Genetic Algorithm-Optimized QSPR Models for Bioavailability, Protein Binding, and Urinary Excretion

Junmei Wang,*,† George Krudy,† Xiang-Qun Xie,§ Chengde Wu,† and George Holland†

Encysive Pharmaceuticals Inc., 7000 Fannin Street, Houston, Texas 77030, and Department of Pharmaceutical & Pharmacological Sciences, College of Pharmacy, University of Houston, 4800 Calhoun Road, Houston, Texas 77204-5037

Received March 15, 2006

In this work, a genetic algorithm (GA) was applied to build up a set of QSPR (quantitative structure−property relationship) models for human absolute oral bioavailability, plasma protein binding, and urinary excretion using the counts of molecular fragments as descriptors. For a pharmacokinetic property, the consensus score of a set of models (20 or 30) was found to improve the correlation coefficient and reduce the standard error significantly. Key fragments that may boost or reduce pharmacokinetic properties were also identified. Databases searches were performed for a set of key fragments identified by bioavailability models. The percentage of hit rates of bioavailability-boosting fragments were significantly higher than those of bioavailability-reducing fragments for MDDR (MDL Drug Data Report), a database of drugs and drug leads entered or entering development. On the other hand, the opposite trend was observed for ACD (Available Chemicals Directory), a database of all kinds of available compounds.

INTRODUCTION

Drug discovery is a very complicated and costly procedure. It is said that 95% of lead compounds fail in the development stages, and half of those failures happen during the clinical trials. Given the fact that only 15% of the drug development cost is preclinical, the failures in clinical trials, therefore, are very costly. About 50% of the failures in clinical trials are the result of poor ADMET (absorption, distribution, metabolism, excretion, and toxicity) and PK (pharmacokinetics) properties.[1,2] It has been gradually realized that a "parallel model" of drug discovery, which optimizes drug potencies, selectivity, and ADMET/PK properties simultaneously, is superior to the traditional serial-cyclical model.[3] Therefore, it is wise to exploit developable drug leads, which have not only high potencies but also good selectivity and ADMET/PK properties, prior to clinical trials in drug discovery. A widely used druglike filter in drug discovery is Lipinski's "rule-of-five", which states that a good drug candidate should have a molecular weight smaller than 500, a calculated log *P* smaller than 5.0, and numbers of hydrogen bond donors and acceptors less than 5 and 10, respectively.[4] It was reported that only four of the top 100 best selling drugs fell outside the rule-of-five in 1998.[3] However, the rule-of-five is only the minimum criterion of a molecule to be druglike. It is very easy for a compound to fall within the rule-of-five but with no potential to lead to a drug. As a matter of fact, 68.7% of the compounds in ACDSD (ACD Screening Database, 2.4 million compounds) and 55% of the compounds in ACD (240 thousand compounds) have no violation of the rule-of-five at all. Therefore, more stringent criteria should be built up to discriminate druglike compounds from the others. Reliable in silico models that predict ADMET/PK properties on a computer can be used as a part of the criteria. There are now a set of programs/software packages available to make predictions on some "physical" ADMET/PK properties, exampled by aqueous solubility, water−oil partition coefficient (log *P*), absorption potential, permeability,[5−8] etc. But there are many fewer mature models for "physiological" ADMET/PK properties, such as oral bioavailability, plasma protein binding, urinary excretion, area under the plasma concentration−time curve (AUC), total body clearance (Cl), volume of distribution, elimination half time ($t_{1/2}$), etc. The main reason is that biological and physiological ADMET/PK properties are typically affected by many factors. For example, the oral bioavailability of a drug is usually less than 100%. Various physiological factors reduce the availability of drugs prior to their entry into the systemic circulation; those factors may include, but are not limited to, poor absorption from the gastrointestinal tract, degradation or metabolism of the drug prior to absorption, and hepatic first pass effect. In addition, each of these factors may vary from patient to patient and some drugs have multiple circulating active forms. As a consequence, it is difficult to build highly predictable models for those properties. The severity is further exaggerated by the lack of high-quality experimental data.

Oral bioavailability is one of the most important pharmacokinetic properties in drug discovery. It represents the percentage of an oral dose that is available to produce pharmacological actions, in other words, the fraction of the oral dose that reaches the arterial blood in an active form. Among the several reasons that lead to a true decrease in bioavailability, drug dissolution and gastrointestinal permeability, which control the rate and extent of drug absorption, are the two most important factors. In general, a drug with high solubility and high membrane permeability is considered

* Corresponding author phone: 713-578-6649; fax: 713-578-6720; e-mail: jwang@encysive.com.
† Encysive Pharmaceuticals Inc.
§ University of Houston.

10.1021/ci060087t CCC: $33.50 © 2006 American Chemical Society
Published on Web 10/06/2006

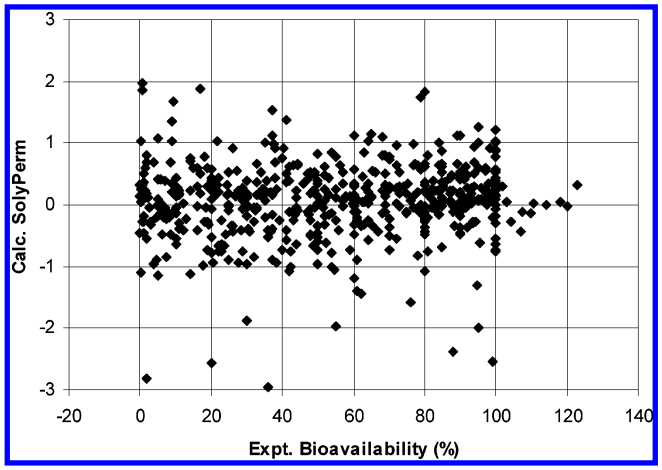


Figure 1. Experimental human bioavailability vs calculated solubility−permeability (SolyPerm) parameters by Volsurf.

Table 1. Experimental Errors of Oral Bioavailability, Plasma Protein Binding, and Urinary Excretion

	no. of data	av unsigned error (%)	RMS error
bioavailability	367	12.1	14.5
plasma protein binding	266	7.8	9.9
urinary excretion	224	3.9	5.6

to be practically exempt from bioavailability problems; a drug having low solubility yet high permeability or high solubility yet low permeability requires careful formulation to improve its dissolution rate, and finally, a drug with poor solubility and permeability is a problematic candidate for administration. In VolSurf,[5] an in silico ADMET software package implemented in Sybyl[9] of Tripos Inc., a hybrid parameter (SolyPerm) that combines the scores of the thermodynamic solubility model and the CACO2 permeation model is applied to classify drug candidates on the basis of both solubility and permeability. The SolyPerm model was constructed using 1833 training-set molecules with PLS (partial least-squares) analysis. The more positive the parameter is, the greater the chance that the compound is bioavailable. The experimental bioavailability of 577 compounds (see next section) versus the calculated SolyPerm parameter with Volsurf is plotted in Figure 1. Unfortunately, there is no obvious correlation between the experimental bioavailability and the SolyPerm parameter. This indicated that more complicated descriptors may be required to build up the relationship.

Plasma protein binding is another important pharmacokinetic property in drug discovery. It is expressed as the percentage of a drug in the plasma that is bound to plasma proteins at the concentrations of the drug that are achieved clinically. Like bioavailability, there are multiple factors that contribute to the experimental errors (about 10%, see Table 1), including the analysis and instruments, as well as the experimental conditions, in addition to the disease states that alter the concentration of albumin or other proteins in plasma that bind drugs. In Volsulf, an in silico QSAR model of predicting binding affinity to HAS (human serum albumin) was constructed with PLS using 408 literature compounds in the training set. The cross validation suggested two significant latent variables (two components) linearly constructed from 94 Volsulf variables can explain most of variances. This PLS model was applied to predict the protein

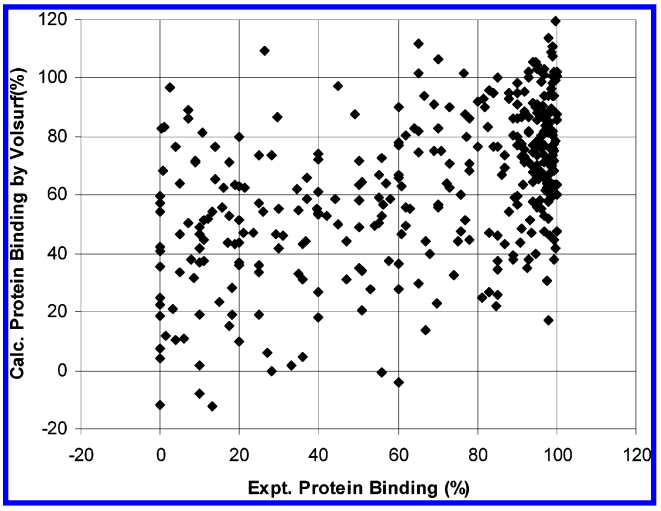


Figure 2. Experimental vs calculated plasma protein binding by Volsurf.

binding of 404 compounds collected by ourselves (see the next section), and the experimental and calculated values were plotted in Figure 2. It is shown that there are still a couple of false positives and negatives, albeit many compounds (especially those having high protein-binding values (>80%)) are correctly predicted.

As soon as drugs are absorbed, the body starts to dispose of them. The elimination pathways depend on the drug and the pathophysiology of the disease. One or more processes such as renal excretion, biliary excretion, and metabolism to either active or inactive metabolites may be involved. The urinary excretion of an unchanged drug is expressed as a percentage of the administered dose eventually excreted unchanged in the urine. This parameter reflects the loss of drug during the procedure of drug administration, and it may be used to assess the relative bioavailability of drugs.

The quantitative structural property (including biological activity) relationship (QSPR) is one of the main means in rational drug design, especially when the structures of receptors are not available. The choice of proper descriptors is the key of a successful QSPR model. There are a great number of descriptors available for use in QSPR studies, ranging from simple molecular weight to molecular surface area to the grid interaction energies between a probe and target molecules.[10] It is relatively easy to build up a predictable QSPR model when the mechanism behind it is simple. However, it is much harder to generate a QSPR model that describes a multiple-mechanism physiological ADMET/PK property only with those commonly used descriptors. In this paper, the counts of molecular fragments were applied as descriptors to model three important ADMET/PK properties, namely, human oral bioavailability, plasma protein binding, and urinary excretion.

The mathematic algorithm applied to build up the correlations between a molecular property and descriptors is the other part of the story. The widely used approaches are regression, partial least-squares fitting, neural networks, etc. Recently, genetic algorithm, an efficient stochastic optimization method, has been widely used by chemists to solve minimization problems such as conformational searches and molecular docking, as well as QSPR model generation.[11−15] GA minimizes a target function through three operations that mimic natural evolution and selection, namely, mutation,

crossover, and selection. First of all, a set of "chromosomes", which may be possible answers to the question, is randomly generated. The "genes" in a chromosome correspond to the variables in question. For each chromosome, the fitness is evaluated by a scoring function. The higher the score, the better the fitness and the closer to the real answer it is. New chromosomes are then generated through crossover and mutation operations. In the subsequent selection operation, chromosomes with high fitness evolve to the next generation, and those having low fitness are allowed to perish. The three operations are iteratively performed until the termination criterion is met. The power of GA lies in its ability to efficiently deal with multiple dimension problems, regardless of the variables being coupled.

In this work, GA has been extensively applied to select key molecular fragments that contribute more to the modeled pharmacokinetic properties. Unlike other statistical methods, GA can build up multiple models that have similar performance. The analysis on the descriptors of a set of models may provide deeper insight into the relative importance of individual descriptors. Moreover, the consensus score of multiple models should be much more reliable than that of any single model in predicting the modeled property for molecules outside the training set.

MATERIALS AND METHODS

1. Experimental Data Source. There were three sources for the pharmacokinetic properties (human oral bioavailability, plasma protein binding, and urinary excretion) studied in this work, namely, the Goodman and Gilman's the Pharmacological Basis of Therapeutics (9th and 10th eds.),[16,17] the Integrity Databases of Prous Science, and some human oral bioavailability data from ref 18. Most of the compounds under study were actual marketed drugs. All the duplicated entries were dropped off for all the three PK properties. To efficiently purge the duplicated entries, the unique sybyl line notations (SLNs)[9] that have a sole expression for a molecule regardless of the sequence order of atoms, were used. When duplications occurred, the priority for the application of the experimental data for the three sources is as follows: the Goodman and Gilman's the Pharmacological Basis of Therapeutics, followed by the Integrity Databases of Porous Science, and after that, ref 18.

2. Descriptors in QSPR Models. Similar to water−oil partition coefficients (log *P*), which can be estimated by the addition of each molecular fragment, atoms/bond, or both contributions, a set of molecular fragments were employed to build up the QSAR models for the three pharmacokinetic properties. For each PK property, the molecular fragments came from three resources. First of all, a cluster analysis was performed on the top 30% of the training set molecules to find common structural fragments that can boost the property in question. In addition, another cluster analysis was performed on the bottom 30% of training set molecules to identify common structural fragments that decrease the values of the property. The cluster analysis intended to identify common substructures for a series of compounds which are represented in a dendrogram. Each node of the dendrogram corresponds to a substructure and the set of compounds containing the substructure. The children branches always contain the common structures of their parents in the upper levels. These two kinds of fragments (boosting and decreasing a property) compose the first part of fragment library to be selected by GA. Second, the rings (both aromatic and aliphatic, up to 12 atoms) and chains (4 to 8 atoms) existing in the training set molecules are the second source of the library. Finally, some simple functional groups, such as hydroxyl, carbonyl, amide, amine, and cyclopropane, etc., compose the third source of the library.

In addition to the fragment descriptors, the following molecular properties, which are descriptors in many QSPR models, were selected as additional descriptors to model the three PK properties: molecular weight, number of atoms, calculated log *P* (clog *P*), calculated molecular refractivity (cmr, serving as a measure of the binding force between the polar portions of an enzyme and its substrate), the number of hydrogen-bond donors, the number of hydrogen-bond acceptors, the number of hydrophobic fragments, the rule-of-five (the score is 4 if all the four conditions are met and is decreased by 1 for each violation). All those descriptors were calculated either with the built-in commands in Sybyl, version 6.9, of Tripos Inc[9] or SPL (Sybyl program language) scripts developed in our group.

3. QSPR Model Generation. A real value-encoded genetic algorithm developed by our group was applied to select 80 and 70 descriptors for oral bioavailability and urinary excretion and for plasma protein binding, respectively. The ratios of the number of data to the number of descriptors were kept at least 6 to avoid over-fittings. The fitness score of each chromosome in GA was the square of the regression coefficient (R^2). The important parameters that controlled the GA performance were listed as the following: *population size*, the number of chromosomes in one generation (100); *chromosome size*, the number of variables in question (80 and 70 for this work); *elite size*, the number of "elite" chromosomes, which entered the next generation directly; (5) *mutation probability*, the probability of performing mutation on each gene of each chromosome (0.05); *crossover probability*, the probability of performing crossover on each chromosome in a population (0.80); *selection methods*, roulette-wheel selection, rank selection, and tournament selection; and *maximum iteration*, the maximum iteration of the optimization (1000). For each parameter, the value in parenthesis was the one applied in this work. It is notable that the selection methods were different at different GA optimization stages: for the first 300 iterations, rank selection was used; then for the next 400 iterations, tournament selection was applied; finally, for the last 300 iterations, roulette-wheel selection was used. GA optimizations were performed 20 times for human bioavailability and urinary excretion and 30 times, for plasma protein binding. Therefore, there were 20, 30, and 20 models generated for the three properties, respectively. The details on how the GA optimization is performed are beyond the scope of this paper.

4. Model Validation and Consensus Score Calculations. For each GA-optimized QSPR model, a 90/10 (90% data in the training set and 10% data in the test set) cross-validation was carried out 1000 times. For each run, 90% of molecules were randomly selected to enter the training set and the others in the test set. Then a regression model was generated with the training set molecules and evaluated by the test set molecules. The average of the square of the regression coefficients for the test set (q^2), which measures the

predictability for a set of GA QSPR models, was then calculated. The consensus score of a molecule was simply defined as the average of the values predicted by a set of GA-QSPR models (20 models for human oral bioavailability and urine excretion and 30 for plasma protein binding). It is expected that a consensus score of a molecule outside the training set is much more reliable than that predicted by any individual model.

5. Hologram QSPR Models. In addition to the GA-QSPR models, a hologram QSPR model for each PK property was also generated using the HQSAR module in Sybyl, version 6.9. HQSAR applies one type of 2D fingerprints, the so-called molecular hologram,[19] as molecular descriptors to build up QSPR models. Details on the theory of HQSAR can be found in the HQSAR manual of Sybyl, version 6.9.[9] The default setting was applied for model generation in this work.

6. Database Searches. Database searches were carried out for a set of key fragments which make substantial contributions to bioavailability. Two MDL[20] databases, MDDR (MDL Drug Data Report) and ACD (Available Chemicals Directory) were searched. The analysis of the hit rates of the two databases can validate the QSAR model to some degree.

RESULTS AND DISCUSSION

1. Performance of the QSPR Models. After the elimination of the duplicated molecules, human oral bioavailability, plasma protein binding, and urinary excretion had 577, 404, and 581 entries, and the mean values are 54.9, 64.9, and 24.9%, respectively. Table 1 lists the statistic results of experimental errors for the three PK properties. The average unsigned error (AUE) and root-mean-square errors (RMSE) were 12.1 and 14.5% for human oral bioavailability, respectively; the AUE and RMSE were 7.8 and 9.9% for plasma protein binding, respectively, and the AUE and RMSE were 3.9 and 5.6% for urinary excretion, respectively. The large experimental errors of these pharmacokinetic properties reflect the fact that the measured PK parameters are affected by multiple factors and partially explain why these properties are difficult to model.

In total, 1249, 941, and 1260 descriptors were generated for human oral bioavailability, plasma protein binding, and urinary excretion, respectively. Except eight descriptors, which were widely used in druglike analysis, the other descriptors were counts of a set of fragments that appeared in the training set molecules. Table 2 lists the performance of QSPR models for the three PK properties. Computational models were constructed by optimizing a set of weights, which measured the relative importance of each fragment type, to reproduce the property in question for the training set molecules. Certainly, the larger the training set results in a more reliable model. Unlike the water−oil partition coefficient, which has up to 10 thousand experimental data, the experimental data for the three PK properties are quite few. To build up valid QSPR models without over-fitting, only a limited number of fragments (n/5, where n is number of data) can be considered for use. Therefore, a smart program should be used to pick up those fragments that make substantial contributions to the property in a study from a big fragment library. Fragment-based descriptors were also used by Andrews et al. to predict human oral bioavailability with a different molecular set.[8] Comparisons were made on the performance of the GA-QSPR models to that of a HQSAR model. The details were presented below.

Table 2. A Summary of GA-QSPR Models, a Consensus Score Model and a HQSAR Model for Human Oral Bioavailability, Plasma Protein Binding and Urinary Excretion

	HQSAR	GA-QSPR	GA-QSPR (consensus)
bioavailability			
no. of models[a]	1	20	1
no. of data	577	577	577
no. of components[b]	6	80	80
mean R^2	0.35	0.55	0.62
mean RMSE (%)	26.4	21.9	20.2
mean cross-validated R^2[c]	0.29	0.42	
mean RMSE for cross validation (%)	27.4	24.6	
plasma protein binding			
no. of models[a]	1	30	1
no. of data	404	404	404
no. of components[b]	5	70	70
mean R^2	0.46	0.82	0.86
mean RMSE (%)	25.4	14.7	13.0
mean cross-validated R^2[c]	0.22	0.66	
mean RMSE for cross validation (%)	30.46	19.5	
urinary excretion			
no. of models[a]	1	20	1
no. of data	581	581	581
no. of components[b]	6	80	80
mean R^2	0.37	0.60	0.65
mean RMSE (%)	23.3	18.3	17.2
mean cross-validated R^2[c]	0.33	0.43	
mean RMSE for cross validation (%)	23.93	21.8	

[a] For a particular pharmacokinetic property, each model has the same data set and the same number of components. [b] For HQSAR, it is the number of principle components in partial least-squares fitting, while for GA-QSPR, it is the number of descriptors in regression analysis. [c] For HQSAR, the cross validation was leave-one-out, while for GA-QSPR, it was 1000 times the 90/10 training and test set split.

For human oral bioavailability, the square of regression coefficient (R^2) and root-mean-square error (RMSE) for the HQSAR model were 0.35 and 26.4%, respectively. In contrast, the performance of 20 GA-QSPR models was significantly better than that of HQSAR: the mean R^2 and mean RMSE values for the 20 models were 0.55 and 21.9%, respectively. Encouragingly, the use of consensus score further improved the performance, and the R^2 and RMSE were 0.62 and 20.2% for the consensus score models, respectively. Since the standard error of the experimental data is 14.5%, this performance is at least acceptable. For the 90/10 cross-validation analysis, the mean R^2 and mean RMSE values for the test sets were 0.42 and 24.6%, respectively. In comparison to the Andrews et al. model, we studied a much larger dataset and tested many more molecular fragments, as well as eight commonly used descriptors, in the QSPR models. Interestingly, many key fragments were identified by both models, such as Frag28 in Table 3.

For plasma protein binding, the R^2 or mean R^2 values were 0.46, 0.82, and 0.86 for the HQSAR, GA-QSPR, and GA-QSPR consensus models, respectively. The RMSE or mean RMSE were 25.4, 14.7, and 13.0%, respectively. This is also

Table 3. List of Key Molecular Fragments and Molecular Properties that Reduce or Boost Bioavailability[a]

	nm	c	v	nf	$c \times v$	sln
Frag1	17	−31.9	1.9	15	−59.6	C[1](C[2](C(C[4]C(C(C(=CCC)CC@4)C)CC@2)CC@1)C)C
Frag2	20	−49.1	1.0	5	−49.1	O[1]CCCC=C@1
Frag3	16	−47.8	1.0	37	−47.8	N[1]CCC@1
CMR	6	−4.2	9.7	577	−40.6	
Frag4	19	−37.9	1.0	6	−37.9	S[1]C[2]=C(NC[5]=C@1C=CC=C@5)C=CC=C@2
Frag5	8	−36.6	1.0	17	−36.6	C[1]=CCC=CC@1
Frag6	18	−32.2	1.0	5	−32.2	N[1]CCCCCC@1
Frag7	9	−8.2	3.8	24	−31.0	N(CC)C(=O)NC
Frag8	20	−28.7	1.0	7	−28.7	SHR1
Frag9	11	−12.9	2.2	40	−27.7	N(CO)C
Frag10	6	−17.7	1.4	15	−24.7	N(CN)C(=CC)C
Frag11	10	−20.3	1.2	15	−24.4	O=P[TAC=4]
Frag12	14	−20.6	1.1	21	−22.6	N(C(C)(C)C)C
Frag13	9	−10.8	1.8	49	−18.9	N(=C(C)C=CC)C
Frag14	15	−7.2	2.2	102	−15.6	NC=C(CC)C
Frag15	11	−2.9	5.1	108	−14.7	O(CC)CCCC
Frag16	7	−5.8	2.2	48	−12.6	O(CC)C(=O)CC
Frag17	11	−2.4	5.1	165	−12.2	C(CCCC)C=C
Frag18	6	−10.3	1.1	23	−11.7	C(=O)NH2
Frag19	7	5.0	2.0	74	10.1	O=CCC(C)C
Frag20	11	5.5	1.9	163	10.3	NC(CC)=O
Frag21	6	4.1	2.6	52	10.4	N(C(C(C)C)C)C
Frag22	6	7.4	1.5	78	11.3	C[1](C=CC=CC=@1)OCC
Frag23	9	7.2	2.1	39	14.7	N(C(=N)C)C
Frag24	6	0.5	29.6	138	15.5	C(CCC)(CC)C
Frag25	6	7.7	2.0	117	15.7	CCNC[4]=CC=CC=C@4
Frag26	13	16.6	1.1	17	17.5	N(CC(O)C)C=C
Frag27	9	11.6	1.6	38	18.0	O(COC)C
Frag28	16	17.1	1.1	22	18.6	C[1]CC@1
Frag29	8	10.9	1.7	11	18.8	O(CC)CC(=C)O
Frag30	7	14.4	1.4	23	20.1	CC(NC=O)=O
Frag31	6	20.9	1.0	13	20.9	C[1]=CC=CC=C@1CCCC@1
Frag32	10	20.1	1.1	14	21.5	O(C(CO)(C)C)C
Frag33	17	20.2	1.1	15	21.6	C[1]=CC=NC(=N@1)N
Frag34	17	27.8	1.0	7	27.8	N[1]=CC=CC[5]=C@1C=CC=C@5
Frag35	7	27.8	1.0	10	27.8	S(=O)(=O)(NCN)C
Frag36	10	27.4	1.1	38	29.6	NC(=O)N
Frag37	10	30.7	1.0	30	30.7	N[1](CC(NC=O)C@1=O)CC(O)=O
Frag38	16	31.4	1.0	21	31.4	O=C(C)CC=O
Rule5	19	9.1	3.7	577	33.2	
Frag39	6	38.7	1.0	13	38.7	N[1]=NNC=N@1
Frag40	10	20.9	2.0	14	41.7	N(CC)(C)COC
Frag41	20	57.4	1.0	5	57.4	C[1](C[2]=C(NC(CN=@1)=O)C=CC(=C@2)Cl)C[14]C=CC=CC=@14
Frag42	11	68.4	1.0	9	68.4	C[1:s=I]C(=CCC=C@1)CCC[9]C@1CCC[13:s=N]C@9CCC@13

[a] nm = the occurrences of one fragment in the 20 (for human oral bioavailability and urine excretion) or 30 (for plasma protein binding) models; c = the average coefficient of one fragment in regression analysis; v = the average count of one fragment for the molecules that have it; nf = the number of training set molecules that have one fragment; $c \times v$ = a measure of the contribution to the PK property in study by one fragment; and sln = the sybyl line notation of this fragment.

an encouraging result given the fact that the standard error of the plasma protein binding data was 10%. This performance was comparable to or better than the protein binding model in volsulf[5] implemented in Sybyl, version 6.9,[18] which gave a leave-one-out R^2 of 0.52 for 338 data in the partial least-squares analysis.

For urinary excretion, GA-QSPR also achieved a better performance than that of HQSAR (0.60 vs 0.37 for R^2 and 18.3 vs 23.3% for RMSE). Not surprisingly, the performance

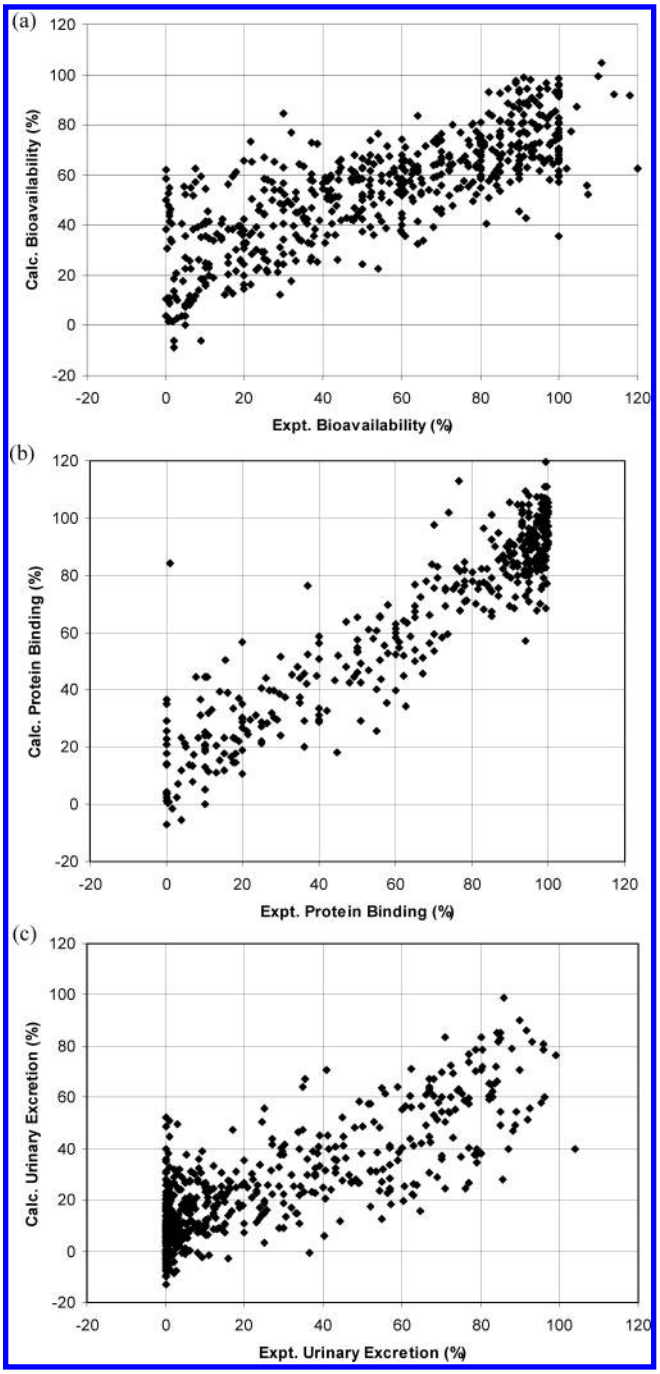


Figure 3. Experimental vs calculated pharmacokinetics properties from the individual QSPR consensus score models: (a) bioavailability, (b) plasma protein binding, and (c) urinary excretion.

of the GA-QSPR consensus model is further improved, and its R^2 and RMSE values were 0.65 and 17.2%, respectively.

The plots of the experimental versus predicted PK properties for the consensus score models are shown in Figure 3. It is noticeable that the experimental data of the plasma protein binding and urinary excretion are not evenly distributed from 0 to 100%.

2. Fragment Analysis. An advantage of GA-QSPR is that it can generate multiple comparable models naturally and a deeper insight into the contribution of each descriptor can be obtained through descriptor analysis. Some key fragments that make large contributions to the three PK properties are listed in Tables 3−5, and their structures are shown in Figure 4. For each fragment listed in those tables, nm represents

Table 4. List of Key Molecular Fragments and Molecular Properties that Reduce and Boost Plasma Protein Binding[a]

	nm	c	υ	nf	$c \times \upsilon$	sln
Frag1	8	−33.1	2.1	9	−69.9	N(CC(=C)CO)C
Natom	23	−1.2	50.8	404	−62.2	
Frag2	11	−48.6	1.2	6	−56.7	N(CC(O)C=O)C
Frag3	13	−17.8	2.9	11	−51.7	S(C(C)C)CCC
Frag4	27	−21.7	1.2	6	−25.3	N[1]=CN=CN@1
Frag5	13	−17.7	1.4	15	−24.7	NCC(CCO)C
Frag6	15	−21.4	1.1	9	−23.8	S(C[2]=CC=C(N)C=C@2)(=O)=O
Frag7	14	−23.4	1.0	12	−23.4	C(C)CC[4]C=CC(OCC)dCC=@4
Frag8	16	−15.7	1.4	49	−21.7	C(C)(C)(C)C
Frag9	11	−15.6	1.3	34	−20.6	N(CCOC)(C)C
Frag10	13	−20.0	1.0	23	−20.0	N[1](CC(NC(CC=C)=O)C@1=O)CC(O)=O
Frag11	9	−19.9	1.0	12	−19.9	NC(CC=O)C
Frag12	14	−6.7	3.0	64	−19.9	N(CC(N)C)CC
Frag13	12	−5.0	3.2	53	−16.0	O(C(C)C)CCC
Frag14	10	−14.8	1.0	17	−14.8	NC(=NC)N
Frag15	13	−13.5	1.1	26	−14.6	C[1]NCCC@1
Frag16	11	−11.6	1.2	9	−14.1	C[1](=CC=CC=C@1)OCC(CNC(C)C)O
Frag17	17	−7.2	1.8	49	−13.3	C[1](C=CC(O)=CC=@1)CC
Frag18	11	−8.3	1.6	17	−13.1	N(C)(CN)CCC
Frag19	27	−5.7	2.3	65	−13.0	N(C=CCC)C
Frag20	29	−7.7	1.6	62	−12.5	C[1]=CC=CC=C@1OCC
Frag21	8	−11.5	1.1	17	−12.2	C(=O)NH2
Frag22	10	−8.6	1.4	62	−12.2	O=CCCO
Frag23	19	−9.9	1.2	69	−11.6	NCC[3]=CC=CC=C@3
Frag24	15	4.5	2.3	123	10.2	OC=CC
Frag25	8	2.7	3.8	53	10.5	NCC(C)CC
Frag26	8	5.6	2.0	218	11.3	OH
Frag27	16	5.6	2.0	170	11.3	NC(=CC)C
Frag28	10	11.9	1.0	90	11.9	C(=O)(OH)R1
Frag29	12	10.2	1.2	13	12.5	C#N
acceptor	12	2.6	5.0	392	12.9	
Frag30	9	6.5	2.3	23	14.6	N(C(COC)C)C
Frag31	12	13.8	1.1	9	15.3	S[1]C=CC=C@1
Frag32	10	6.2	2.7	20	16.8	O=CCC(C)C=C
Clogp	30	6.3	2.7	404	16.9	
Frag33	20	2.9	6.0	61	17.4	OC(CCO)C
Frag34	12	10.7	1.9	61	19.8	N(C(CN)C)CC
Frag35	15	8.4	2.4	8	19.8	ClCCNCC
Frag36	8	22.5	1.0	7	22.5	S(CCC(=C)C)C
Frag37	8	23.8	1.0	8	23.8	N(C(C)(C)C)C=O
Frag38	14	2.3	10.7	159	24.2	C(CC)(C)C
Frag39	15	24.3	1.0	20	24.3	CN[2]CCN(C)CC@2
Frag40	8	24.8	1.0	27	24.8	N[1]CCNCC@1
Frag41	25	2.3	11.2	314	25.1	C=CC=CC
Frag42	17	26.0	1.1	21	27.2	C[1](S(N)(=O)=O)C=CC=CC=@1
Frag43	15	27.6	1.1	26	30.8	S(C[2]=CC=CC=C@2)(=O)=O
Frag44	10	31.2	1.0	7	31.2	C[1](C(C)=CC=C)C(=C(NC(=C@1C(OC)=O)C)C)C(OC)=O
Frag45	13	36.1	1.0	12	36.1	S[1]CNCC@1
Rule5	30	10.1	3.6	403	36.4	
Frag46	24	39.7	1.0	8	39.7	N[1]C(C(NC(C)=O)C@1=O)SCC(=C@1C(O)=O)CSCdNN
Frag47	8	51.0	1.1	11	55.6	NCC(=C)CO
CMR	20	9.1	10.2	404	91.9	

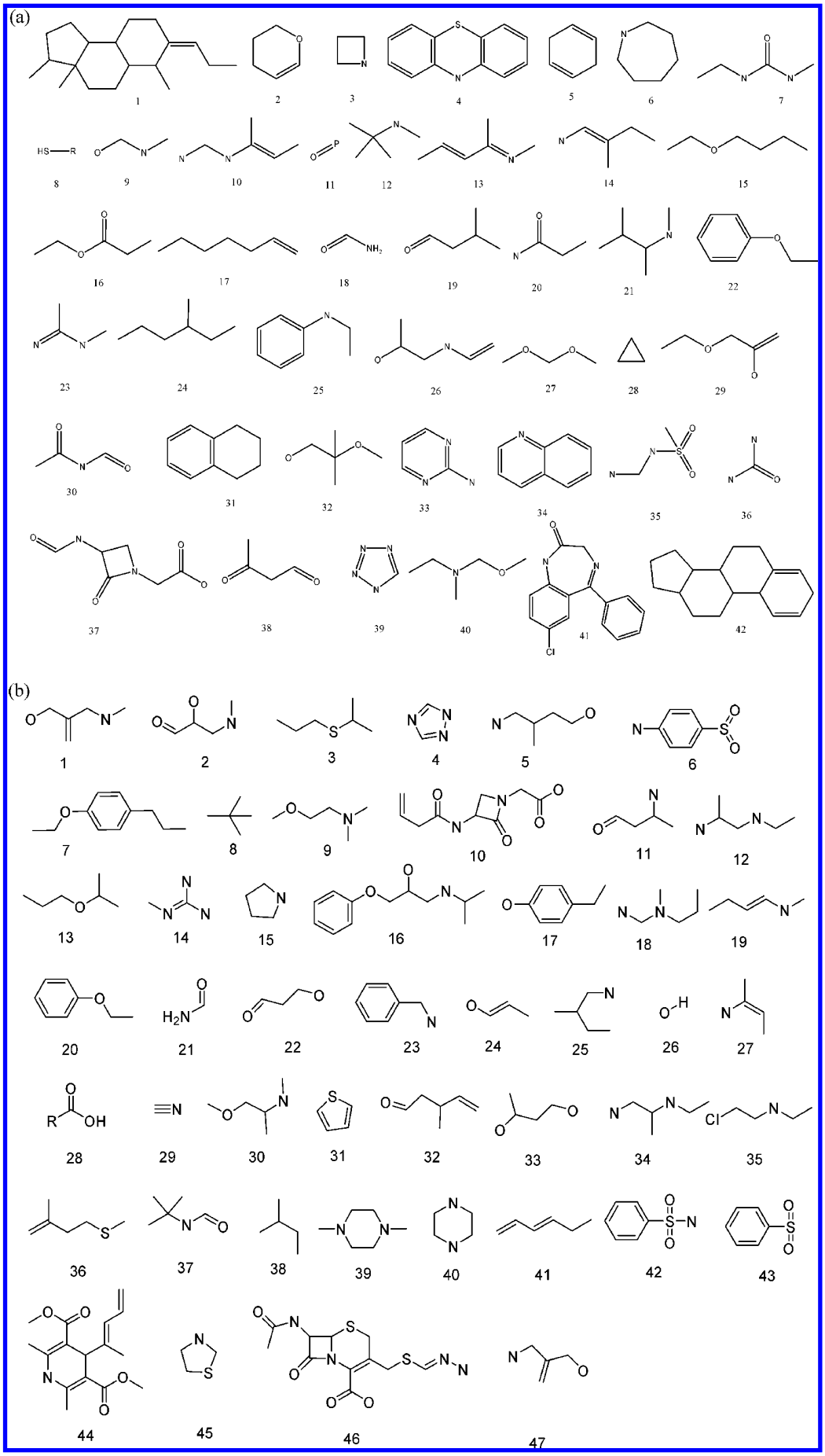
[a] nm = the occurrences of one fragment in the 20 (for human oral bioavailability and urine excretion) or 30 (for plasma protein binding) models; c = the average coefficient of one fragment in regression analysis; υ = the average count of one fragment for the molecules that have it; nf = the number of training set molecules that have one fragment; $c \times \upsilon$ = a measure of the contribution to the PK property in study by one fragment; and sln = the sybyl line notation of this fragment.

Table 5. List of Key Molecular Fragments and Molecular Properties that Reduce and Boost Urinary Excretion[a]

	nm	c	υ	nf	$c \times \upsilon$	sln
Frag1	15	−23.9	4.0	8	−95.5	P(=O)CPO
Frag2	20	−33.0	1.1	12	−35.7	NC(OC)CO
Frag3	11	−26.3	1.1	7	−30.0	N(C(O)(C)C)CC
Frag4	7	−10.7	2.5	15	−26.3	S(CCCC)CC
Frag5	6	−14.4	1.6	16	−22.4	NCC(OC)CO
MW	7	−0.1	378.1	581	−22.2	
Frag6	12	−11.7	1.8	41	−21.1	S(CNC)CCC
Frag7	6	−16.0	1.3	17	−20.6	O=C[2]C=CCC=C@2
Frag8	11	−6.1	3.2	13	−19.7	S(=O)(CC)C
Frag9	9	−13.6	1.3	36	−17.0	NC(=CC=O)C
Frag10	16	−6.7	2.5	17	−16.9	N(CCNCN)C
Frag11	6	−6.9	2.0	35	−14.0	N(C(C=C)C)C
Frag12	6	−3.9	3.3	12	−12.6	FC(F)CC
hydrophobe	12	−4.0	3.1	557	−12.3	
Frag13	8	−5.5	2.1	20	−11.3	C(NC[3]=CC=CC=C@3)(=CC)C
Frag14	9	−4.8	1.7	84	−8.0	C=CC[3]=CC=CC=C@3
Frag15	6	−4.2	1.6	185	−6.7	C(C=C)C=C
Frag16	6	−2.7	2.4	250	−6.5	N(C)C=O
Frag17	6	−4.5	1.3	150	−6.0	C(=O)N(R1)R2
Frag18	8	4.6	1.6	71	7.2	O(CC)CC(=C)C
Frag19	14	1.7	4.6	358	7.7	N(R1)R2
Frag20	6	4.9	1.7	73	8.5	NC=CCC=CC
Frag21	6	7.1	1.3	152	9.4	N(CCO)(C)C
Frag22	17	9.9	1.1	57	11.1	NCOC
Frag23	9	11.0	1.1	27	11.8	NC(CC=O)C
Frag24	15	11.5	1.1	58	12.5	N[1]CCCC@1
Frag25	13	12.8	1.2	39	14.8	S(=O)(N)=O
Frag26	11	8.7	1.7	36	14.9	N(CC=C(N)C)C
Frag27	10	9.8	1.7	62	16.7	SCCCNCC
Frag28	16	12.2	1.4	65	16.7	SCCN
Frag29	10	11.1	1.5	17	17.0	NCN=CC=CN
Frag30	9	5.5	3.4	34	18.8	O(CO)CCO
Frag31	13	7.4	2.6	64	19.3	S(CC)CCC
Frag32	6	19.6	1.0	6	19.6	N[1]=CC=CC[5]=C@1C=CC=C@5
Frag33	18	26.9	1.0	11	26.9	NC(=O)C=NOC
Frag34	13	28.0	1.0	6	28.0	N(C(=C)CO)C=C
Frag35	18	26.1	1.2	17	32.3	N[TAC=4]
Frag36	8	23.2	1.4	7	33.1	N[1]CCC[4]=C(C=CC=C@4)C@1
Frag37	9	33.4	1.0	8	33.4	N[1](CCNCC@1)C[7]=CC[9]=C(C=C@7)CC=CN@9
Frag38	16	33.9	1.0	6	33.9	N[1]=CN=CC[5]=C@1NC=N@5
Frag39	12	34.6	1.0	7	34.6	C[1]=NN=NNH@1
Frag40	13	9.5	4.0	6	38.0	P(=O)(O)C(P)(O)C
Frag41	20	41.5	1.0	9	41.5	N(C(=O)CO)C=C
Frag42	9	43.6	1.0	38	43.6	N[1](CC(NC(C)=O)C@1=O)CC(O)=O
Frag43	18	48.5	1.0	8	48.5	C[1](C=CNC(N=@1)=O)N
Frag44	17	52.9	1.0	7	52.9	C[1](C[2]N(C(=C(CSC[8]N(N=NN=@8)C)CS@2)C(O)=O)C@1=O)NC(C)=O
Frag45	15	38.8	1.5	17	57.0	O=P[TAC=4]

[a] nm = the occurrences of one fragment in the 20 (for human oral bioavailability and urine excretion) or 30 (for plasma protein binding) models; c = the average coefficient of one fragment in regression analysis; υ = the average count of one fragment for the molecules that have it; nf = the number of training set molecules that have one fragment; $c \times \upsilon$ = a measure of the contribution to the PK property in study by one fragment; and sln = the sybyl line notation of this fragment.

the occurrences of this fragment in the 20 (for human oral bioavailability and urine excretion) or 30 (for plasma protein binding) models, c is the average coefficient of this fragment in regression analysis, υ is the average count of this fragment for the molecules that have it, and $c \times \upsilon$ is a measure of the contribution to the PK property in study by this fragment. A more positive $c \times \upsilon$ value results in a higher contribution of the fragment to boost the PK value. On the other hand, a


(a)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
(b)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

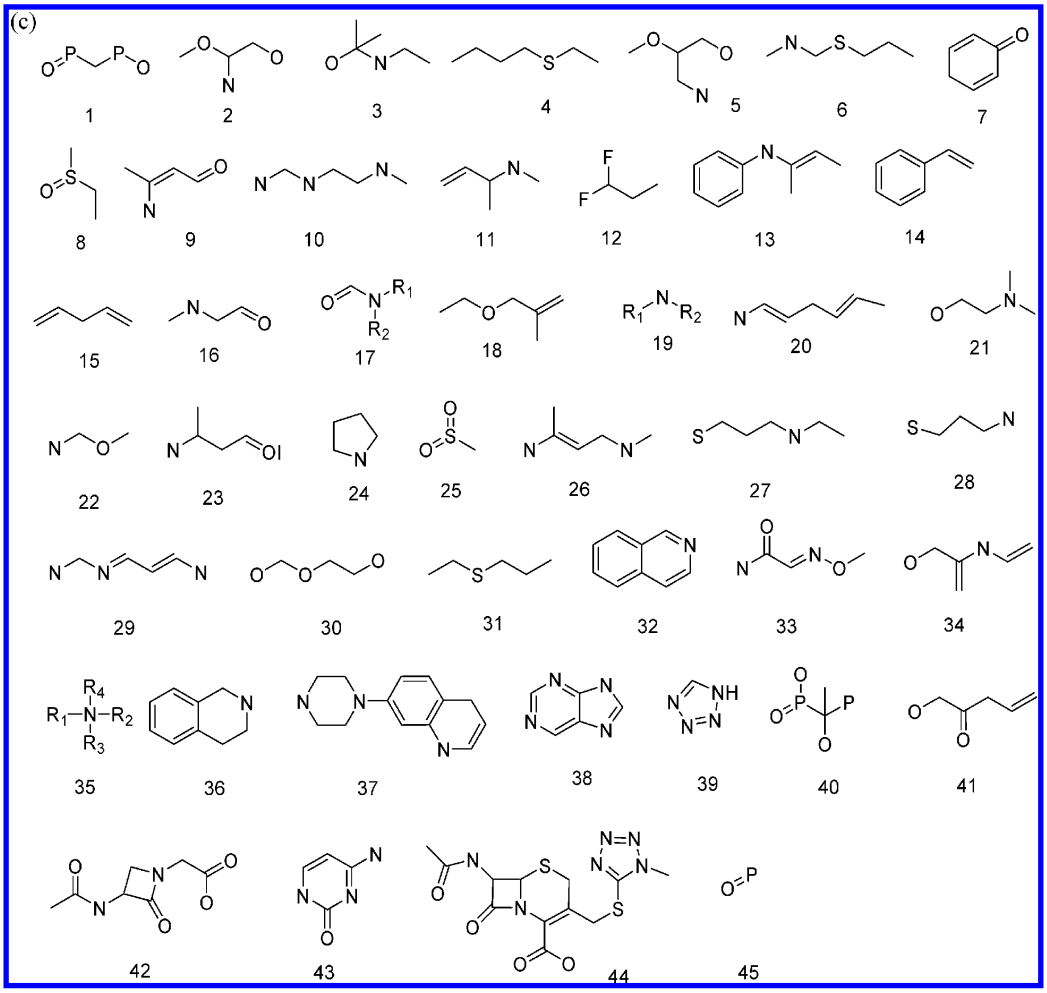


Figure 4. Key fragments that reduce and boost pharmacokinetics properties: (a) bioavailability, (b) plasma protein binding, and (c) urinary excretion.

more negative $c \times v$ value leads to a higher contribution of the descriptor to reduce the PK value. nf is the number of training set molecules that have this fragment; sln is the sybyl line notation of this fragment.[9] A descriptor, which is essential to a pharmacokinetic property, should have not only a large $c \times v$ value but also large nm and nf values. In other words, descriptors that are occasionally selected in one or two models or that do not apply to most of training set molecules (such as rare fragments only show up in one or two molecules) are unlikely important to the property in this study. For example, the descriptor of rule-of-five, which has large $c \times v$, nm, and nf values (33.2, 19, 577, respectively), is essential for human oral bioavailability. It makes sense since a molecule that obeys the Lipinski's rule typically has good absorption potential and solubility, and thus, it is likely to have a good bioavailability. In another example, Frag24 is more important than Frag25 for plasma protein binding, even though their $c \times v$ values are almost the same, since Frag24 shows up in 15 GA-QSPR models out of 30 and 123 molecules have it, while Frag25 just appears in 8 models and only 53 molecules have this fragment (see Table 4). It is emphasized that the inclusion or removal of one fragment may cause changes on other descriptors; therefore, the net contribution should be calculated with all the descriptors in the model being considered.

3. Hit Rate Analysis. Database searches were performed for MDDR, a database of marketed drugs and drug leads that entered development, and ACD, a database containing a great variety of available compounds. The total entries of the two databases are 141k and 237k, respectively. The hit rates in percentage of 12 simple fragments that boost (boost set) and 12 that reduce (reduce set) bioavailability are shown in Figure 5. For the MDDR database, the total percentage hit rates of 12 fragments are 31.2 and 24.1% for the boost set and reduce set, respectively. Whereas for the ACD database, the total percentage hit rates are 8.5 and 10.3% for the boost set and reduce set, respectively. It is not surprising that MDDR has a higher percentage hit rate than ACD in both sets. However, the ratio of the total hit rates for the two databases is significantly larger for the boost set than for the reduce set (3.7 vs 2.3). About 38% of molecules in the MDDR database were found having at least one of the 24 fragments. Interestingly, fragment 3 (azetidine) and fragment 7 (1-ethyl-3-methyl-urea) in the reduce set have a larger percentage hit rates for MDDR than for ACD. This is not a surprise since bioavailability is not the only property to be concerned in drug development, and the two fragments may be important for other properties, for instance activity and selectivity. If the two fragments are not considered, the total percentage hit rate of the reduce set is only 14.9% for MDDR, marginally larger than that for ACD (10.3%) and significantly smaller than that of the boost set for MDDR (31.2%). The conclusion is that the fragments identified in the bioavailability models are common building blocks in

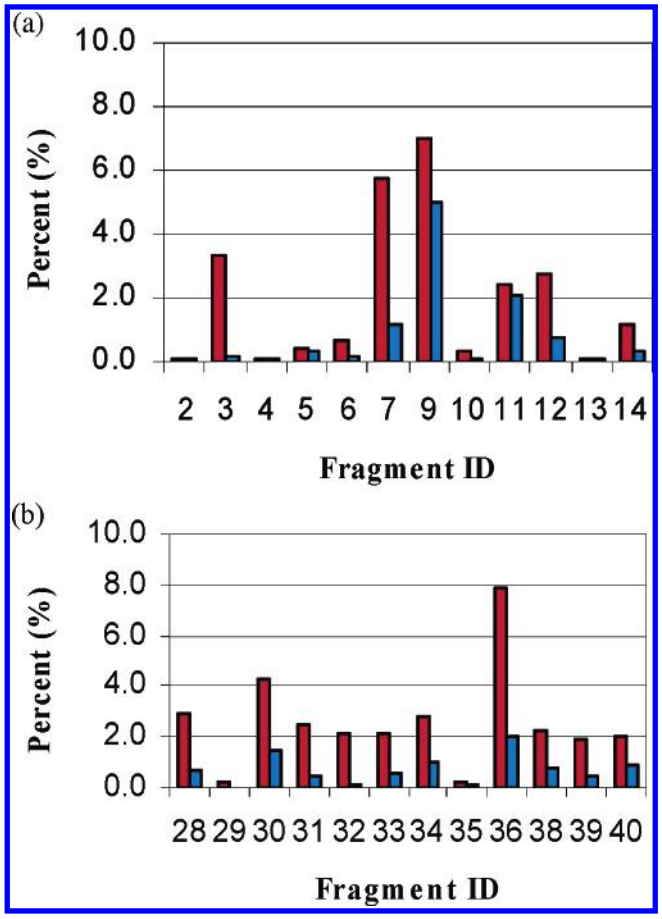


Figure 5. Hit rate (in percentage) of small fragments that boost or reduce bioavailability upon searching MDDR (red) and ACD (blue) databases: (a) 12 fragments that reduce bioavailability and (b) 12 fragments that increase bioavailability.

druglike molecules, and the boost set fragments are more likely to occur in druglike molecules than those in reduce set.

Hit rate analysis was also conducted for plasma protein binding and urinary excretion. For the boost set of plasma protein binding (Frag36−Frag47 in Table 4), the total percentage hit rates are 175.6 and 100.8% for the MDDR and ACD databases, respectively. Whereas for the reduce set (Frag1−Frag12 in Table 4), the total percentage hit rates are 37.7 and 18.7% for MDDR and ACD, respectively. It is not a surprise that the total hit rate of the boost set is much larger than that of reduce set, given the fact that the mean plasma protein binding of 404 drugs is 65%. For urinary excretion, the total percentage hit rates of 13 boost set fragments (Frag33−Frag45 in Table 5) are 11.7 and 5.7% for MDDR and ACD, respectively, while the total percentage hit rates of 13 reduced set fragments (Frag1−Frag13 in Table 5) are 42.1 and 17.2% for MDDR and ACD, respectively. The much larger total hit rate of the reduce set than that of the boost set is reasonable since the mean urinary excretion of 581 drugs is 24.9%. Unlike oral bioavailability, both plasma protein binding and urinary excretion have similar ratios of the total hit rates between the MDDR and ACD databases for the boost set and the reduce set (1.7 versus 2.0 for plasma protein binding and 2.0 versus 2.4 for urinary excretion). This is understandable since bioavailability is one major optimized parameter in drug discovery, whereas the plasma protein binding and urinary excretion are not in most cases. Finally, about 96% of the molecules in MDDR database contain at least one of the 24 key plasma protein binding fragments mainly because Frag41 and Frag38 are common fragments in druglike molecules. In contrast, about 39% molecules in MDDR database contain at least one of the 26 key urinary excretion fragments.

4. Further Development. As an intelligent optimization algorithm, GA can deal with a great number of descriptors. Unlike stepwise regression and partial least-squares fitting, GA can efficiently select a subset of descriptors to build up a set of QSPR comparable models. Although many other optimization methods can generate multiple models with different initial conditions, GA can produce multiple models by simply performing optimizations many times since random numbers are involved in GA optimizations. The relative importance of a descriptor is assessed not only by its coefficients but also by the frequency of the descriptor showing up in a set of models. Thus, GA-QSPR provides a deeper insight into how the modeled molecular property is being affected by its descriptors.

GA is a very flexible algorithm that has many adjustable parameters, which include population size, elite size, mutation probability, and crossover probability, as well as selection methods. Although there are some guidelines on how to set parameters for different kinds of problems (for examples, a large mutation rate could help GA to escape from local minima and is suitable to solve multiple minima problems), the users must set the parameters prior to carrying out GA optimizations. Evidently, it is a better idea to optimize those GA parameters rather than arbitrarily setting ones to achieve better performance (better score, fewer iterations) for a given problem. Therefore, it is necessary for us to continue to optimize the GA parameters and test different selection methods for the algorithm to achieve good performance in building up QSPR models using molecular fragments as descriptors.

Although the strategy of applying GA to select a set of fragments to model pharmacokinetic properties is justified by the encouraging results, the weakness of this methodology should be pointed out here. The GA-QSPR models may not be very robust since they are based on a training set of less than 1000 compounds and the fragment library does not enumerate all the key fragments occurring in drugs outside the training set. This is the reason the cross-validated correlation coefficients of the three PK models are considerably smaller than the nonvalidated ones. It is expected that more experimental data will appear in the literature so that our models can be further improved.

Despite this deficiency, the fragment-based descriptors provide an intuitional way to amend a molecular property by including or avoiding some molecular fragments. This information cannot be easily obtained from other descriptors including those applied in HQSAR.

CONCLUSIONS

In this work, we successfully applied a genetic algorithm (GA) program to build up a set of QSPR models for human oral bioavailability, plasma protein binding, and urinary excretion using molecular fragment counts as descriptors. The consensus score model, based on a set of comparable GA-QSPR models, was found to improve the correlation

coefficient and reduce the standard errors significantly: the squares of regression coefficients and root-mean-square errors of the consensus score models were 0.62 and 20.2%, 0.86 and 13.0%, and 0.65 and 17.2% for human oral bioavailability, plasma protein binding, and urinary excretion, respectively. Key fragments that could boost and reduce pharmacokinetic properties were also identified. The result of percentage hit rate analysis of database searches for fragments in the bioavailability models was consistent to the human sense that druglike molecules overall have higher bioavailability than those that are not druglike.

ACKNOWLEDGMENT

We are grateful to acknowledge the research support from NCSA (MCB000013N (J.W.)).

Supporting Information Available: The complete lists of descriptors (eight basic molecular properties and molecular fragments that make contribution to the modeled pharmacokinetic property) and their coefficients for the three pharmacokinetic models. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Beresford, A. P.; Selick, H. E.; Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discovery Today* **2002**, *7*, 109−116.
(2) Egan, W. J.; Merz, K. M., Jr.; Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.
(3) Blake, J. F. Chemoinformatics−predicting the physicochemical properties of 'drug-like' molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104−107.
(4) 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.* **2001**, *46*, 3−26.
(5) Cruciani, G.; Pastor, M.; Clementi, S. Handling information from 3D grid maps for QSAR studies. In *Molecular Modeling and Prediction of Bioactivity*; Gundertofte, K., Jørgensen, F. E., Eds.; Kluwer Academic/ Plenum Publishers: New York, 2000; pp 73−82.
(6) Duffy, E. M.; Jorgensen, W. L. Prediction of properties from simulations: Free energies of solvation in hexadecane, octanol, and water. *J. Am. Chem. Soc.* **2000**, *122*, 2878−2888.
(7) Xing, L.; Glen, R. C. Novel methods for the prediction of log P, pK_a, and log D. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.
(8) 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.
(9) *Sybyl User Manual*; Tripos Inc.: St. Louis, MO, 1995.
(10) Guba, W.; Cruciani, G. Molecular field-derived descriptors for the prediction of pharmacological data. In *Molecular Modeling and Prediction of Bioactivity*; Guntertofte, K., Jørgensen, F. S., Eds.; Kluwer: New York, 2000; pp 89−95.
(11) Wang, J.; Kollman, P. A. Automatic parametrisation of force field by systematic search and genetic algorithms. *J. Comput. Chem.* **2001**, *22*, 1219−1228.
(12) Hou, T.; Wang, J.; Xu, X. Application of genetic algorithm on the structure activity correlation study of a group of non-nucleoside HIV-1 RT inhibitors. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 303−310.
(13) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(14) Xiao, Y. L.; Williams, D. E. Genetic algorithms for docking of actinomycin d and deoxyguanosine molecules with comparison to the crystal structure of actinomycin d-deoxyguanosine complex. *J. Phys. Chem. B* **1994**, *98*, 7191−7200.
(15) Bowie, J. U.; Eisenberg, D. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 4436−4440.
(16) Benet, L. Z.; Øie, S.; Schwartz, J. B. Design and optimization of dosage regimens: Pharmacokinetic data. In *Goodman and Gilman's the Pharmacological Basis of Therapeutics*, 9th ed.; Hardman, J. G., Limbird, L. E., Eds.; McGraw-Hill: New York, 1996; pp 1707−1792.
(17) Thummel, K. E.; Shen, D. D. Design and optimization of dosage regimens: Pharmacokinetic data. In *Goodman & Gilman's the Pharmacological Basis of Therapeutics*, 10th ed.; Hardman, J. G., Limbird, L. I., Eds.; McGraw-Hill Companies, Inc.: New York, 2001; pp 1917−2023.
(18) Sietsema, W. K. The absolute oral bioavailability of selected drugs. *Int. J. Clin. Pharmacol., Ther. Toxicol.* **1989**, *27*, 179−211.
(19) Barnard, J. M.; Downs, G. M. *Fingerprint Descriptor Package*, version 3.1; Barnard Chemical Information Ltd.: Sheffield, U.K., 1995.
(20) MDL Information Systems Inc.: San Ramon, CA, 2004.

CI060087T