QSAR Model for Drug Human Oral Bioavailability[1]

Fumitaka Yoshida[†] and John G. Topliss*

Division of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

Received February 15, 2000

The quantitative structure−bioavailability relationship of 232 structurally diverse drugs was studied to evaluate the feasibility of constructing a predictive model for the human oral bioavailability of prospective new medicinal agents. The oral bioavailability determined in human adults was assigned one of four ratings and analyzed in relation to physicochemical and structural factors by the ORMUCS (**o**rdered **mu**lticategorical **c**lassification method using the **s**implex technique) method. A systematic examination of various physicochemical parameters relating primarily to absorption, and structural elements which could influence metabolism, was carried out to analyze their effects on the bioavailabilty classification of drugs in the data set. Lipophilicity, expressed as the distribution coefficient at pH 6.5, was found to be a significant factor influencing bioavailability. The observation that acids generally had better bioavailability characterizitics than bases, with neutral compounds between, led to the formulation of a new parameter, $\Delta \log D$ ($\log D_{6.5} - \log D_{7.4}$), which proved to be an important contributor in improving the classification results. The addition of 15 structural descriptors relating primarily to well-known metabolic processes yielded a satisfactory QSAR equation which had a correct classification rate of 71% (97% within one class) and a Spearman rank correlation coefficient (R_s) of 0.851, despite the diversity of structure and pharmacological activity in the compound set. In leave-one-out tests, an average of 67% of drugs were correctly classified (96% within one class) with an R_s of 0.812. The relationship formulated identified significant factors influencing bioavailability and assigned them quantitative values expressing their contribution. The predictive power of the model was evaluated using a separate test set of 40 compounds, of which 60% (95% within one class) were correctly classified. Since the necessary physicochemical parameters can be calculated or estimated and the structural descriptors are obtained from an inspection of the structure, the model enables a rough estimate to be made of the prospective human oral bioavailability of unsynthesized compounds. Also, the model has the advantage of transparency in that it indicates which factors may affect bioavailabilty and the extent of that effect. This could be useful in designing compounds which are more bioavailable. Refinement of the model is possible as more bioavailability data becomes available. Potential uses are in drug design, prioritization of compounds for synthesis, and selection for detailed studies of early compound leads in drug discovery programs.

Introduction

In the design of new drugs intended for oral use, enhancement of the oral bioavailability of an active lead compound is a subject of great importance. All too often promising new drug candidates fail because of inadequate bioavailability. Oral bioavailability involves several factors, such as gastrointestinal transit and absorption, chemical stability in the gastrointestinal tract, and the first pass effect of gut wall and liver metabolism. Although various QSAR studies of congeneric compounds have been reported concerning different processes affecting oral bioavailability,[2−4] an overall quantitative relationship between the oral bioavailability of structurally diverse compounds and their physicochemical/structural properties has proved to be an elusive goal due to the complexity of the factors involved. However, the ability to predict the approximate human oral bioavailability of compounds from their physicochemical properties and structure prior to synthesis would be of great practical benefit in the design of new drugs. Recently, Hirono et al.[5] published a study of the quantitative property−bioavailability relationships for 188 noncongeneric diverse organic medicinals. The compounds were divided into three groups, nonaromatics, aromatics, and heteroaromatics, and separate equations were formulated for each group which were statistically reliable and satisfactory. However, in addition to the need for prior classification of the compounds into one of the three groups, the lipophilicity of the compounds was not separately identified as a factor, although many studies have reported that this is one of the most important properties which determines absorption and metabolism.[2−4] Also some of the descriptors used were not clearly linked to those known to influence bioavailability through absorption and metabolism.

Our goal was to construct a simple method involving a single equation for predicting the approximate human oral bioavailability of new "drug-like" compounds based on the summation of a contribution value for various physicochemical and structural features related to

[†] Present address: Kumiai Chemical Industry Co., Ltd., Research and Development Dept., 4-26, Ikenohata 1-chome, Taito-Ku, Tokyo 110, Japan.

* Author to whom correspondence should be addressed. Tel: 734-763-4998. Fax: 734-763-2022. E-mail: jgt@umich.edu.

10.1021/jm0000564 CCC: $19.00
Published on Web 06/07/2000

Table 1. Bioavailability Classification

	class 1	class 2	class 3	class 4
rating	1	2	3	4
bioavailability (%)	≤20	20−49	50−79	≥80

absorption and metabolism. If such a relationship could be successfully established, it would not only permit the estimation of the approximate bioavailability of candidate compounds prior to synthesis but would also provide information concerning the modification of physicochemical and structural factors necessary for the enhancement of oral bioavailability. In this work we have attempted to determine the important factors influencing bioavailability and give them quantitative values expressing the contribution to bioavailability employing the ORMUCS (**o**rdered **mu**lticategorical **c**lassification method using the **s**implex technique) method. This method is a modified form of discriminant analysis[6] designed for use in QSAR work involving the type of noncontinuous activity (property) data utilized in this study (see the Methods section). The compounds used in the analysis consisted of 130 bases, 58 neutral compounds, and 44 acids for a total of 232. Compound selection for the analysis was based on the accessibility of human bioavailability data and physicochemical data and the objective of having as diverse a group of drugs as possible with regard to pharmacological activity class and physicochemical characteristics. In the model development, compounds with experimentally determined partition coefficients were predominantly employed in order to provide the most accurate assessment of the contribution of this parameter.

Methods

Bioavailability. Bioavailability data in healthy human subjects were collected from the literature.[7] When two or more values for a drug were available or a value had a range, the averaged value was employed. However, for benzodiazepines, there were large differences in data originating from different sources, so the datum of ref 6a was used for consistency. In the case of drugs whose bioavailability is dose-dependent, the data obtained at the normal therapeutic dose was used. The bioavailability value of each drug was represented as a rating according to the ranges shown in Table 1 to reflect the fact that there was variability in the data which was determined under somewhat different experimental conditions by many investigators. The ranges were set to represent reasonable and useful categories for differentiation of bioavailability.

There were 37 drugs in class 1, 54 in class 2, 63 in class 3, and 78 in class 4.

In this paper bioavailability represents the percentage of an administered dose of a parent compound reaching the systemic circulation after oral administration. Types of drugs excluded from the analysis were prodrugs, whose bioavailability values were for an active metabolite; unstable compounds, such as nitroglycerine; those existing in the neutral form mainly as zwitterions at physiological pH, such as ACE inhibitors; and quaternary ammonium compounds.

Physicochemical and Structural Parameters. As an overall measure of lipophilicity, the log *P* value (*n*-octanol/water) was used. These were mostly taken from the literature[8] or otherwise calculated using the CLOGP program[9] or the MLOGP method[10] or estimated from that of a similar compound. For ionizable compounds, values of log *D* (log distribution coefficient) at a given pH were calculated[11] from log *P* and literature or estimated pK_a values[12] using a simply constructed program in Excel. In this analysis, the log *D* values at the pH of blood and the small intestine were considered to be of primary relevance. The pH of the blood was taken as 7.4 and the pH of the small intestine as 6.5.[13] Other physicochemical parameters considered which may relate to absorption were hydrogen-bonding potential, molecular weight, and solubility.

Possible structural descriptors relating to metabolism considered were those for readily hydrolyzable entities such as esters, lactones, and carbamates; readily oxidized moieties such as dihydropyridines and thiols; phenols and alcohols (conjugation/oxidation); sulfonamides (conjugation); aromatic, benzylic, and allylic oxidation; aromatic/heterocyclic amines (acetylation); N and O dealkylation; ketones (reduction); aromatic nitro groups (reduction) and amides (cleavage). A key factor with regard to the effect of metabolism of a particular structural entity on bioavailabity is the rate at which this takes place. Thus, some well-known metabolic transformations may occur too slowly to have a material effect on bioavailability.

ORMUCS Method. The ORMUCS method was developed by Takahashi et al.[14] and is useful for the development of discriminant functions for modeling ordered classes. It is an adaptive least squares (ALS) related approach using a simplex technique for the derivation of a single discriminant function and was shown to be more stable than the ALS method in this respect. A detailed description and applications of the ORMUCS method have been reported.[14−16] Only the principal features are presented here. The discriminant function of classification is

$$S(\mathbf{X}) = \mathbf{W}\mathbf{X} = \sum_{i=1}^{d} w_i x_i \qquad (1)$$

described by eq 1 where x_i is the ith component of $\mathbf{X}$ represented as a pattern vector in the d-dimensional measurement space, and w_i is the weight assigned to the component. The vector $\mathbf{W}$ consists of w_1, w_2, ..., w_d, and is called a weight vector. The rules of classification are defined by the following equation:

$$C^{(k)}_{\min} \le S(\mathbf{X}) < C^{(k)}_{\max} \quad \text{implies } \mathbf{X} \in \text{class } k \qquad (2)$$

In eq 2, $C^{(k)}_{\min}$ and $C^{(k)}_{\max}$ are the lower and upper limits of the discriminant score for the category k, respectively. These rules are also adopted to develop the discrimination function. The optimization procedure of the weight vector in this function as shown in eq 1 is described in detail elsewhere.[14] Two criteria were employed for the evaluation of the weight vector. One is the recognition rate R shown below.

$$R = N_c/N_t \qquad (3)$$

N_c is the number of samples correctly classified using the weight vector, and N_t is the total number of samples. The other is the perceptron function J.

$$J = \sum_{X_i \in S_E} |\mathbf{W}\cdot\mathbf{X}^{(k)}_i - C^{(k)}_{\text{lim}}| \qquad (4)$$

where

$$C^{(k)}_{\text{lim}} = C^{(k)}_{\min} \quad \text{for} \quad \mathbf{W}\cdot\mathbf{X}^{(k)}_i < C^{(k)}_{\min}$$

$$C^{(k)}_{\text{lim}} = C^{(k)}_{\max} \quad \text{for} \quad \mathbf{W}\cdot\mathbf{X}^{(k)}_i > C^{(k)}_{\max} \qquad (5)$$

In eq 5, $\mathbf{W}$ is the weight vector to be evaluated, $\mathbf{X}_i$ is an element of the set of patterns misclassified, S_E, and k is the class to which the pattern $\mathbf{X}_i$ belongs. $C^{(k)}_{\text{lim}}$ is the limited value of the class k. Among the weight vectors, a larger value of R is more desirable. If the values of R for one or more weight vectors are equal to one another, the one with the smallest value of J should be selected as the best weight vector.

Computation. All computations were carried out on a Power Macintosh 7100/80AV. The original ORMUCS program

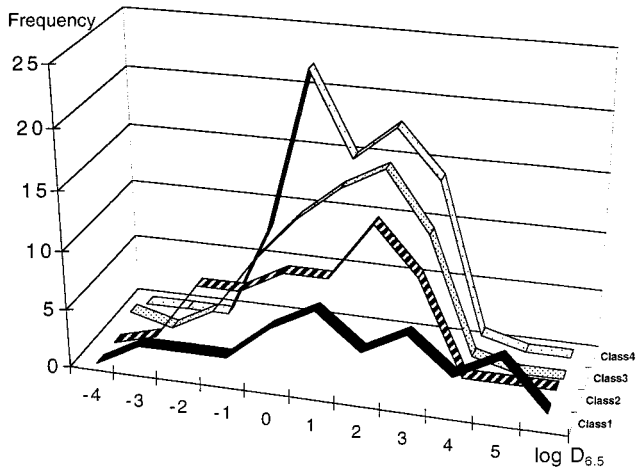

Figure 1. A histogram expressing the distribution of log $D_{6.5}$ within each bioavailability class.

Table 2. log $D_{6.5}$ Values by Bioavailability Class

bioavailability class	log $D_{6.5}$ range	% compounds
4	−2.0 to 3.0	99
3	−2.0 to 3.0	94
2	−2.0 to 3.0	81
1	−2.0 to 3.0	65

Table 3. Composition of Data Set by Bioavailability Class and Compound Type

bioavail. class	base	neutral	acid	total
1	20	17	0	37
2	46	3	5	54
3	45	11	7	63
4	19	27	32	78
total	130	58	44	232

developed by Dr. T. Takahashi was converted for use on the Power Macintosh 7100/80AV with the assistance of Mr. E. Kouno.

Results and Discussion

Relationship between Bioavailability and Lipophilicity. Numerous QSAR studies have reported that the lipophilicity of a compound plays a significant role in its absorption in the gastrointestinal tract and metabolism in the gut wall and liver.[2−4] The most relevant measure of lipophilicity with regard to oral absorption by passive diffusion is probably the distribution coefficient (log *D*) at pH 6.5, which is the pH of the small intestine, where absorption mostly takes place.[13] In the present study the distribution of log *D* (pH 6.5) within each bioavailability class is illustrated in Figure 1. It can be seen that the distribution peak broadens as bioavailability decreases. Also, there is some tendency for compounds with higher log *D* values to have poorer bioavailability. Further insight is provided by the data in Table 2, which shows that over 99% of highly bioavailable compounds have log $D_{6.5}$ values in the range −2.0 to 3.0. Also (Table 3) it was observed that, in general, bases showed poorer bioavailability characteristics than acids. Thus, whereas bases were about equally divided between the less bioavailable compounds (classes 1 and 2) and the more bioavailable compounds (classes 3 and 4), some 89% of acids were in the higher bioavailability categories. Neutral compounds were split about two to one, favoring higher bioavailability. This led to the formulation of a new parameter, Δ log *D* (= log $D_{6.5}$ − log $D_{7.4}$), which expresses the difference between the fractions of the neutral form at two given pH values, 6.5 and 7.4, for an ionizable compound. Compounds with positive values of Δ log *D* are acidic, whereas compounds with negative values are basic. Thus, compounds with positive Δ log *D* values will tend to have better bioavailability characteristics, other things being equal. This may reflect the fact that these compounds have a higher fraction of neutral compound present (the form in which the compound is absorbed) at the relevant absorption pH (6.5) and, once absorbed, a lower fraction of neutral compound present (pH 7.4) that would be subject to liver metabolism (in contrast with the ionized form). As is apparent from eqs 6 and 7 in Table 4, employment of the Δ log *D* parameter in combination with the parabolic form of log $D_{6.5}$ resulted in a marked improvement in the classification results. Thus, from eq 7 the optimum value of log $D_{6.5}$ for the overall processes involving absorption and first-pass effect is around 0.7. It has been reported[17] that absorption from the intestinal tract of rats generally shows a parabolic relationship with log *P* (or log *D* for ionizable compounds) and an optimal value of around 2.0 for log *P*. The optimum value from the present analysis is lower, reflecting possible species differences and the fact that bioavailability involves metabolism, which tends to increase with increasing lipophilic character.

Table 4. Initial Classification Results

eq	log $D_{6.5}$	(log $D_{6.5}$)^2	Δ log *D*	intercept	*n*	n_{mis}	R_s
6	0.207	−0.253		5.039	232	131(68)	0.342
7	0.207	−0.145	1.496	4.906	232	118(32)	0.562

QSAR Model. Proceeding from the results discussed in the preceding section, other physicochemical parameters identified previously were examined along with the log $D_{6.5}$ and Δ log *D* terms in an attempt to further improve the QSAR model. Hydrogen-bonding potential and molecular weight at high values are known to adversely affect absorption;[18] however, these were not material since an insignificant number of compounds in the data set had such high values. A drug may be poorly absorbed if its water solubility is very low, and this effect will be dose dependent. An example is mercaptopurine, which has a very low solubility (<0.1 mg/mL) and a bioavailability of 12% at an approximate dose of 360 mg.[19] The inclusion of dose and solubility data in the classification analysis was unnecessary since solubility correlates well with log $D_{6.5}$ for the vast majority of compounds. In addition to exploration of the physicochemical parameters, structural descriptors relating to metabolism listed previously were examined as possible variables in the model. A descriptor was not incorporated unless it was present in at least six compounds. These descriptors were added to the model if they resulted in a significant improvement of the classification results as judged by the number of compounds misclassified by one category, those misclassified by two categories, and the Spearman rank correlation coefficient. Descriptors were added to the model in order of their importance in improving the classification results. In general, descriptor definitions were made as broad as possible and, with the exception of descriptors

Table 5. The QSAR Model

$$S(\mathbf{X}) = \Sigma w_i s_i \qquad (8)$$

no.	descriptors s_i	weight w_i	CI[a]	n[b]
1	log $D_{6.5}$	−0.027	0.05	232
2	$(\log D_{6.5})^2$	−0.046	0.25	232
3	Δ log D (log $D_{6.5}$ − log $D_{7.4}$)	0.370	0.23	232
4	phenolic OH[c] (excluding di-ortho-subst)	−1.032	0.45	22
5	SO_2NH_2	−1.014	0.17	7
6	alcoholic OH (excluding *tert*-OH)[d]	−0.177	0.09	59
7	hydrolysis: esters,[e] lactones, β-lactams, alkyl carbamates	−1.074	0.37	24
8	aromatic *p*-hydroxylation[f]	−0.599	0.26	33
9	$ArCH_2-R$[g] (excluding di-ortho subst Ar)	−0.235	0.12	47
10	allylic oxidation (C−C=C)[h]	−0.201	0.09	13
11	*tert*-alicyclic amine (no ring heteroatoms)[i]	−0.340	0.10	24
12	XCCNR (R = Me, Et; X = N,O, Ar, C=C)[j]	−0.410	0.15	28
13	readily oxidized moieties: thiols, dihydropyridines	−1.137	0.24	11
14	ketones[k]	−0.493	0.10	15
15	NO_2 on a benzene ring (excluding *ortho* subst)	−0.148	0.03	7
16	$ArNH_2$, $ArNHNH_2$, $ArCONHNH_2$, $ArC(=NH)NH_2$ as pK_a value[l]	−0.034	0.04	16
17	HOCCNH *tert*-alkyl, HOCCN< (cyclic rings)	0.210	0.05	16
18	benzodiazepine (with no additional fused rings)	0.231	0.05	10
	constant	4.358		

n = 232 (four classes); boundaries, 2.0, 3.0, 4.0; recognition, n_{mis} = 67(8), R_s = 0.851 (p < 0.0001); leave-one-out, n_{mis} = 76(10), R_s = 0.812 (p < 0.0001)

[a] Contribution index. The product of the weighting coefficient and standard deviation for each descriptor. [b] n = total number of each of the descriptors used in the analysis. [c] Except with *o*-CO_2H, *o*-$CONH_2$ and *o*-CH_2OH substituents, which can undergo intramolecular hydrogen bonding with the phenolic OH group. [d] For steroids this descriptor is 0 for 11-β-OH substituents (steric hindrance) and 2 for 17-β-OH substituents (unless *tert*) due to high susceptibility to first-pass metabolism. [e] Weighting is 0.5, where the carbon α to the carbonyl is tertiary or where the carbonyl can undergo intramolecular hydrogen bonding with a nearby group. [f] Applies where there is an open para position with respect to the activating groups OR, $N(R)R_1$, $NHC(=O)R$ (R,R_1 = H, alkyl, aryl, aralkyl) with no ring substituents beyond one ortho to the activating group. The activating group and an ortho substituent may be part of a fused ring. [g] R = H, CH_2X, where X = C or H and is not attached to a polar atom. [h] Excluding steroidal ring A dienones and allylic substructures of the types C=C−C−X and C=C−C−C−X, where X is a polar atom. [i] Excluding ring systems with a bridged N atom such as quinine and quinidine. [j] Weighted by 2 for O=CCNR; see the text. [k] Excluding α,β-unsaturated dienones, diaryl-ketones, and ketones with a heteroatom attached at the α-position. Corrected (−0.5) for α-branching on the aliphatic side and α,β-unsaturation. [l] If a molecule contains two or more amino groups the most basic group with no ortho substitution is selected.

17 and 18, all were based on well-known metabolic pathways. However, it was necessary to refine these to reflect various factors, mainly steric and electronic effects, likely to affect metabolism, to achieve optimal significance of the descriptor in the model. As part of this process metabolic data on individual or subclasses of compounds were considered. Also, although most of the structural limitations in the descriptor definitions were based on understood factors, some were empirically derived.

The end result of this process was eq 8 with 18 descriptors, which is presented in Table 5. In Table 6, the classification results and compound scores of individual compounds from eq 8, which determine their classification, are shown. In this equation, n is the number of compounds in the analysis, and n_{mis}, the number misclassified. The figure in parentheses shows the number misclassified by two ratings. R_S is Spearman's rank correlation coefficient. The weight vector, w_i, denotes the weighting of a particular descriptor in contributing to the total classification score. Descriptors take the value of 1 unless otherwise indicated. For the structural descriptors (except no. 16) the weight vector is multiplied by the number of times it is present in a compound. The classification score for a compound is the sum of the contributions from all of the individual descriptors which apply to that compound. The contribution index, CI, is the product of the weight vector and the standard deviation and indicates the degree of contribution of a descriptor to discrimination between the different classification categories. Higher values denote a higher degree of contribution and importance in achieving the discrimination. Boundary numbers represent the scores of boundary points between the four bioavailabilty classes. Compound scores of >4 predict a compound to belong to bioavailability class 4, scores between 3 and 4 correspond to bioavailability class 3, scores between 2 and 3 indicate bioavailability class 2, and a score of below 2 predicts bioavailability class 1. From eq 8 in Table 5 it can be seen that the QSAR model correctly classifies 165 of the 232 compounds (71%) in the data set with 8 of the 67 misclassified compounds off by two categories. Thus, 224 of 232 compounds (97%) are correctly classified ± one class. The latter is an appropriate measure to include as a criterion in judging the model given the large range of some bioavailabilty data, which may straddle adjoining classes. The Spearman rank correlation coefficient of 0.851 is highly significant. The statistics for the leave-one-out tests, which relate to the predictive power of the equation, are also favorable. The degree of improvement achieved by the addition of structural descriptors relating to metabolism can be seen by comparing eq 8 with eq 7. A more detailed discussion of the classification results is presented in a later section.

The important role of lipophilicity terms in the final model is shown by their CI values. From eq 8 the optimum log $D_{6.5}$ is −0.3 and a progressive negative impact on bioavailability is seen as values move away from this level. A log $D_{6.5}$ value of 3.0 results in a reduction of the bioavailability score of 0.5 and a value of 5.0 reduces the score by 1.3. The Δ log D descriptor

Table 6. Compound Classification and Scores

							class		
no.	drug	type[a]	log $D_{6.5}$	Δ log D^b	pK_a	structural descriptors[c]	obsd	calcd	score (calcd)
1	adrenaline	B	−3.49	−0.90	9.9	4(2), 6, 12	1	1	0.90
2	alprenolol	B	0.40	−0.89	9.2	6, 8, 9, 10	1	2	2.80
3	clomethiazole	N	2.12	0.00		9	1	3	3.86
4	coumarin	N	1.39	0.00		7, 9	1	2	2.92
5	dobutamine	B	−0.48[e]	−0.90	9.5	4(3), 9	1	1	0.70
6	domperidone	B	2.83	−0.75	7.7	8, 11	1	2	2.70
7	dopamine	B	−3.37	−0.90	8.9	4(2), 9	1	1	1.29
8	epanolol	B	−0.40	−0.77	7.8[f]	4, 6, 8	1	2	2.27
9	estradiol	N	4.01	0.00		4, 6(2), 9	1	1	1.89
10	felodipine	N	3.22[e]	0.00		7, 13	1	1	1.58
11	hydralazine	B	0.14	−0.61	7.3	8(1.5), 16(7.3)	1	2	2.98
12	isradipine	N	1.72[e]	0.00		7,13	1	1	1.96
13	ketamine	B	1.14	−0.69	7.5	12(2), 14(0.5)	1	2	2.95
14	lofepramine	B	4.58[e]	−0.32	6.7[f]	8(2), 9(2), 12(2)	1	1	0.66
15	lovastatin	N	4.26	0.00		6, 7(2), 10(2)	1	1	0.68
16	mebendazole	N	2.83	0.00		7	1	2	2.84
17	meptazinol	B	0.71[e]	−0.88	8.7	4,11	1	2	2.62
18	mercaptopurine	B	−1.12	−0.72	7.6	13	1	2	2.93
19	nabumetone	N	2.83[d]	0.00		9, 14	1	3	3.18
20	nalbuphine	B	−0.63	−0.88	8.7	4, 6, 11	1	2	2.48
21	naloxone	B	0.64	−0.81	7.9	4, 11	1	2	2.65
22	nimodipine	N	0.73	0.00		7, 13, 15	1	1	1.96
23	nisoldipine	N	1.58	0.00		7, 13	1	1	1.99
24	nitrendipine	N	0.97	0.00		7, 13, 15	1	1	1.93
25	phenolphthalein	N	2.41	0.00		4(2), 7	1	1	0.89
26	probucol	N	7.29[e]	0.00			1	1	1.70
27	prochlorperazine	B	2.51	−0.69	7.5	8, 12	1	2	2.73
28	progesterone	N	3.87	0.00		10, 14	1	2	2.87
29	selegiline	B	0.62[e]	−0.89	9.2[f]	12	1	3	3.58
30	simvastatin	N	4.68	0.00		6, 7(2), 10(2)	1	1	0.49
31	sumatriptan	B	−2.07	−0.90	9.4[f]	12	1	3	3.47
32	tacrine	B	−0.59	−0.90	9.8	9, 16(9.8)	1	3	3.45
33	terbutaline	B	−1.82[d]	−0.89	8.8	4(2), 6, 17	1	1	1.89
34	testosterone	N	3.32	0.00		6(2), 10, 14(0.5)	1	2	2.96
35	tetrahydrocannabinol	N	4.49[e]	0.00		4, 9, 10(3)	1	1	1.43
36	venlafaxine	B	−2.29	−0.89	9.2	12	1	3	3.44
37	xamoterol	B	−0.80[d]	−0.78	7.8[f]	4, 6	1	2	2.85
38	acebutolol	B	−0.99	−0.89	9.2	6, 14	2	3	3.34
39	alprazolam	B	1.10	−0.16	6.2[f]	9	2	3	3.98
40	amitriptyline	B	2.14	−0.90	9.4	9(2), 12	2	2	2.88
41	chlorpromazine	B	2.49	−0.89	9.2	8	2	3	3.07
42	cisapride	B	1.79[e]	−0.75	7.7[f]	11, 16(0.60)	2	3	3.52
43	clemastin	B	1.65	−0.89	9.2[f]	11	2	3	3.52
44	chlorothiazide	A	−0.45	0.57	6.7	5	2	3	3.56
45	desipramine	B	1.20	−0.90	10.2	8(2), 9(2)	2	2	2.26
46	dextropropoxyphene	B	2.20	−0.89	9.2[f]	7	2	2	2.67
47	diltiazem	B	1.47	−0.75	7.7	7, 12	2	2	2.46
48	diprafenone	B	0.59	−0.89	9.2[f]	6, 8, 9, 14, 17	2	2	2.70
49	doxepin	B	1.55[e]	−0.82	8.0	8, 12	2	2	2.89
50	encainide	B	−0.34[e]	−0.90	10.2	8, 9, 11	2	2	2.86
51	ethinyl estradiol	N	3.67	0.00		4, 9	2	2	2.37
52	etilefrine	B	−2.86[d]	−0.90	9.8[f]	4, 6, 12	2	2	2.10
53	famotidine	B	−1.27	−0.52	7.1	5	2	3	3.11
54	fluorouracil	A	−0.89	0.10	8.0		2	4	4.38
55	imipramine	B	1.80	−0.90	9.5	8(2), 9(2)	2	2	2.16
56	indoramine	B	1.56	−0.75	7.7	8, 11	2	2	2.99
57	isoprenaline	B	−2.76	−0.88	8.7	4(2), 6	2	1	1.51
58	isotretinoin	A	4.30	0.90	4.5[f]	10(4)	2	2	2.91
59	labetalol	B	0.35	−0.90	9.3	6, 9	2	3	3.60
60	lidocaine	B	0.88	−0.79	7.9	9, 12(2)	2	2	2.95
61	lorcainide	B	1.44	−0.88	8.7[f]	11	2	3	3.56
62	medifoxamine	B	−0.43[e]	−0.89	9.2[f]	8(2), 12	2	2	2.40
63	metoprolol	B	−1.32	−0.90	9.7	6	2	3	3.80
64	mianserin	B	2.78[e]	−0.52	7.1	8, 9, 12	2	2	2.52
65	midazolam	B	1.37	−0.14	6.1	9	2	3	3.95
66	moricizine	B	2.73	−0.21	6.4	7, 8	2	2	2.19
67	morphine	B	−2.34	−0.90	9.6	4, 6, 11	2	2	2.29
68	nadolol	B	−2.46	−0.90	9.7	6(3), 17	2	3	3.49
69	nafcillin	A	−0.84[e]	0.90	2.7	7	2	3	3.61
70	naltrexone	B	0.84	−0.84	8.1	4, 11	2	2	2.62
71	nicardipine	B	1.18	−0.72	7.6[f]	7, 12, 13, 15	2	1	1.23
72	nicotine	B	−0.34	−0.82	8.0	11	2	3	3.72
73	oxacillin	A	−1.32	0.90	2.8	7, 9	2	3	3.34
74	oxprenolol	B	−0.72	−0.90	9.3	6, 8, 9	2	3	3.01
75	pentazocine	B	1.05	−0.88	8.8	4, 10(2), 11	2	2	2.18
76	pentoxifylline	N	0.29	0.00		14	2	3	3.85
77	phenylephrine	B	−3.02	−0.90	9.8	4, 6, 12	2	2	2.06
78	pimozide	B	5.42	−0.62	7.3	8, 11	2	1	1.68

Table 6 (Continued)

no.	drug	type[a]	log $D_{6.5}$	Δ log D[b]	pK_a	structural descriptors[c]	class obsd	class calcd	score (calcd)
79	pirenzepine	B	−1.51	−0.83	8.1	8, 12	2	2	2.98
80	prenalterol	B	−1.61[d]	−0.89	9.2[f]	4, 6	2	2	2.74
81	promethazine	B	2.21	−0.89	9.1	8(2), 12	2	2	2.13
82	propafenone	B	0.15[e]	−0.89	9.2[f]	6, 8, 9, 14	2	2	2.52
83	propanolol	B	−0.02	−0.90	9.5	6, 8(1.5)	2	2	2.95
84	ritodrine	B	−1.00[d]	−0.89	9.0	4(2), 6	2	1	1.77
85	salbutamol	B	−2.69[d]	−0.90	9.3	6(2), 17	2	3	3.62
86	scopolamine	B	0.11	−0.72	7.6	6, 7(0.5), 11, 12	2	2	2.62
87	spironolactone	N	2.78	0.00		7(2), 10, 14(0.5)	2	1	1.53
88	sulpiride	B	−2.04	−0.89	9.0	5, 11, 12	2	2	2.13
89	thioridazine	B	2.90	−0.90	9.5	8, 11	2	2	2.62
90	triazolam	B	2.26	−0.14	6.2[f]	9	2	3	3.78
91	verapamil	B	1.39	−0.89	8.9		2	3	3.90
92	acetaminophen	N	0.51	0.00		4	3	3	3.30
93	acetylsalicylic acid	A	−1.81	0.90	3.5	7	3	3	3.51
94	amantadine	B	−1.86	−0.90	10.8		3	3	3.91
95	amiodarone	B	2.29	−0.29	6.6	9, 12	3	3	3.30
96	amlodipine	B	0.78	−0.90	9.5[f]	7(2), 13	3	1	1.77
97	atenolol	B	−2.66	−0.90	9.3	6	3	3	3.59
98	atropine	B	−1.37	−0.90	9.7	6, 7(0.5), 11, 12	3	2	2.51
99	bepridil	B	1.10	−0.90	9.3[f]	8, 11	3	3	3.00
100	betamethasone	N	1.94	0.00		6	3	3	3.96
101	bevantolol	B	1.11	−0.87	8.4	6, 9	3	3	3.54
102	brotizolam	B	2.63	−0.14	6.2[f]	9	3	3	3.68
103	bufuralol	B	1.00	−0.89	9.0	6, 9, 17	3	3	3.75
104	captopril	A	−1.78[d]	0.90	3.7	13	3	3	3.45
105	chloramphenicol	N	1.14	0.00		6(2), 15	3	3	3.77
106	chlorpheniramine	B	0.69	−0.89	9.2		3	3	3.99
107	chlorphthalidone	N	0.22[d]	0.00		5	3	3	3.34
108	cimetidine	B	−0.08	−0.38	6.8	9	3	3	3.99
109	clomipramine	B	2.49	−0.89	9.2[f]	8, 9(2)	3	2	2.61
110	clopenthixol	B	2.71	−0.83	8.1[f]	6, 17	3	3	3.67
111	clozapine	B	1.81	−0.81	8.0	8, 12	3	2	2.85
112	codeine	B	−0.28	−0.80	7.9	6, 11	3	3	3.55
113	dexamethasone	N	1.83	0.00		6	3	3	3.98
114	diclofenac	A	2.69	0.89	4.8	8	3	3	3.68
115	dicloxacillin	A	−0.79	0.90	2.8	7, 9	3	3	3.37
116	diphenhydramine	B	0.77	−0.89	9.0	12	3	3	3.57
117	doxazosin	B	1.14[e]	−0.25	6.5	8, 16(6.5)	3	3	3.35
118	enoximone	N	2.32[e]	0.00		9	3	3	3.81
119	ethambutol	B	−4.02	−0.90	9.5	6(2)	3	3	3.24
120	finasteride	N	3.03	0.00			3	3	3.85
121	fluoxetine	B	0.92	−0.90	10.2[f]		3	3	3.96
122	flupenthixol	B	2.88	−0.83	8.1	6, 17	3	3	3.62
123	fluvoxamine	B	−0.13[e]	−0.90	9.5[f]		3	4	4.24
124	furosemide	A	−0.02	0.90	4.7	5	3	3	3.68
125	haloperidol	B	1.42	−0.86	8.3	11, 14	3	3	3.08
126	hydrochlorothiazide	A	−0.19	0.42	7.0	5	3	3	3.50
127	levobunolol	B	−0.40	−0.90	9.3	6, 14, 17	3	3	3.57
128	levomepromazine	B	1.98	−0.89	9.2[f]	8	3	3	3.19
129	maprotiline	B	0.52	−0.90	10.5		3	3	3.99
130	meperidine	B	0.25	−0.88	8.7	7(0.5), 11	3	3	3.15
131	metoclopramide	B	−0.18	−0.90	9.3	12, 16(0.6)	3	3	3.60
132	moclobemide	B	1.26[e]	−0.15	6.2		3	4	4.19
133	nifedipine	N	0.42	0.00		7(2), 13	3	2	2.13
134	nifurtimox	N	0.08	0.00			3	4	4.36
135	nitrazepam	N	2.25	0.00		15, 18	3	4	4.15
136	norethisterone	N	2.97	0.00		10	3	3	3.67
137	nortriptyline	B	1.08	−0.90	9.7	9(2), 12	3	3	3.06
138	omeprazole	B	2.21	−0.02	5.2[f]	9	3	3	3.83
139	ondansetron	B	1.13[e]	−0.61	7.3[f]	9, 14(0.5)	3	3	3.56
140	paroxetine	B	−1.96[e]	−0.90	11.2[f]		3	3	3.90
141	penbutolol	B	1.35	−0.90	9.3	6, 8, 17	3	3	3.34
142	perphenazine	B	2.88	−0.77	7.8	6, 8, 17	3	3	3.05
143	pindolol	B	−0.55	−0.88	8.8	6	3	3	3.85
144	prazosin	B	1.04[e]	−0.25	6.5	16(6.5)	3	3	3.96
145	primaquine	B	−2.19[e]	−0.90	10.3[f]		3	3	3.86
146	procainamide	B	−1.82	−0.89	9.2	12, 16(3.5)	3	3	3.39
147	procyclidine	B	1.30[e]	−0.88	8.8[f]	11	3	3	3.58
148	quinidine	B	1.08	−0.86	8.3	6, 10	3	3	3.58
149	raclopride	B	0.43	−0.89	9.2	11, 12	3	3	3.26
150	ranitidine	B	−1.44	−0.84	8.2		3	3	3.99
151	timolol	B	−0.99	−0.90	9.3	6, 17	3	4	4.04
152	triamterene	B	0.80	−0.15	6.2	16(6.2)	3	4	4.04
153	urapidil	B	0.90	−0.52	7.1	8	3	3	3.50
154	zofenoprilat	A	−0.64[e]	0.90	3.7[f]	13	3	3	3.55
155	allopurinol	N	−0.55	0.00			4	4	4.36
156	amobarbital	A	2.05	0.13	7.8		4	4	4.16

Table 6 (Continued)

no.	drug	type[a]	log $D_{6.5}$	Δ log D[b]	pK_a	structural descriptors[c]	class obsd	class calcd	score (calcd)
157	amrinone	B	−0.72	−0.02	5.2[f]	16(3.0)	4	4	4.24
158	betaxolol	B	−0.07	−0.90	9.4	6	4	3	3.85
159	bisoprolol	B	−0.83	−0.89	9.2[f]	6	4	3	3.84
160	bumetanide	A	−1.84[e]	0.90	4.0	5	4	3	3.57
161	caffeine	N	−0.07	0.00			4	4	4.36
162	carbamazepine	N	2.45	0.00			4	4	4.01
163	carteolol	B	−1.35	−0.89	9.2[f]	6, 17	4	4	4.01
164	chlorambucil	A	2.29	0.82	5.8		4	4	4.12
165	chlordiazepoxide	B	2.43	−0.01	4.8	18	4	4	4.25
166	chloroquine	B	0.33	−0.90	9.9		4	4	4.01
167	chlorpropamide	A	0.76	0.89	5.0		4	4	4.64
168	cibenzoline	B	−0.37[e]	−0.90	10.3		4	4	4.03
169	clobazam	N	0.95	0.00		18	4	4	4.52
170	clonazepam	N	2.41	0.00		15, 18	4	4	4.11
171	clonidine	B	0.06	−0.82	8.0		4	4	4.05
172	cyclophosphamide	N	0.63	0.00			4	4	4.32
173	desmethyldiazepam	N	2.93	0.00		18	4	4	4.11
174	diazepam	N	2.99	0.00		18	4	4	4.09
175	diazoxide	A	1.20	0.03	8.5	9	4	4	4.03
176	diflunisal	A	0.94	0.90	3.0		4	4	4.62
177	disopyramide	B	−1.08	−0.90	10.4		4	4	4.01
178	ethanol	N	−0.31	0.00		6	4	4	4.19
179	ethosuximide	N	−0.33[d]	0.00			4	4	4.36
180	flecaninide	B	0.24	−0.90	9.3		4	4	4.02
181	flurbiprofen	A	1.81	0.90	4.2		4	4	4.49
182	fluconazole	N	−0.11[d]	0.00			4	4	4.36
183	flucytosine	N	−1.65[d]	0.00		16(2.9)	4	4	4.18
184	flunitrazepam	N	2.06	0.00		15, 18	4	4	4.19
185	gemfibrozil	A	1.47[e]	0.89	4.8[f]	9(2)	4	4	4.08
186	glipizide	A	1.31	0.81	5.9	9	4	4	4.32
187	glyburide	A	1.85	0.87	5.3		4	4	4.47
188	hexobarbital	A	1.48	0.04	8.3	10(2)	4	3	3.83
189	ibuprofen	A	2.18	0.88	5.2		4	4	4.40
190	indomethacin	A	2.27	0.90	4.5		4	4	4.39
191	isoniazide	N	−0.70	0.00		16(2.1)	4	4	4.28
192	isosorbide 2-nitrate	N	−0.40	0.00		6	4	4	4.18
193	isosorbide 5-nitrate	N	−0.15	0.00		6	4	4	4.18
194	ketoprofen	A	1.21	0.89	4.6		4	4	4.59
195	ketolorac	A	−1.34[d]	0.90	3.5		4	4	4.64
196	lorazepam	N	2.51	0.00		6, 18	4	4	4.05
197	mabuterol	B	−0.03[e]	−0.90	9.7[f]	6, 17	4	4	4.06
198	methadone	B	1.16	−0.90	9.3	14(0.5)	4	3	3.69
199	methylprednisolone	N	1.66[d]	0.00		6	4	4	4.01
200	metronidazole	N	−0.02	0.00		6, 9, 17	4	4	4.16
201	mexiletine	B	−0.55	−0.89	9.2	9(2)	4	3	3.56
202	naproxen	A	1.04	0.90	4.2		4	4	4.61
203	nitrofurantoin	N	−0.47	0.00			4	4	4.36
204	nizatidine	B	−0.08	−0.38	6.8		4	4	4.22
205	oxaprozin	A	3.41	0.84	5.8[f]		4	4	4.04
206	oxazepam	N	2.24	0.00		6, 18	4	4	4.20
207	phenobarbital	A	1.43	0.21	7.5		4	4	4.30
208	phenylbutazone	A	1.06	0.90	4.4		4	4	4.61
209	phenylethylmalonamide	N	0.13	0.00			4	4	4.35
210	phenytoin	A	2.47	0.10	8.0		4	4	4.05
211	prednisolone	N	1.62	0.00		6	4	4	4.02
212	prednisone	N	1.46	0.00		6, 14(0.5)	4	3	3.80
213	primidone	N	0.91	0.00			4	4	4.30
214	probenecid	A	0.11	0.90	3.4		4	4	4.69
215	protriptyline	B	0.34	−0.84	8.2		4	4	4.03
216	quinine	B	0.34	−0.88	8.8	6, 10	4	3	3.64
217	salicylic acid	A	−1.24	0.90	3.0	8	4	4	4.05
218	sulfadiazine	A	−0.39	0.65	6.5	16(2.3)	4	4	4.52
219	sulfamethoxazole	A	0.11	0.83	5.8	9, 16(2.3)	4	4	4.35
220	sulfinpyrazone	A	−1.40	0.90	2.8		4	4	4.64
221	sulfisoxazole	A	−0.50	0.89	5.0	9, 16(2.3)	4	4	4.37
222	temazepam	N	2.19	0.00		6, 18	4	4	4.13
223	tenoxicam	A	0.56	0.88	5.1[f]		4	4	4.66
224	theophylline	A	−0.02	0.02	8.8		4	4	4.37
225	tocainide	B	−0.56	−0.78	7.8	9(2)	4	3	3.60
226	tolbutamide	A	1.11	0.87	5.3	9	4	4	4.36
227	tolmetin	A	−0.21	0.90	3.5	9	4	4	4.46
228	trazodone	B	1.47	−0.33	6.7		4	4	4.10
229	trimethoprim	B	0.13	−0.57	7.2	16(3.5)	4	4	4.02
230	valproic acid	A	1.04	0.89	4.8		4	4	4.61
231	warfarin	A	1.28	0.88	5.1	14	4	4	4.08
232	zalcitabine	N	−1.33	0.00		6, 16(3.5)	4	4	4.01

[a] A = acid, B = base, N = neutral. Compounds not significantly ionized at pH 6.5 and 7.4 (Δ log D = 0.00) are designated as neutral. [b] log $D_{6.5}$ − log $D_{7.4}$. [c] See Table 5 and the main text. Numbers in parentheses represent scaling factors. [d] Derived from the CLOGP value. [e] Derived from the MLOGP value. [f] Estimated.

adjusts the log *D* of a compound caused by the change in pH following intestinal absorption and it is a very significant contributor to the discrimination. Presumably, partitioning from blood to liver tissues increases with log $D_{7.4}$, resulting in increased compound metabolism. In this regard it has been shown[4] that the induction of P448 in rat liver by miscellaneous compounds in vivo correlated strongly with lipophilicity. Also, several QSAR studies of liver microsomal oxidation demonstrated that the metabolic reactions were significantly dependent on lipophilic character.[4] In addition, acidic and neutral comounds have the advantage of better gastric absorption compared to basic compounds.

Reviewing the individual structural descriptors contained in eq 8, it is apparent that the presence of phenolic OH and SO_2NH_2 groups have a marked and highly significant effect in reducing oral bioavailability, as seen from the weighting vectors and CI values for these descriptors. Phenolic hydroxyl groups are susceptible to transformation in the gut wall and liver and can be conjugated in several ways. Sulfonamide groups are known to be metabolized via *N*-acetylation by *N*-acetyltransferase and can be conjugated. In addition, it has been reported[20] that OH, SO_2NH_2, and $NHCOCH_3$ groups on a benzene ring negatively affected the rate of intestinal absorption in rats. Also, a QSAR study has demonstrated that the absorption rate in rats decreased in relation to the number of phenolic hydroxyl groups present.[21] The negative effects on oral absorption of these groups is consistent with findings that H-bond-donor acidity is a significant factor in decreasing membrane permeability and the absorption of compounds from the gastrointestinal tract.[22,23]

As expected, descriptor 7, covering hydrolytic cleavage (enzymatic or other) of susceptible functions such as esters, lactones, β-lactams, and alkyl carbamates is a strong and highly significant contributor in reducing bioavailability. Alcoholic OH groups (descriptor 6) are numerous in the data set and, as expected, reduce bioavailability, presumably through conjugation and oxidation, although the effect is quite moderate.

Various metabolic carbon oxidative processes are covered by descriptors 8−10, all playing a significant role in reducing bioavailability. Of these, hydroxylation of an activated aromatic ring (descriptor 8) has the largest effect, with aryl methyl and allylic groups (descriptors 9 and 10) contributing to a lesser extent. Descriptor 8 was extended beyond the definition given in the Table 5 footnote to include propranolol and hydralazine, which undergo extensive aromatic ring oxidation.[24] The value of the descriptor for propranolol was estimated as 1.5 on the basis of the reported HOMO levels[25] for aniline (−8.53 eV), 1-aminonaphthalene (−8.11 eV), and benzene (−9.44 eV), which indicate the relative ease of oxidation. The value for hydralazine was similarly estimated as 1.5.

N-Dealkylation is covered by descriptors 11 and 12. As can be seen from the definitions for these in Table 5, the structural requirements are quite limiting, reflecting the fact that this process does not always occur rapidly enough to significantly affect bioavailability. Descriptor 12 for the XCCNMe (or Et) moiety (where X = O,N) appears to capture an electronic effect which facilitates the N-demethylation (or deethylation) process. As far as can be determined, there are no publications describing the electronic effect directly, but two interesting examples were found where this may play a role. The demethylation rates for tertiary methylamines ($R_1N(Me)R_2$) were analyzed, and a QSAR was derived with log *P* and pK_a.[26,27] The amine with R_1 = $C(Me)_2C(=O)Me$ was omitted from the analysis because it was demethylated very rapidly, despite the hydrophilic substituent. Also, in a study of the N-dealkylation of N-alkylamphetamines, replacement of an *N*-ethyl by a 2-cyanoethyl group to give *N*-cyanoethylamphetamine markedly increased N-dealkylation.[28] In the present analysis the descriptor was important to account for the rapid dealkylation of certain compounds, such as lidocaine and ketamine.

Table 7. Classification Results by Category

class	no.	correct	mis(1)[a]	mis(2)[b]	mis(3)[c]
all	232	165	59	8	0
4	78	69	9	0	0
3	63	52	10	1	
2	54	28	25	1	
1	37	16	15	6	0

[a] Misclassified by one category. [b] Misclassified by two categories. [c] Misclassified by three categories.

Other readily oxidized entities, thiols and dihydropyridines, are accounted for by descriptor 13, which has a marked influence on bioavailability, as shown by the high weight vector. Ketones, covered by descriptor 14, are generally metabolically transformed by reduction, and this can have a quite pronounced effect on bioavailability. Aromatic nitro groups (descriptor 15) are also reduced and contribute to the equation but with marginal significance, as denoted by the very low CI number.

Descriptor 16 accounts for the susceptibility of aromatic and heterocyclic amines, hydrazines, hydrazones, and amidines to metabolic acetylation and oxidation, which are scaled by their pK_a values, reflecting their relative reactivity in this process.

The presence of certain types of β-amino alcohol moieties, primarily found in β-adrenergic and some antipsychotic drugs (descriptor 17), results in a positive bioavailability contribution in the model. The origin of this is not clear but may be due in part to a correction of the alcohol descriptor arising from an electronic effect of the amino function. Another positive contributor is descriptor 18 for the presence of the benzodiazepine ring system, which may be correcting for too large a negative effect from the log $D_{6.5}$ term. The contribution index for both of these descriptors in the model is small, and if the descriptors were dropped from the model, the effect on the overall classification results would be minor.

In this analysis, compounds metabolized rapidly through specific metabolic pathways not shared by other compounds in the data set are likely to be misclassified. For example, fluorouracil, which was misclassifed by two ratings, is known to be rapidly reduced to dihydrofluorouracil.[29] Such cases could not be generalized with respect to analysis descriptors, due to the few examples available. In contrast, since many amines were present in the data set, detailed descriptors to discriminate amines could be specified.

Classification Results. A detailed breakdown of the classification results by category is presented in Table

Table 8. Classification Results − Outcome by Compound Prediction[a]

compounds predicted to be in class	no.	correct	mis(1)	mis(2)	mis(3)
all	232	165	59	8	0
4	76	69	6	1	0
3	87	52	29	6	−
2	47	28	19	0	−
1	22	16	5	1	0

[a] See footnotes *a*−*c* of Table 7.

7. Overall, the bioavailability of 71% of the compounds were correctly classified and 97% were correct to within one class. There was a striking difference in classification accuracy depending on how bioavailable the compounds were. Some 88% of class 4 compounds were correct (100% within one class), and a similar result was obtained for class 3 compounds, with 83% correct. Accuracy dropped off for class 2 and class 1 compounds to 52% and 43% correct, respectively. Only in the case of class 1 compounds were a significant number (16%) misclassified by two categories.

The accuracy of the classification of compounds with lower bioavailability will be negatively influenced by unrecognized structural entities leading to rapid metabolism. It is difficult to identify these and include them in the model when representative examples are few. Another factor may be the smaller number of compounds in the lower bioavailability classes since the ORMUCS method sets classifications to maximize the overall number of correct classifications. Thus, there will be a weighting favoring class 4 and 3 compounds.

Of the 67 misclassified compounds, 48 were overpredicted and 19 underpredicted. Only class 2 and 3 compounds can be misclassified in either direction. Considering just these, 21 class 2 compounds were overpredicted and five underpredicted. For class 3, the numbers were six and five, respectively. A bias toward overprediction is to be expected from the factors discussed in the preceding paragraph. In terms of classification errors, it is better in a practical sense for these to be in the direction of overestimation. This reduces the chance that compounds with adequate bioavailability will be overlooked. To illustrate the point, only six compounds were misclassified as belonging to the lowest bioavailability category.

Another way of viewing the results is presented in Table 8. This arranges the data according to the compounds predicted to be in each class as opposed to those actually present in each class. It can be seen that 91% of the compounds predicted to be in class 4 were correctly classified (99% within one class). Together the

Table 9. Test Set Results

							class		
no.	drug	type[a]	log $D_{6.5}$	Δ log D^b	pK_a	structural descriptors[c]	obsd	calcd	score (calcd)
1	budenoside	N	3.20	0.00		6,7	1	2	2.55
2	fenoterol	B	−1.17[d]	−0.90	9.3[f]	4(3), 6, 9	1	1	0.49
3	flumazenil	B	1.01	−0.14	6.2[f]	7, 12	1	2	2.75
4	nivadipine	N	1.99[d]	0.00		7(2), 13, 15	1	1	1.76
5	metaproterenol	B	−2.72[d]	−0.90	9.3	4(2), 6	1	1	1.52
6	rimiterol	B	−3.13[d]	−0.90	9.9[f]	4(2), 6	1	1	1.42
7	terguride	B	2.05[e]	−0.32	6.8[f]	9, 11, 12	1	3	3.35
8	raloxifene	B	1.08[e]	−0.90	9.0[f]	4(2), 11	1	1	1.54
9	ajmaline	B	2.42[e]	−0.25	6.5[f]	6(2), 8, 12	2	2	2.57
10	atovaquone	N	3.96[e]			4	2	2	2.51
11	carvedilol	B	1.31	−0.90	9.3[f]	8(2), 6	2	2	2.61
12	tizanidine	B	−0.68[e]	−0.81	8.0[f]		2	4	4.03
13	rizatriptan	B	−0.53[e]	−0.90	9.0[f]	12	2	3	3.51
14	fluvastatin	A	1.42[e]	0.90	4.5[f]	6(2), 8	2	3	3.68
15	terbinafin	B	3.30[e]	−0.86	8.5[f]	10	2	3	3.25
16	amiloride	B	−3.33[e]	−0.90	8.7	16(2.0)	3	3	3.61
17	bromfenac	A	1.39[e]	0.90	4.3	8, 16(3.0)	3	3	3.86
18	cerevastatin	A	1.06[e]	0.90	4.5[f]	6(2)	3	4	4.26
19	dexfenfluramine	B	0.40[e]	−0.90	10.0	9	3	3	3.77
20	methotrexate	A	−4.55	0.90	3.8	9, 16(6.2)	3	3	3.65
21	mibefradil	B	1.25[e]	−0.90	9.3[f]	9(2), 7	3	2	2.38
22	mirtazapine	B	2.01[e]	−0.55	7.1[f]	9, 12	3	3	3.27
23	olanzapine	B	−0.02[e]	−0.81	8.0[f]	9, 12	3	3	3.41
24	riluzole	B	1.23[e]	−0.25	6.5[f]	16(6.5)	3	3	3.94
25	risperidone	B	0.74[e]	−0.90	8.7[f]	11	3	3	3.64
26	tramodol	B	0.85	−0.90	9.4		3	3	3.97
27	zidovudine	N	0.05			6, 10	3	3	3.98
28	citalopram	B	0.45[e]	−0.90	9.2		4	4	4.00
29	dofetilide	B	0.44	−0.52	7.0	12	4	3	3.74
30	etodolac	A	1.96	0.90	4.7	8, 9(2)	4	3	3.39
31	guanafacine	B	0.46[e]	−0.90	8.7		4	4	4.00
32	lamotrigine	B	1.91[e]	−0.05	5.7	16(5.7)	4	3	3.93
33	lansoprazole	N	1.73[e]				4	4	4.17
34	levonorgestrel	N	3.87[e]			10, 14	4	3	3.12
35	milrinone	B	0.68[e]	−0.02	5.2[f]	9	4	4	4.08
36	nevirapine	N	1.92			9	4	3	3.90
37	pirmenol	B	−0.53[e]	−0.90	10.2[f]		4	4	4.03
38	pramipexole	B	−2.90[e]	−0.90	11.0[f]	9, 16(4.5)	4	3	3.35
39	rilmenidine	B	−0.94[e]	−0.90	9.3[f]		4	4	4.01
40	terazosin	B	0.29	−0.52	7.1	16(7.1)	4	3	3.91

[a−f] See Table 6.

results from Tables 7 and 8 indicate a high degree of reliability for compound predictions in the ≥80% bioavailability range. In the case of class 3 compounds, 87 were predicted to belong to this class, with 52 correctly classified, compared to 63 actually present resulting in a lower correct classification rate of 60% (93% ± one class) in contrast to the 83% from Table 7. The implication here is that a high proportion of actual class 3 compounds will be identified but a significant number of others will be incorrectly assigned to this class. Those compounds predicted to be class 2 compounds were 60% correct (100% ± one class), a somewhat higher percentage than in Table 7, due to the lower number predicted to be in this category. Only 22 of 37 actual compounds were placed by the model in class 1. Of these, a high proportion (73%) was correct. This suggests that the class 1 predictions made by the model would be reasonably accurate but that a significant proportion of compounds in the ≤20% category would be overestimated.

Test Set Results. To evaluate the performance of the QSAR model, a separate compound test set of 40 additional compounds was subsequently assembled (Table 9) and the bioavailability classifications of these were calculated manually using eq 8 of Table 5. This gave 24 compounds (60%) correctly classified with 38 (95%) correct to within one class. As in the original data set, performance was somewhat better with classes 4 and 3 (64% correct, 96% ± one class) than with the lower bioavailability categories 2 and 1 (53% correct, 87% ± one class). Thus, although there was some loss of accuracy for the higher bioavailability categories, results from the test compounds were generally in line with those from the model data set.

Limitations. The QSAR model would be expected to perform well for predicting the bioavailability of compounds within the general universe of structural types represented in the data set from which the model was derived. Indeed, this was demonstrated using the test set compounds. However, for compounds outside of this realm (see listing of these types in the earlier section on bioavailabilty), this would not be the case. Also, high molecular weight compounds (>500) and those with strong hydrogen-bonding capability are known to have reduced absorption potential, and these types of compounds do not have significant representation in the data set. Such compounds can be identified as problematic for bioavailability using criteria described by Lipinsky[18] in his "rule of five". Peptides and peptide-like compounds, which are rare as orally active drugs, are not in the data set, so the model is not designed to handle these. The model assumes absorption by passive diffusion and so would not be appropriate for the small minority of compounds absorbed through other mechanisms.

Structural descriptors in the model have a single averaged weighting value for all of the compounds containing that particular feature, but it is clear that for certain compounds the actual effect might represent a wide divergence from this average value. In addition, structural factors affecting the bioavailabilty of particular compounds through rapid metabolism may not have been included in the model because too few examples were present in the data set.

Conclusions

We believe that the feasibility of constructing a QSAR model for predicting the approximate human oral bioavailabilty of prospective new medicinal agents has been clearly demonstrated. The model, developed from human oral bioavailability data on a diverse set of 232 drugs, performed well, and a significant level of discrimination in bioavailabilty level was achieved despite the inherent complexity involved. The model can be readily employed for so-called "drug-like" compounds commonly worked with in drug discovery projects. Animal data, which does not always correlate with human data, is not utilized, and predictions do not require experimental data and thus can be made for unsynthesized compounds. Predictions of compound human oral bioavailability do not require the use of the ORMUCS program, which is only employed for the development of the model. They can be easily made using a hand calculator (or through a simple calculation program) of the compound score using the model equation and the relevant descriptors for that compound. The model has the advantage of transparency in that it indicates which factors may affect bioavailability and the extent of that effect, thus providing the basis for designing improved compounds. Refinement of the model is possible with the availability and incorporation of more compound bioavailability data, and future work along these lines is planned. The model could be used in drug discovery projects as an input for decision making and priority setting concerning new compounds to be synthesized and the selection of existing compounds for detailed workup.

Acknowledgment. We thank Dr. Takahashi and Dr. Sasaki of the Toyohashi University of Technology for providing the ORMUCS program and Mr. Kuono for converting the program for use on the Power Macintosh 7100/80AV. The authors are indebted to the KI Chemical Research Institute Co., Ltd., Shizuoka, Japan, for financial support.

References

(1) Presented in part at the Corwin Hansch Award Symposium, Division of Computers in Chemistry, ACS National Meeting, Anaheim, CA, March 21−25, 1999.
(2) Austel, V.; Kutter, E. In *Quantitative Structure−Activity Relationships of Drugs*; Topliss J. G., Ed.; Academic Press: New York, 1983; pp 437−496.
(3) Lien, E. J. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1975; Vol V, Chapter 3.
(4) Hansch, C.; Leo, A.; Hoekman, D., Eds.: *Exploring QSAR−Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995; Chapter 8.
(5) Hirono, S,; Nakagome, I.; Hirano, H.; Matsushita, Y.; Yoshi, F.; Moriguchi, I. Noncongeneric structure−pharmacokinetic property correlation studies using fuzzy adaptive least-squares: Oral bioavailability. *Biol. Pharm. Bull.* **1994**, *17* (2), 306.
(6) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches: Methods and Principles in Medicinal Chemistry*; Mannhold, R., Krogsgaard-Larson, P., Timmerman, H., Eds.; VCH: Weinheim, 1993; Vol 1, pp 100−101.
(7) (a) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 9th ed.; Hardman, J. G., Limbird, L. E., Molinoff, P. B., Ruddon, R. W., Gilman, A. G., Eds.; McGraw-Hill: New York, 1996: pp 1712−1792. (b) *Handbook of Clinical Pharmacology*; Bochner, F., Carruthers, G., Kampman, J., Steiner, J., Azarnoff, D. L., Eds.; Little, Brown and Company: Boston/Toronto 1983. (c) *Clin. Pharmacokinet.* **1990**−**1994**, 18−25. (d) *Handbook of Clinical Pharmacokinetic Data*; Jack, D. B., Ed.; Macmillan: New York, 1992. (e) *Physicians Desk Reference*, 53rd ed., Medical Economics Co.: 1999.

(8) Hansch, C.; Leo, A.; Hoekman, D. *Exploring QSAR—Hydrophobic, Electronic and Steric Constants*; American Chemical Society: Washington, DC, 1995.
(9) CLOGP program, ver. 1.0.0, Biobyte Corp.
(10) Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40* (1), 127−130.
(11) Scherrer, R. A.; Howard, S. M. Use of distribution coefficients in quantitative structure−activity relationships. *J. Med. Chem.* **1977**, *20*, 53−58.
(12) (a) *Wilson and Grisvold's Textbook of Organic Medicinal and Pharmaceutical Chemistry*, 9th ed.; Delgado, J. N., Remers, W. A., Eds.; J. B. Lipincott: Philadephia, 1991; pp 877−881. (b) *AHFS 94 Drug Information*; American Society of Hospital Pharmacists, Inc.: 1994. (c) *Principles of Medicinal Chemistry*, 4th ed.; Foye, W. O., Lemke, T. L., Williams, D. A., Eds.; Williams and Wilkins: Baltimore, 1995; pp 948−960. (d) *The Merck Index*, 12th ed., Budavari, S., Ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 1996. (e) *Physicians Desk Reference*, 52−54nd eds.; Medical Economics Co.: 1998−2000.
(13) Dressman, J. B.; Amidon, G. L.; Fleisher, D. J. Absorption potential: Estimating the fraction absorbed for orally administered compounds. *J. Pharm. Sci.* **1985**, *74*, 588−589.
(14) Takahashi, Y.; Miyashita, Y.; Abe, H.; Sasaki, S. A new approach for ordered multicategorical classification using simplex technique. *Bunseki Kagaku* **1984**, *33*, E487-E494.
(15) Sasaki, S.; Takahashi, Y. Application of pattern recognition techniques to assessment of hazards of chemical substances. *Hazard Asses. Chem.* **1986**, *5*, 61−101.
(16) Takahashi, Y.; Sasaki, S. Application of the ORMUCS method to structure−activity studies on the fungicidal activity of mepronil derivatives. *Quant. Struct.-Act. Relat.* **1987**, *6*, 17−21.
(17) Lien, E. J. Structure−activity relationships and drug disposition. *Ann. Rev. Pharmacol. Toxicol.* **1981**, *21*, 31−61.
(18) Lipinsky, 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 Deliv. Rev.* **1997**, *23*, 3−25.
(19) Zimm, S.; Collins, J. M.; O'Neill, D.; Chabner, B. A.; Poplack, D. G. Inhibition of first-pass metabolism in cancer chemotherapy: Interaction of 6-mercaptopurine and allopurinol. *Clin. Pharmacol. Ther.* **1983**, *34*, 810−817.
(20) Nogami, H.; Hanano, M.; Yamada, H. Studies on absorption and excretion of drugs. X. Relation between chemical structure and absorption rate. (2). Substituent constant for absorption rate coefficient of foreign organic compounds.*Chem. Pharm. Bull.* **1968**, *16*, 580−585.
(21) Tai, A. W.; Lien, E. J. Structure−activity relationship in membrane transport and drug absorption. *Acta Pharm. Jugoslav.* **1980**, *30*, 171−182.
(22) Burton, P. S.; Conradi, R. A.; Hilgers, A. R. (B) Mechanisms of peptide and protein absorption. (2) Transcellular mechanisms of peptide and protein absorption: Passive aspects. *Adv. Drug Del. Rev.* **1991**, *7*, 365−386.
(23) von Geldern, T. W.; Hoffman, D. J.; Kester, J. A.; Nellans, H. N.; Dayton, B. W.; Calzadilla, S. V.; Marsh, K. C.; Hernandez, L.; Chiou, W.; Dixon, D. B.; Wu-Wong, J. R.; Opgenorth, T. J. Azole endothelin antagonists. 3. using Δ log *P* as a tool to improve absorption. *J. Med. Chem.* **1996**, *39*, 982−991.
(24) (a) Walle, T.; Gaffney, T. E. Propanolol metabolism in man and dog: Mass spectrometric identification of six new metabolites. *J. Pharmacol. Exptl. Ther.* **1972**, *182*, 83−92. (b) Reference 12c, p 136.
(25) Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test. 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.***1992**, *19*, 37−52.
(26) Hansch, C. Quantitative relationships between lipophilic character and drug metabolism. *Drug Metab. Rev.* **1972**, *1*, 1−13.
(27) Hansch, C.; Steward, A. R.; Iwasa, J. The use of substituent constants in the correlation of demethylation rates. *J. Med. Chem.* **1965**, *8*, 868−870.
(28) Beckett, A. H.; Shenoy, E. V. B.; Salmon, J. A. The influence of replacement of the *N*-ethyl group by the cyanoethyl group on the absorption, distribution and metabolism of (±)-ethylamphetamine. *J. Pharm. Pharmacol.* **1972**, *24*, 194−202.
(29) Reference 7a, Chapter 51, p 1250.

JM0000564