bioavailability is reasonable since both permeability and solubility are closely related to this parameter.[18,26]

In our analysis, four properties, including TPSA, H-bond donor count (N_{HBD}), H-bond acceptor count (N_{HBA}), and total H-bond count (N_{HB}), are usually used to represent hydrophilicity. As shown in Figure 1, 90% of compounds in the database have a TPSA smaller than 160 Å^2, a N_{HBD} smaller than 5, a N_{HBA} smaller than 9, and a N_{HB} smaller than 13. So TPSA < 160 Å^2, $N_{\text{HBD}} < 5$, $N_{\text{HBA}} < 9$, or $N_{\text{HB}} < 13$ may be used as drug-likeness filters to distinguish drugs from nondrugs. As shown in Figure 4, the properties related to hydrophilicity show better correlations ($r = -0.287$ for TPSA, $r = -0.261$ for n_{HBD}, $r = -0.319$ for n_{HBA}, and $r = -0.298$ for n_{HB}) with %*F* than those related to hydrophobicity.

We then examine the performance of molecular properties to distinguish the two distributions of high and low bioavailability for the compounds with strong metabolism ([%FA − %*F*] ≥ 20) and those with weak metabolism ([%FA − %*F*] < 20). First, we split the 513 compounds with both bioavailability and absorption data into two subgroups: Class-HM of 183 compounds with (%FA − %*F*) ≥ 20 and Class-LM of 330 compounds with (%FA − %*F*) < 20. Then, for each class the two distributions for each molecular descriptor are generated, and Student's *t*-test is employed to evaluate the significance of the difference between the means. The *p*-values of the two

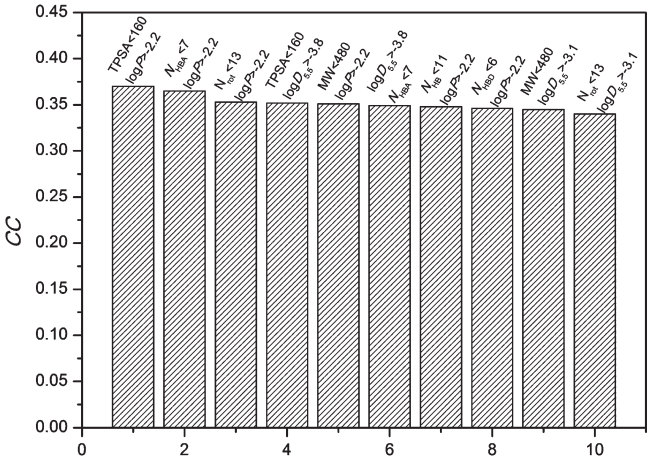


Figure 6. Matthews correlation coefficients (CC) of the best 10 rules for bioavailability classification.

distributions for Class-HM and Class-LM are listed in Table 1. As shown in Table 1, the molecular properties achieve a much better classification capability for Class-LM than that for Class-HM. In addition, for Class-HM, no single molecular descriptor shows significant *p*-value. That is to say, no single molecular descriptor can give good classification for compounds with strong metabolism.

3. Simple Rules To Predict Oral Bioavailability. The analysis in the previous section shows that a single molecular property is not enough for classifying oral bioavailability. However, there are studies suggesting that empirical rules based on molecular properties are useful for the early phase of drug discovery.[4,5] Veber proposed two simple rules to identify compounds with accepted oral bioavailability: A rat oral bioavailability of ≥20% is considered acceptable. TPSA ≤ 140 Å^2 (or sum of H-bond donors and acceptors ≤12) combined with $N_{\mathrm{rot}} \leq 10$ is an efficient and selective criterion.[4] On the basis of our current data, we re-evaluate Veber's rule. For the human data, 80.8% of molecules with bioavailability exceeding or equal to 20% meet the Veber's rule (TPSA ≤ 140 Å^2 and $N_{\mathrm{rot}} \leq 10$); however, only 56.6% of molecules with bioavailability less than 20% meet this criterion (TPSA > 140 Å^2 or $N_{\mathrm{rot}} > 10$).

Furthermore, we evaluate the performance of 35 simple rules with two molecular properties. The performance of bioavailability classification for these rules is shown in Table S1 in the Supporting Information. One interesting result is that, when TPSA and N_{rot} are used in the same rule, the cutoff value for TPSA is 140 Å^2, which is the same as the results reported by Veber et al.[4] and Clark.[31] Our results validate that the TPSA of 140 Å^2 is a good filter for predicting absorption-related properties. However, our results also show that the best value for N_{rot} is 13, not 10 reported by Veber.[4] The difference is attributed to that the human data used here is different from the rat data used by Veber et al.[4] Figure 6 shows the Matthews correlation coefficients of the best 10 rules for bioavailability classification. The best rule has two descriptors: TPSA < 160 Å^2 and $\log P > -2.2$. The performance of this rule is better than another rule: TPSA < 160 Å^2 and $\log D > -3.8$. This result is interesting because $\log D_{5.5}$ shows a better correlation with %*F* than $\log P$, but when combined with TPSA, $\log D_{5.5}$ shows worse performance for classification. According to the Matthews correlation coefficients, the best 10 rules show a similar performance. We conclude that some descriptors in the rules can be replaced by the others.

Table 2. Performance of 13 MLR Models for Predicting the Oral Bioavailability of the Training Set

no.	*n*[a]	*r*[b]	r_{adj}[c]	*q*[d]	RMSE[e]	*F*[f]
1	10	0.54	0.53	0.52	28.79	364.35
2	20	0.60	0.58	0.57	27.55	497.05
3	30	0.64	0.63	0.61	26.43	635.58
4	40	0.66	0.64	0.61	26.12	690.57
5	50	0.70	0.68	0.66	24.90	870.58
6	60	0.73	0.70	0.68	24.11	1011.76
7	70	0.73	0.71	0.68	23.97	1057.62
8	80	0.75	0.72	0.68	23.50	1161.52
9	90	0.76	0.73	0.70	23.17	1246.69
10	100	0.78	0.75	0.71	22.57	1390.08
11	110	0.79	0.75	0.72	22.30	1476.54
12	120	0.79	0.75	0.68	22.24	1519.63
13	130	0.81	0.77	0.69	21.67	1682.42

[a] *n* represents the number of descriptors used in the model. [b] *r* represents the regression coefficient. [c] r_{adj} represents the adjusted regression coefficient. [d] *q* represents the leave-one-out (LOO) cross-validation coefficient. [e] RMSE represents the root-mean-square error. [f] *F* represents the variance ratio.

Finally we test the best rule as shown in Figure 6: TPSA < 160 Å^2 and $\log P > -2.2$, to classify bioavailability for 330 compounds in Class-LM and 183 compounds in Class-HM. The classification results (sensitivity 93.3%, 249/267 and specificity 68.3%, 43/63) for Class-LM show satisfying performance. However, the classification results of the simple rule for Class-HM show poor performance, especially for the 50 compounds with %*F* < 20, indicated by a sensitivity of 87.3% (131/150) and a specificity of 10% (5/50). Overall, the simple rules give much better predictions for compounds with a weak metabolism than those with strong metabolism. Therefore, these simple rules can be used as effective predictors for intestinal absorption, but not for oral bioavailability. However these simple rules can be practically used in drug discovery because practically we are most concerned about the false prediction of bioavailable compounds. When the compounds are discarded on the basis of the prediction, there is a slim chance that those compounds are tested by experiments. If we use these simple rules in drug discovery, it is possible that some compounds with low bioavailability are predicted to be highly bioavailable. Most compounds with high bioavailability can be correctly predicted.

4. Predictions Models Based on GFA. In practical applications, the quantitative prediction models are usually more attractive than the qualitative prediction model, and this motivates us to develop reliable quantitative prediction models based on our new data.

First, we construct a multiple linear regression (MLR) model using seven molecular properties, including MW, TPSA, N_{rot}, $N_{\text{rule-of-5}}$, N_{HBD}, N_{HBA}, and $\log D_{5.5}$. The linear correlation coefficient (*r*) and the LOO cross-validation correlation coefficient (*q*) for the training set are 0.38 and 0.36, respectively, which are marginally better those between %*F* and N_{rot}. This suggests that the MLR model with seven important molecular properties does not give good predicted results. The reason for the poor prediction of oral bioavailability is that the hepatic metabolism cannot be effectively represented by these simple molecular properties. Therefore, the structural fingerprints are introduced to characterize the substructures closely related to specific metabolism processes.

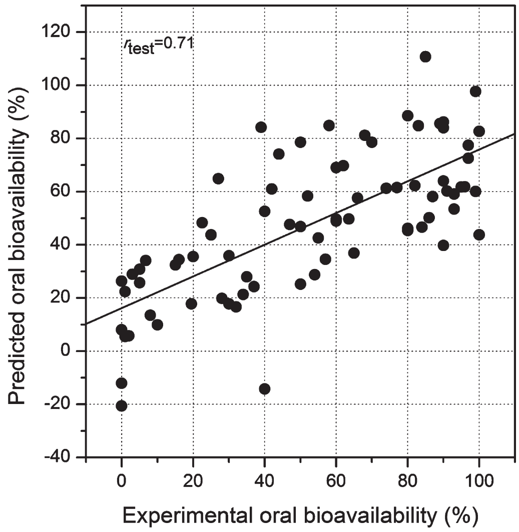


Figure 7. Scatter plot of the experimental oral bioavailability versus the predicted values for the tested compounds.

Then we build the MLR models using GFA based on molecular properties and structural fingerprints. The GFA technique can automatically select the meaningful fragments that make substantial contributions to oral bioavailability from a huge pool of fingerprints. The number of the descriptors used in the MLR models is changed from 10 to 130 as a step of 10. Table 2 summarizes the performance of 13 prediction models for oral bioavailability. The change of the r and q versus the number of descriptors is shown in Figure S1 in the Supporting Information. As shown in Table 2 and Figure S1, when the number of the descriptors increases, the r values increase too, but the q values do not always increase. When the number of the descriptors increases from 10 to 60, the q values increase rapidly from 0.52 to 0.68; then, the q values increase slightly from 0.68 to 0.72 when the number of the descriptors increases from 70 to 110. When the number of the descriptors increases further, the q values do not increase anymore. It is obvious that an equation including more than 120 descriptors is enough.

Validation is crucial in any QSAR modeling. Here, the actual prediction powers of models 6−11 are validated by an external test set of 80 compounds. According to the predictions, three compounds, including emepronium, avitriptan, and mycophenolic acid, could not be predicted well by all these six models (prediction error is usually larger than 50%). Therefore, these three compounds are considered as outliers and not included to estimate the actual prediction powers of these models. The r values for the test set are 0.68, 0.60, 0.65, 0.60, 0.66, and 0.67, respectively, and the RMSE values are 24.68%, 27.82%, 26.46%, 28.32%, 25.88%, and 26.06%, respectively. According to the prediction on the external test set, models 6, 10, and 11 are better than the others. Usually, selecting a single model while discarding the remaining models is not the most advantageous choice, and the average from the outputs of the multiple models is more reliable. So we use a consensus scoring scheme to predict the oral bioavailability of the tested molecules by averaging the predictions given by models 6, 10, and 11. Encouragingly, the consensus score improves the performance, and the r_{test} and RMSE are 0.71 and 23.55%, respectively. The scatter plot of the experimental oral bioavailability versus the predicted values for the tested compounds is shown in Figure 7.

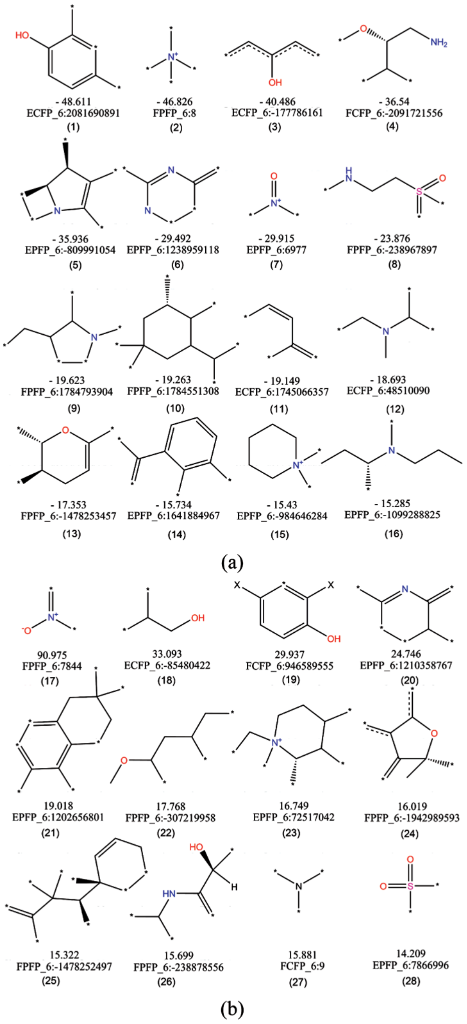


Figure 8. The 2-D structures, the fingerprint index, and the coefficients of (a) the 16 most important that can reduce bioavailability and (b) the 12 most important fingerprints that can boost oral bioavailability.

5. Analysis of the Structural Fingerprints in the Prediction Models. In the best MLR models given by GFA, model 6 in Table 2, only two molecular properties, log $D_{6.5}{}^2$ and TPSA, have been selected as important descriptors. That is to say, the other 58 descriptors in model 6 are structural fingerprints. The importance of each structural fingerprint is estimated by the respective coefficients. The most important 16 or 12 fingerprints with a high impact on oral bioavailability are shown in Figure 8.

By analyzing the fingerprints with negative coefficients shown in Figure 8a, we find that some of them are closely related to important metabolic pathways, for example, the phenolic OH and alcoholic OH (fragments 1 and 3), *tert*-alicyclic amine (fragments 5, 9, 12, and 16), and groups for allylic oxidation (fragment 11).[8] However the interpretation of these important fingerprints is not straightforward because not all fingerprints in Figure 8a are directly related to metabolism and some of them are involved in other important processes for bioavailability. For example, the groups with at least one charged nitrogen (fragments 2, 7, 15) are well-known to have significant negative impact on the diffusion of drugs.[26]

The positive coefficients of the fingerprints shown in Figure 8b indicate that these fingerprints can boost the intestinal absorption.

However, after careful analysis, we find that not all fingerprints are directly related to the increase of intestinal absorption. Many fingerprints shown in Figure 8b can be considered as the correction factors for the fingerprints shown in Figure 8a. For example, as we mentioned above, the groups with charged nitrogen (fragment 2) can reduce the intestinal absorption significantly. However, not all groups with charged nitrogen give the same negative contributions. After including the positive contribution of fragment 23 as shown in Figure 8b, the whole contribution of fragment 23 (−30.1) is less negative than the other structures with fragment 2 (−46.8). Similarly we observe that fragment 19 is a positive correction factor for fragments 1 and 3, fragment 17 is a positive correction factor for fragment 7, fragment 28 is a positive correction factor for fragment 8, and fragment 27 is a positive correction factor for fragments 5, 9, 12, and 16. These fragments have a positive contribution to the intestinal absorption.

CONCLUSIONS

Here we report an extensive database of oral bioavailability including 1014 drugs. On the basis of the large set of bioavailability data, the relationships between simple molecular properties and oral bioavailability have been studied. We generate a set of simple rules for bioavailability classification. These simple rules only show good performance for classifying intestinal absorption but not for classifying oral bioavailability. We successfully applied GFA to build up a set of MLR models using several molecular properties and structural fingerprints as descriptors. The qualities of these models are validated by the good prediction results on the training and test sets. The consensus score model, based on three comparative MLR models, is found to give the best prediction for the tested compounds: the correlation coefficient is 0.71, and the root-mean-square error is 23.55%. Our analysis of the key fingerprints provides a deep insight into the mechanism how the modeled molecular bioavailability is affected by its structural features.

The prediction of oral bioavailability is worse than that of intestinal absorption as we reported previously;[26] therefore, it is still necessary for us to improve the prediction models for oral bioavailability further. In this work, we use a training set of 916 to build up the prediction models. But our results show that the fingerprints generated based on the training set only cover a limited chemical space, and it is possible that predictions for not covered or newly developed classes of drugs may not be accurate. Therefore, the most important thing is to improve the reliability and extensiveness of the database of human oral bioavailability. It is necessary to check the reliability of the data from different sources carefully because the oral bioavailability data usually shows significant diversity from one source to another. It will be necessary to collect more experimental data into our database in the future. A more extensive database can help us to identify new fingerprints related to metabolism and absorption and construct a more universal prediction model which can cover various structural types.

ASSOCIATED CONTENT

Supporting Information. The classification capability of 35 simple rules with two molecular properties in Table S1. The change of the *r* and *q* versus the number of descriptors for the training set is shown in Figure S1. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Tingjun Hou: Institute of Functional Nano & Soft Materials (FUNSOM) and Jiangsu Key Laboratory for Carbon-Based Functional Materials & Devices, Soochow University, Suzhou, Jiangsu 215123, China. E-mail: tjhou@suda.edu.cn or tingjunhou@hotmail.com. Phone: +86-512-65882039. Fax: +86-512-65882846. Jian Zhang: The Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiao Tong University, School of Medicine, Shanghai 200025, China. E-mail: jian.zhang@sjtu.edu.cn. Phone: +86-21-63846590-776922.

ACKNOWLEDGMENT

The project is supported by the National Science Foundation of China (Grant No. 20973121), National Basic Research Program of China (973 Program) (Grant No. 2011CB504001), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). We thank Dr. Hong Liang in Maeja Pharmaceutical Inc. for pointing out some errors in the oral bioavailability database.

REFERENCES

(1) Hou, T. J.; Li, Y. Y.; Zhang, W.; Wang, J. M. Recent Developments of In Silico Predictions of Intestinal Absorption and Oral Bioavailability. *Comb. Chem. High Throughput Screening* **2009**, *12*, 497–506.

(2) Hou, T.; Wang, J. Structure - ADME relationship: still a long way to go? *Expert Opin Drug Metab. Toxicol.* **2008**, *4*, 759–770.

(3) Hou, T. J.; Xu, X. J. Recent development and application of virtual screening in drug discovery: An overview. *Curr. Pharm. Des.* **2004**, *10*, 1011–1033.

(4) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.

(5) Martin, Y. C. A bioavailability score. *J. Med. Chem.* **2005**, *48*, 3164–3170.

(6) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? *J Chem. Inf. Model.* **2007**, *47*, 460–463.

(7) Andrews, C. W.; Bennett, L.; Yu, L. X. Predicting human oral bioavailability of a compound: Development of a novel quantitative structure-bioavailability relationship. *Pharm. Res.* **2000**, *17*, 639–644.

(8) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem.* **2000**, *43*, 2575–2585.

(9) Ma, C. Y.; Yang, S. Y.; Zhang, H.; Xiang, M. L.; Huang, Q.; Wei, Y. Q. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method. *J. Pharm. Biomed. Anal.* **2008**, *47*, 677–682.

(10) Moda, T. L.; Montanari, C. A.; Andricopulo, A. D. Hologram QSAR model for the prediction of human oral bioavailability. *Bioorg. Med. Chem.* **2007**, *15*, 7738–7745.

(11) Wang, J. M.; Krudy, G.; Xie, X. Q.; Wu, C. D.; Holland, G. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. *J. Chem. Inf. Model.* **2006**, *46*, 2674–2683.

(12) Goodman, L. S.; Gilman, A.; Brunton, L. L. Teton Data Systems (Firm). *Goodman & Gilman's the pharmacological basis of therapeutics.* http://ezproxy.baylor.edu/login?url=http://online.statref.com/document.aspx?FxId=75&DocID=1&grpalias=BUSN (accessed May 11, 2011).

(13) *SYBYL molecular simulation package*, 2004. http://www.tripos.com (accessed May 13, 2011).

(14) Halgren, T. A.; MMFF, V. I. MMFF94s option for energy minimization studies. *J. Comput. Chem.* **1999**, *20*, 720–729.

(15) *Discovery Studio 2.5 Guide*; Accelrys Inc.: San Diego, 2009. http://www.accelrys.com.

(16) *ACDLABS v9.0*, 2005. http://www.acdlabs.com.

(17) Lien, E. J. Structure-activity relationships and drug disposition. *Annu. Rev. Pharmacol. Toxicol.* **1981**, *21*, 31–61.

(18) Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266–275.

(19) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, *10*, 682–686.

(20) *Pipeline Pilot 7.5*; Accelrys Inc.: San Diego, 2009.

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

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

(23) Friedman, J. H. *Multivariate Adaptive Regression Splines*, Technical Report 102; Stanford University: Stanford, CA, 1988.

(24) 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*, 775–781.

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

(26) Hou, T. J.; Wang, J. M.; Zhang, W.; Xu, X. J. ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* **2007**, *47*, 208–218.

(27) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(28) 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*, 337–349.

(29) Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.

(30) 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*, 1585–1600.

(31) 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*, 807–814.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper was published to the Web on May 16, 2011, with an error to the Introduction and the Results and Discussion section. The paper will repost with the Issue on June 6, 2011.