Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches

Marlene Kim[†,‡], **Alexander Sedykh**[§], **Suman K. Chakravarti**[Φ], **Roustem D. Saiakhov**[Φ], and **Hao Zhu**[†,‡,*]

[†]Department of Chemistry, Rutgers University, Camden, New Jersey 08102

[‡]The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102

[§]Division of Chemical Biology and Medicinal Chemistry, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599

[Φ]Multicase Inc., 23811 Chagrin Boulevard, Suite 305, Beachwood, Ohio 44122

Abstract

Purpose—Oral bioavailability (%F) is a key factor that determines the fate of a new drug in clinical trials. Traditionally, %F is measured using costly and time -consuming experimental tests. Developing computational models to evaluate the %F of new drugs before they are synthesized would be beneficial in the drug discovery process.

Methods—We employed Combinatorial Quantitative Structure-Activity Relationship approach to develop several computational %F models. We compiled a %F dataset of 995 drugs from public sources. After generating chemical descriptors for each compound, we used random forest, support vector machine, *k* nearest neighbor, and CASE Ultra to develop the relevant QSAR models. The resulting models were validated using five-fold cross-validation.

Results—The external predictivity of %F values was poor (R^2=0.28, n=995, MAE=24), but was improved (R^2=0.40, n=362, MAE=21) by filtering unreliable predictions that had a high probability of interacting with MDR1 and MRP2 transporters. Furthermore, classifying the compounds according to the %F values (%F<50% as "low", %F 50% as 'high') and developing category QSAR models resulted in an external accuracy of 76%.

Conclusions—In this study, we developed predictive %F QSAR models that could be used to evaluate new drug compounds, and integrating drug-transporter interactions data greatly benefits the resulting models.

Keywords

oral bioavailability; intestinal membrane transporter; QSAR; drugs

INTRODUCTION

Drug oral bioavailability is the fractional extent of the drug dosage that finally reaches the therapeutic site of action and is quantitatively symbolized as %F (1). In many cases, most of the orally administered drug is metabolized and eliminated before reaching systemic blood circulation (1). Therefore, poor bioavailability may cause a new drug to fail clinical trials,

[*]Corresponding author: 315 Penn St., Rutgers University, Camden, NJ 08102, Telephone: (856) 225-6781, hao.zhu99@rutgers.edu.

even if it has high efficacy in previous *in vitro* and/or *in vivo* tests. The traditional process for measuring the %F of a drug is expensive, costly, and time-consuming. Using computational methods as an alternative to calculating the %F of new drug candidates, even before synthesizing the compound, would be advantageous by saving resources and provides a promising alternative to traditional experimental protocols.

To date there are many computational oral bioavailability models that are available (2–11). Some are based on Quantitative Structure-Activity Relationship (QSAR) models that predict the oral bioavailability of new compounds directly from the molecular structure. Table I lists several major QSAR studies on oral bioavailability. In 2000, Andrews *et al*. developed a computational oral bioavailability model using linear regression. This model was able to predict highly bioavailable compounds accurately, but had poor performance for low bioavailable compounds (2). Moda *et al*. developed hologram QSAR oral bioavailability models that predicted %F using fragment descriptors. However, poorly soluble and non-oral bioavailable drugs were excluded intentionally from the modeling set (3). Ma *et al*. used a Combinatorial QSAR (Combi-QSAR) approach to develop an oral bioavailability classification model. Although the unbalanced accuracy for a five-fold cross-validation of their modeling set was 80%, the specificity (correct predictive rate for inactive compounds) was only 20% due to the high imbalance between actives and inactives (4). More recently, Tian *et al*. attempted to create multiple linear regression human oral bioavailability models by combining molecular properties and structural fingerprints with genetic function approximation. The predictivity of the reported model was acceptable ($R^2_{ext}=0.50$), but the structural fingerprints used to generate the training set does not apply to all drug classes. This limits the applicability of the model for predicting new classes of compounds (5).

In addition to the QSAR models mentioned above, previous research suggests that the rule-based models, such as the rule-of-five (6), are not sufficient enough for evaluating the oral bioavailability of drugs (7–9). Nevertheless such empirical rules are useful for qualitative assessment and we list in Supplemental Table I several *rules* previously developed for assessing drug oral bioavailability and absorption. In 2002, Veber *et al*. studied the molecular properties and *in vivo/in vitro* pharmacokinetic parameters that affect oral bioavailability (7). The authors concluded that the molecular properties of the drug, target receptor, cell membrane, and transporter proteins should all be studied during drug development. Ignoring one factor can result in poor bioavailability (7). More recently, property-based rules for bioavailability (5) and parameters needed for optimal oral bioavailability classification (10) were evaluated. There are certain physical properties that contribute to oral bioavailability, but these parameters are better at predicting intestinal absorption (5,7,10). Recently, Paixão used *in vitro* test results as parameters to develop an oral bioavailability model (11). Incorporating *in vitro* data helped improve the prediction accuracy of the resulting models.

In this study, we developed several novel models of human oral bioavailability of pharmaceutical drugs. After compiling over one thousand drugs and their experimental %F values, we corrected the data entry errors using both automatic tools and manual curation steps. We utilized the Combi-QSAR approach to develop several computational oral bioavailability models. A series of individual category (CTG) and continuous (CNT) models were developed and validated using a five-fold cross-validation. To improve the predictivity of the resulting QSAR models, we tried to integrate Human Intestinal Transporter (HIT) interactions into the final predictions. This hybrid approach was able to exclude compounds with considerable prediction errors from the final predictions. Our predictive Combi-QSAR oral bioavailability models can be used to assess and evaluate new drug candidates. Furthermore, similar approaches could be developed and utilized to model other complex biological activities for drug and drug like molecules.

METHODS

Human Oral Bioavailability Dataset

The human oral bioavailability dataset was compiled from various public and private sources (3,5,8,12–17). Originally it contained over 1,300 entries. Several tools (CASE Ultra, Chem Axon Standardizer, Chem Axon Structure Checker) were used for chemical structure curation and standardization. For duplicate entries, one was removed. For stereoisomers, the structure of the compound with the highest activity was kept. For salts, the chemical structure was neutralized. Mixtures were separated and the largest component was kept. All metals, metaloorganics, and inorganic entries were removed.

We also carefully evaluated the experimental %F values in our dataset. It was common to find different %F values for the same compound among different sources. We selected the %F values reported in, *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, over the %F values reported from other sources, because the bioavailability data in this book was curated and harmonized by experienced medicinal chemists (12). In other cases, the values were harmonized if the range of the %F values were less than 10 for the same compound. If the %F value for the salt and neutral forms were different, the %F value for the neutral form was kept. For compounds with disparate %F values, the experimental studies that reported the values were carefully evaluated. After comparing sources, the %F from the study that clearly defined the method for determining the %F value was selected. A total of 995 unique compounds remained for the following modeling process after the curation.

After harmonizing the %F values, the compounds were classified as low bioavailable (%F<50, n=455) and high bioavailable (%F 50, n=540). There is no universal criterion to define high and/or low bioavailable compounds. We used %F=50% as an arbitrary classification threshold in this study since it could also balance the ratio of two classifications in the dataset. Non-oral drugs (%F=0), e.g. compounds commonly administered by intramuscular or intravenous injection, were included in the low bioavailable group. Also, using sigmoid function, we transformed the %F values to $logK(\%F)$, a pseudo-equilibrium constant, as it has more balanced distribution of values and could afford improved models.

$$logK(\%F) = log\left(\frac{\%F}{100 - \%F}\right) \quad (1)$$

The distribution of all 995 compounds based on the %F values is displayed in Figure 1. Supplemental Table II lists all 995 compounds, the oral bioavailability values, and the corresponding references.

Chemical Descriptors

Chemical descriptors for each compound were generated using 2-D chemical descriptors from Dragon ver. 6.0 (Talete SRL, Milano, Italy) and Molecular Operating Environment (MOE) ver. 2011.10. Dragon descriptors included constitutional indices, ring descriptors, topological indices, walk and path counts, connectivity indices, matrix-based descriptors, autocorrelations, Burden eigenvalues, edge adjacency indices, functional group counts, atom-centered fragments, atom-type, E-state indices, atom pairs, molecular properties, and drug-like indices. MOE descriptors included physical properties, structural keys, E-state indices, topological polar surface area, and topological indices. Initially, the Dragon and MOE software generated 3,753 and 186 descriptors, respectively. Since many Dragon descriptors in this dataset were redundant, the number of Dragon descriptors was reduced by

removing low variance (standard deviation <0.01 or missing values) and highly correlated (r >0.95) descriptors. The remaining 1,597 Dragon and 186 MOE descriptors were range-scaled to [0,1] and used in the modeling process except for CASE Ultra, which has its own built-in fragment descriptors.

Modeling Approaches

In this study, the implementation of the Random Forest (RF) (18) and Support Vector Machine (SVM) (19–22) algorithms available in R.2.15.1 (23) were used. The *k* Nearest Neighbor (*k*NN) models (24) were built using Chembench (chembench.mml.unc.edu).

CASE Ultra

CASE Ultra is a QSAR expert system and can automatically generate a predictive model from a training set of non-congeneric compounds with associated biological activity data. The training set usually contains examples of both active and inactive chemicals and the algorithm identifies positive and deactivating alerts (structural fragments statistically related to activity and inactivity) after processing them. These alerts form a CASE Ultra model that can be used to predict activity of a test chemical (25,26).

Combinatorial QSAR Modeling Workflow

The entire Combinatorial QSAR modeling workflow is shown in Figure 2. Individual models were developed using Dragon (denoted by the prefix "D") or MOE descriptors and either RF, SVM, or *k*NN modeling methods. CASE Ultra was used to develop a single CTG model. This resulted in seven different CTG, four different CNT-%F, and four different CNT-logK(%F) models. The individual CTG models were D-RF, D-SVM, D-*k*NN, MOE-RF, MOE-SVM, MOE-*k*NN, and CASE Ultra. The individual CNT-%F and CNT-logK(%F) models were D-RF, D-SVM, MOE-RF, and MOE-SVM. The results for each CTG model and CNT model were averaged to generate the corresponding consensus CTG, CNT-%F, and CNT-logK(%F) predictions, which will be further referred to as consensus models (Figure 2).

All models were validated using five-fold external cross-validation. Briefly, the oral bioavailability dataset was randomly divided into five equal subsets. One subset was used as the validation set (20%) and the other four subsets (80%) were used as the training set. The training set was used to develop the models and the models were validated by the left-out validation set. The procedure was repeated five times so that each compound was in a validation set. Additional details about the modeling approaches can be found elsewhere (27,28).

Universal Statistical Figures of Merit for All Models

Since various modeling approaches and different descriptors were used in the modeling process, universal statistical metrics were needed to evaluate the performance of the models developed individually. The results were harmonized by 1) using sensitivity (percentage of high oral bioavailable drugs predicted correctly), specificity (percentage of low oral bioavailable drugs predicted correctly), and CCR (correct classification rate or balanced accuracy) for CTG models; and 2) Pearson's multiple linear correlation coefficient (R) and mean absolute error (MAE) for CNT models. These parameters are defined as followed:

$$\% sensitivity = \left(\frac{true\ positives}{true\ positives + false\ negatives} \right) 100 \quad \text{(2)}$$

$$\% specificity = \left(\frac{true\ nagatives}{true\ negatives + false\ positives} \right) 100 \quad (3)$$

$$\% CCR = \left(\frac{sensitivity + specificity}{2} \right) 100 \quad (4)$$

$$R^2 = \frac{regression\ sum\ of\ squares}{total\ sum\ of\ squares} \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| predicted\ value_i - true\ value_i \right| \quad (6)$$

Integrating Human Intestinal Transporters Interactions of Compounds into Oral Bioavailability Predictions

We recently reported a QSAR study for predicting interactions for different HITs (29). These HIT models were used to generate the transporter interaction scores for the drugs in our oral bioavailability dataset. Interactions between molecules and HITs depend on the size, shape, charge, and the chemical properties of the molecule (30). Most of the compounds in our dataset have aromatic rings, bulky groups, and are ionizable. Compounds with these features are commonly removed from the enterocytes by the efflux transporters Multidrug Resistance Protein 1 (MDR1) and Multidrug Resistance-Associated Protein 2 (MRP2), which could decrease their oral bioavailability (30). Therefore, we used the interaction parameters of MDR1 and MRP2 to filter predictions of compounds from our models.

RESULTS

Overview of Dataset

We did a comprehensive analysis on the chemical structures and relevant bioavailability data from the public databases used in this study. This comparison revealed that only 80% of the entries in current oral bioavailability databases are accurate. There were discrepancies between reports from different sources, affecting both molecular structures and %F values. For some compounds, the substituent groups were placed at incorrect positions. Supplemental Table III lists several examples of incorrect chemical structures that were identified from the original sources and corrected.

Furthermore, Buxton *et al*. indicated that it would be normal for different sources to report different %F values for the same compound (1). However, the compounds with disparate %F values needed to be harmonized for modeling purpose. Furthermore, we found that errors from the reported %F values occurred when a source incorrectly used the neutral names and salt forms of a molecule interchangeably. All of the errors were carefully examined and corrected.

The structural similarities between the compounds in the dataset can be analyzed by performing a principal component analysis on the chemical descriptors. After generating the principal components using the 186 MOE descriptors for all of the compounds in the database, we selected the top three most important components to create a three-dimensional plot (Figure 3) for all 995 compounds. These three principal components capture around

50% of the variance in our database. This plot could be viewed as the chemical structure space covered by all the compounds in our oral bioavailability dataset. More detailed plots of the chemical structure space can be found in Supplemental Figures 1–3. According to this analysis, there are about 10 structural outliers that are dissimilar to the majority of the compounds. Most of these compounds represent non-bioavailable or low bioavailable drugs, including antibiotics, neuronal drugs, and intravenous drugs. Some previous studies showed that removing structural outliers before the modeling process was beneficial to the results of the QSAR models (3,5). In this study, we kept these outliers since they are only a small portion (∼0.1%) of the whole dataset. Furthermore, removing the outliers did not improve the resulting models (data not shown).

Category Models

We developed seven individual and one consensus model by using two bioavailability categories ("low", %F<50% and "high", %F 50%; see Methods). The five-fold external cross-validation results for all CTG models are shown in Figure 4. The sensitivity, specificity, and CCR for the individual models ranged from 59–72%, 61–70%, and 62–70%, respectively. The D-SVM model had the lowest predictivity (CCR=62%). The MOE-RF model had the highest specificity and CCR of 70%. The MOE-kNN model had the highest sensitivity of 72%. Compared to the best individual model, the consensus model showed similar statistics, with sensitivity, specificity, and CCR as 72%, 69%, and 70%, respectively. The model obtained from our commercial modeling software, CASE Ultra, had intermediate results with sensitivity, specificity, and CCR all as 65%.

Furthermore, we implemented Consensus Prediction Thresholds (CPT), as mentioned in one of our previous studies (31), to the prediction results by using different low bioavailable and high bioavailable thresholds. The prediction results from each individual model had continuous scores that ranged from 0 to 1. The 0.5 mark was initially used as the single threshold to distinguish compounds predicted as low bioavailable (CPT<0.5) and high bioavailable (CPT 0.5). Using stricter thresholds, the compounds that were predicted around 0.5 should be considered as "inconclusive." We removed these inconclusive predictions by using different CPTs to define low bioavailable and high bioavailable compounds. Two CPTs were defined: 1) <0.4 as low bioavailable and >0.6 as high bioavailable (CPT-1 scheme); 2) <0.3 as low bioavailable and >0.7 as high bioavailable (CPT-2 scheme).

Implementing CPT-1 and CPT-2 schemes enhanced the predictivity of the individual and consensus CTG models. For the individual CTG models with CPT-1 and CPT-2, the sensitivity, specificity, and CCR ranges were between 61–87%, 50–82%, and 59–83%, respectively (results not shown here). In the consensus CTG model, the sensitivity, specificity, and CCR were 78%, 74%, 76%, respectively for CPT-1 and 82%, 77%, 79%, respectively for CPT-2 (Figure 5). As the tradeoff for excluding compounds with inconclusive predictions, using CPT-1 and CPT-2 decreased the consensus model coverage to 71% and 46%, respectively.

Continuous Models

We also developed four individual and one consensus model for the CNT-%F and CNT-logK(%F) bioavailability datasets. The results for both types of models are shown in Table II. The statistics for the four individual CNT-%F models were relatively poor (R^2=0.13–0.30 and MAE=∼24–53). Using Applicability Domain (AD) to remove unreliable predictions of structurally dissimilar compounds, as described previously (24), did not give significant improvement to our models (results not shown). We therefore did not use AD for the analysis in this study. Compared to the individual models, the consensus CNT-%F model

was also close to the upper boundary ($R^2=0.28$ and MAE=$\sim$24). To verify the statistical significance of all the models (in comparison to random chance performance), a two-way ANOVA test with a confidence level of 95% was performed for each model (32). The obtained *p* values were lower than 0.05.

The statistics for the four individual CNT-logK(%F) models were similar ($R^2=0.11$–0.30 and MAE=$\sim$23–28). The consensus CNT-logK(%F) model was also close to the upper boundary ($R^2=0.25$ and MAE=24). The obtained *p* values were lower than 0.05. Nevertheless, the distribution of errors was very different for the CNT-logK(%F) model compared to %F scale (Figure 6). Compounds with very low and very high %F values were predicted more accurately by the CNT-logK(%F) model.

Integrating Human Intestinal Transporter Parameters into the CNT-%F Bioavailability Model

HITs are an important factor in intestinal absorption, which greatly affects oral bioavailability and other pharmacokinetic properties of their substrate (33). Figure 7 depicts the transportation of drug molecules by MDR1, MRP2, and by passive diffusion in an enterocyte. It is known that both MDR1 and MRP2 are responsible for the active efflux of drug molecules from the enterocyte to the lumen (30). For this reason, a drug with low passive diffusion, but high substrate affinity to MDR1 and/or MRP2 is not likely to be highly bioavailable. We have also considered Breast Cancer Resistant Protein, another major efflux transporter, but its imputed interactions did not enhance our results (data not shown) and we excluded it from further analysis. We used four MDR1 and MRP2 model prediction results (29) to calculate the probability of interaction (POI) for the compounds in our dataset. Then, the mean probability of interaction (MPOI) for MDR1-s, MDR1-i, MRP2-s, and MRP2-i for drugs in various bioavailability ranges were calculated (Figure 8), where *s* and *i* represent substrates and inhibitors, respectively.

We established a rule that the drugs with a POI value greater than the MPOI value of orally non-bioavailable drugs in each transporter models should not have a predicted %F value greater than 10. Table III lists examples of drugs, with large %F prediction errors, and their HIT classifications. To simplify the discussion, the HIT predictions for each compound were classified as 0 (POI<MPOI) or 1 (POI>MPOI) using the rule we created. For example, all the HIT predictions and the predicted %F for Tirofiban were 1 and 54%, respectively. Therefore, Tirofiban was considered an outlier and was subsequently removed from the final model. On the other hand, Procaine had a high prediction error and could not be removed, because all the HIT predictions were classified as 0. In this case, the low bioavailability of this drug may be due to other HITs, metabolism, or other reasons. Table III lists examples of compounds with high prediction errors that were successfully removed (No. 1–3), and missed (No. 4–6) by our rule. The predictivity of compounds, that are substrates of the two transporters, could not be improved by this rule. These type of compounds with large prediction errors (e.g. compounds 4 and 5) may be due to other factors, such as metabolic stability. For example, Procaine (%F=0, Pred. %F=66) was predicted as a false positive and is metabolized by an esterase in the liver (34). The bioavailability model will be expected to be further improved by integrating metabolism-related parameters, such as CYPs interactions The compound Lymecycline (%F=99, Pred. %F=3), in Table III is a specific case. It was predicted to be the substrate of the two transporters, but it is actually a high bioavailable drug. Lymecycline is water-soluble at physiological pH and is readily absorbed through the gastrointestinal tract (35).

By using the HIT interaction rule described above to remove unreliable predictions, we were able to improve the prediction accuracy of the current CNT-%F models, especially the consensus model. We did not incorporate HIT interactions into the CNT-logK(%F) models.

Since the overall results for the two types of models were similar, doing so would have been redundant. The results for integrating various HIT parameters into the consensus CNT-%F model are listed in Table IV. It is noticeable that the use of four HIT parameters affects the predictions differently. However, the best results were obtained from combining all four transporters parameters. The R^2 coefficient enhanced from 0.28 to 0.40 and the MAE reduced from 24 to 21. Subsequently, using HIT parameters reduced the prediction coverage to 30%. A two-way ANOVA (α=95%) (32) and Bootstrap Non-Parametric Permutation (N=10,000; α=95%) (36,37) analysis revealed that the observed improvements are statistically significant. We therefore conclude that integrating HIT information with the oral bioavailability models was a valid approach. We noticed that the relationship between %F and drug interactions with MDR1 and/or MRP2 is non-monotonic. Some HIT combinations were better than others and incremental improvements were not always achieved when integrating another HIT parameter. This is understandable as there is overlap in substrate specificity between different efflux transporters (29).

DISCUSSION

Although the results for the CNT-%F and CNT-logK(%F) models are relatively low, each model has their advantages. We determined the MAE for the various %F ranges for both of the models (Figure 6). For predicting compounds with extreme %F values (%F 20% and %F 90%), the CNT-logK(%F) models performed better. For the mid %F ranges (%F=20–90%), the CNT-%F model yielded more accurate results. Both types of models can be used to predict oral bioavailability. Using the CNT-logK(%F) model can be advantageous if higher accuracy is needed for very low or very high bioavailability ranges. However, combining the results of the CNT-%F and CNT-logK(%F) models did not result in better statistics (data not shown).

Interpretation of QSAR models

There are many factors that affect oral bioavailability. Some examples are intestinal absorption, water solubility, and lipophilicity (1). These parameters can be modified to increase or decrease oral bioavailability by slightly changing certain chemical features on a compound. We evaluated chemical structures potentially related to oral bioavailability by analyzing Dragon descriptors. Dragon descriptors contain more diverse structural descriptors compared to MOE, so they more are practical for the model interpretations. We calculated the average values of the most important structural Dragon descriptors for both the 100 least bioavailable drugs (%F=0– 10) and 100 most bioavailable drugs (%F=90–99) (Figure 9). There were more descriptors related to low bioavailable drugs than high bioavailable drugs. For example, compounds with high %F normally have aromatic groups (descriptor ARR). Compounds with multiple aromatic rings like, Anthracene and Naphthalene, can readily pass through biological membranes, which facilitate their absorption and increase their bioavailability (38). Compared to aromatic rings, drugs with aliphatic carbon chains (descriptors nCsp3 and C-009) were likely to have lower %F since these kinds of drugs are poorly soluble in water, which greatly lowers their bioavailability (39). An example of this type of drug is Docosanol (%F=0%). There are several descriptors that refer to the Lipinski rule of five (6). According to the Lipinski rule of five, the existence of over five hydrogen bond donors and/or acceptors may cause the decrease of the drug bioavailability (6). The descriptor nO refers to the number of oxygen atoms (oxygen is a potential hydrogen bond acceptor or donor), and the descriptor nHDon refers to the number of hydrogen donor groups. Our modeling results support this hypothesis since there are greater descriptor values for low bioavailable drugs than high bioavailable drugs. On the other hand, the presence of aromatic halogens (descriptor nARX) was prevalent in highly bioavailable drugs. It was reported that the existence of an appropriate number of aromatic

halogens can enhance the lipophilicity and aqueous solubility of a drug, two properties critical for absorption and bioavailability (40). Descriptors nArNR2 and N-071 represent tertiary aromatic amines and aromatic amines, respectively. Drugs with aromatic amines can be readily absorbed through the gastrointestinal tract (41,42). Beta-Lactams (descriptor nBeta-Lactams) tend to have low %F due to their low lipophilicity which makes it difficult to passively diffuse across the intestinal membrane. Beta-lactams that have high %F are typically transported by intestinal influx transporters like Peptide Transporter 1 (43), which increases %F.

Some features, which are considered to be important for bioavailability, represent complex mechanisms. For example, N-alkylation (refer to the descriptor nN+) is a common procedure used to increase the aqueous solubility of drug molecules which have low bioavailability, such as Bupivacaine (%F=0%) (44). However, this also reduced lipophilicity and the net effect on the oral bioavailability is hard to measure. In our dataset this descriptor was considered to be relevant to low bioavailability since this feature was found mostly in low bioavailable compounds. The arylsulfonamide moiety (represented by the descriptor nSO2N) was associated with high oral bioavailability. A similar fragment descriptor was also identified by CASE Ultra as the top biophore. There were 23 drugs in our dataset that contained this structural feature and their average %F was 77%. Examples of these drugs are shown in Table V. Methods for improving the oral bioavailability of sulfonamides have been studied for many decades. Previous studies found that the nitrogen atom of this fragment (as shown in Table V) plays an important role in binding to the receptor and is critical to membrane permeability and bioavailability (45,46). However, the potential mechanisms that are relevant to the bioavailability of sulfonamides are still not well understood.

Cytochrome P450 (CYP) enzymes have a crucial impact on the metabolic stability of a drug (47). Some descriptors in Figure 9, such as the number of hydrogens (nH), hydrogens attached to sp^3 carbon atoms (H-052), and number of sp^3 carbon atoms (nCsp3), were found to be correlated with low bioavailability. It was reported that CYP enzymes hydroxylate the C-H bond on sp^3 carbon atoms (47). Thus, these three descriptors may represent the structural features with low metabolic stability. Interestingly, halogenated hydrocarbons are also susceptible to oxidative dehalogenation by CYP enzymes (47). However, our descriptor analysis shows that aromatic halogens are related to high bioavailability (likely via enhancing membrane permeability). This relationship could be further explored in the future.

CONCLUSION

In this study, we first compiled a database containing 995 unique human oral bioavailable drugs. The diverse drugs in this data set include molecules with both low and high bioavailability. We harmonized the %F values and evaluated all chemical structures to ensure that the data in our database is accurate.

The bioavailability database was used to develop both CTG and CNT models by using various modeling approaches. The consensus predictions show better performance than individual models for both CTG and CNT models. Although the results of CNT models are relatively poor, we were able to use HIT parameters to improve the model prediction accuracy. Correctly using HIT parameters based on the transport direction allowed us to remove some compounds with high predictions errors. Efflux transporters that transport drugs out of the enterocytes can limit the oral bioavailability of their drug-substrates. In this study we found that the two efflux transporters, MDR1 and MRP2, were important for enhancing the oral bioavailability predictions in our models.

All of the models developed in this study can be used to evaluate the bioavailability of new drug candidates. The analysis of the important descriptors in the resulting models showed the relationships between several types of chemical structures and drug oral bioavailability. This type of knowledge could be useful for designing new drug molecules with suitable oral bioavailability. The use of HIT parameters was beneficial to the model predictions. We have confirmed that HITs need to be a component in future bioavailability models. Future directions of *in silico* oral bioavailability modeling should also take into consideration interactions with the CYP enzymes. Similar methods could be developed and employed to model other complex bioactivities of drugs and drug-like molecules.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

ABBREVIATIONS

%F	Oral bioavailability
AD	Applicability domain
ANOVA	Analysis of variance
CCR	Correct classification rate (balanced accuracy)
CNT	Continuous activity scale
CPT	Consensus prediction threshold
Combi-QSAR	Combinatorial quantitative structure-activity relationship
CTG	Category activity scale
CYP	Cytochrome P450
D	Dragon descriptors
HIT	Human intestinal transporter
***k*NN**	*k n*earest neighbor
MAE	Mean absolute error
MDR1	Multidrug resistance protein 1 (P-gp, ABCB1)
MOE	Molecular Operating Environment
MPOI	Mean probability of interaction
MRP2	Multidrug resistance-associated protein 2 (ABCC2)
POI	Probability of interaction
QSAR	Quantitative structure-activity relationship
R^2	Coefficient of determination

| **RF** | Random forest |
| **SVM** | Support vector machine |

REFERENCES

1. Buxton, IL.; Benet, LZ. Chapter 2. Pharmacokinetics: The Dynamics of Drug Absorption, Distribution, Metabolism, and Elimination. In: Brunton, LL.; Chabner, BA.; Knollmann, BC., editors. Goodman & Gilman's The Pharmacological Basis of Therapeutics. 12th ed.. New York: McGraw-Hill; 2011.

2. Andrews CW, Bennett L, Yu LX. Predicting human oral bioavailability of a compound: development of a novel quantitative structure-bioavailability relationship. Pharm Res. 2000; 17:639–644. [PubMed: 10955834]

3. Moda TL, Montanari CA, Andricopulo AD. Hologram QSAR model for the prediction of human oral bioavailability. Bioorgan Med Chem. 2007; 15:7738–7745.

4. Ma CY, Yang SY, Zhang H, Xiang ML, Huang Q, Wei YQ. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method. J Pharmaceut Biomed. 2008; 47:667–682.

5. Tian S, Li Y, Wang J, Zhang J, Hou T. ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in human based on molecular properties and structural fingerprints. Mol Pharm. 2011; 8:841–851. [PubMed: 21548635]

6. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliver Rev. 2001; 46:3–26.

7. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem. 2002; 45:2615–2623. [PubMed: 12036371]

8. Hou T, Wang J, Zhang W, Xu X. ADME evaluation in drug discovery. 6. Can oral bioavailability in humans be effectively predicted by simple molecular property-based rules? J Chem Inf Model. 2007; 47:460–463. [PubMed: 17381169]

9. Martin YC. A bioavailability score. J Med Chem. 2005; 48:3164–3170. [PubMed: 15857122]

10. Varma MVS, Obach RS, Rotter C, Miller HR, Chang G, Steyn SJ, El-Kattan A, Troutman MD. Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. J Med Chem. 2010; 53:1098–1108. [PubMed: 20070106]

11. Paixão P, Gouveia LF, Morais JAG. Prediction of the human oral bioavailability by using in vitro and in silico drug related parameters in a physiologically based absorption model. Int J Pharm. 2012; 429:84–98. [PubMed: 22449410]

12. Thummel, KE.; Shen, DD.; Isoherranen, N. Appendix II. Design and Optimization of Dosage Regimens: Pharmacokinetic Data. In: Brunton, LL.; Chabner, BA.; Knollmann, BC., editors. Goodman & Gilman's The Pharmacological Basis of Therapeutics. 12th ed.. New York: McGraw-Hill; 2011.

13. Hou T, Li Y, Zhang W, Wang J. Recent development of in silico predictions of intestinal absorption and bioavailability. Comb Chem High T Scr. 2009; 12:497–506.

14. Zhu J, Wang J, Li Y, Hou T. Recent progress of in silico predictions of oral bioavailability. Comb Chem High T Scr. 2011; 14:362–375.

15. CASE Ultra 1.4.4.6 32 bit.

16. PubChem. http://pubchem.ncbi.nlm.nih.gov/

17. Benet LZ, Broccatelli F, Oprea TI. BDDCS applied to over 900 drugs. AAPS J. 2011; 13:519–547. [PubMed: 21818695]

18. Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

19. Vapnik, VN. In the Nature of Statistical Learning Theory. New York: Springer; 2000.

20. Kovatcheva A, Golbraikh A, Oloff S, Xiao TD, Zheng W, Wolschann P, Buchbauer G, Tropsha A. Combinatorial QSAR of ambergris fragrance compounds. J Chem Inf Comput Sci. 2004; 44:582–595. [PubMed: 15032539]

21. Kovatcheva A, Golbraikh A, Oloff S, Feng J, Zheng W, Tropsha A. QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. SAR QSAR Environ Res. 2005; 16:93–102. [PubMed: 15844445]

22. Votano JR, Parham M, Hall LM, Hall LH, Kier LB, Oloff S, Tropsha A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure information representation. J Med Chem. 2006; 49:7169–7181. [PubMed: 17125269]

23. Dalgaard, P. Introductory Statistics with R. New York: Springer; 2008.

24. Zheng W, Tropsha A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. J Chem Inf Comput Sci. 2000; 40:185–194. [PubMed: 10661566]

25. Chakravarti SK, Saiakhov RD, Klopman G. Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. J Chem Inf Model. 2012; 52:2609–2618. [PubMed: 22947043]

26. Saiakhov RD, Chakravarti SK, Klopman G. Effectiveness of CASE Ultra expert system in evaluating adverse effects of drugs. Mol Inf. 2013; 32:87–97.

27. Tropsha A, Golbraikh A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. Curr Pharm Design. 2007; 13:3494–3504.

28. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. J Comput Aided Mol Des. 2003; 17:241–253. [PubMed: 13677490]

29. Sedykh A, Fourches D, Duan J, Hucke O, Garneau M, Zhu H, Bonneau P, Tropsha A. Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. Pharm Res. 2013; 30:996–1007. [PubMed: 23269503]

30. Giacomini, KM.; Sugiyama, Y. Chapter 5. Membrane Transporters and Drug Response. In: Brunton, LL.; Chabner, BA.; Knollmann, BC., editors. Goodman & Gilman's The Pharmacological Basis of Therapeutics. 12th ed.. New York: McGraw-Hill; 2011.

31. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, Clark J, Connelly MC, Sigal M, Hodges D, Guiguemde A, Guy RK, Tropsha A. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model. 2013; 53:475–92. [PubMed: 23252936]

32. Fisher, RA. The Design of Experiments. 8th ed.. Edinburgh: Oliver and Boyd; 1966.

33. Shugarts S, Benet LZ. The role of transporters in the pharmacokinetics of orally administered drugs. Pharm Res. 2009; 26:2039–2054. [PubMed: 19568696]

34. Inoue M, Morikawa M, Tsuboi M, Ito Y, Sugiura M. Comparative study of human intestinal and hepatic esterases as related to enzymatic properties and hydrolyzing activity for ester-type drugs. Japan J Pharmacol. 1980; 30:529–535. [PubMed: 7206363]

35. Dubertret L, Alirezai M, Rostain G, Lahfa M, Forsea D, Niculae BD, Simola M, Horvath A, Mizzi F. The use of lymecycline in the treatment of moderate to severe acne vulgaris: a comparison of the efficacy and safety of two dosing regimens. Eur J Dermatol. 2003; 13:44–48. [PubMed: 12609781]

36. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904; 15:72–101.

37. Kornbrot, D. Pearson Product Moment Correlation. In: Everitt, B.; Howell, D., editors. Encyclopedia of Statistics in Behavioral Science. Hoboken, New Jersey: John Wiley & Sons; 2005.

38. Utvik TIR, Johnsen S. Bioavailability of polycyclic aromatic hydrocarbons in the north sea. Environ Sci Technol. 1999; 33:1963–1969.

39. Herman RB. Theory of hydrophobic bonding. II Correlation of hydrocarbon solubility in water with solvent cavity surface area. J Phys Chem. 1972; 76:2754–2759.

40. Birnbaum L. The role of structure in the disposition of halogenated aromatic xenobiotics Environ. Health Persp. 1985; 61:11–20.

41. Stillwell WG, Turesky RJ, Sinha R, Skipper PL, Tannenbaum SR. Biomonitoring of heterocyclic aromatic amine metabolites in human urine. Cancer Lett. 1999; 143:145–148. [PubMed: 10503894]

42. Holland RD, Gehring T, Taylor J, Lake BG, Gooderham NJ, Turesky RJ. Formation of a mutagenic heterocyclic aromatic amine from creatinine in urine of meat eaters and vegetarians. Chem Res Toxicol. 2005; 18:579–590. [PubMed: 15777097]

43. Saitoh H, Gerard C, Aungst BJ. The secretory intestinal transport of some beta-lactam antibiotics and anionic compounds: a mechanism contributing to poor oral absorption. J Pharmacol Exp Ther. 1996; 278:205–211. [PubMed: 8764353]

44. Nielsen AB, Frydenvang K, Liljefors T, Buur A, Larsen C. Assessment of the combined approach of *N*-alkylation and salt formation to enhance aqueous solubility of tertiary amines using bupivacaine as a model drug. Eur J Pharm Sci. 2005; 24:85–93. [PubMed: 15626581]

45. Sawa M, Mizuno K, Harada H, Tateishi H, Arai Y, Suzuki S, Oue M, Tsujiuchi H, Furutani Y, Kato S. Tryptamine-based human β_3-adrenergic receptor agonists. Part 3: improved oral bioavailability via modification of the sulfonamide moiety. Bioorg Med Chem Lett. 2005; 15:1061–1064. [PubMed: 15686912]

46. Wu C, Decker ER, Blok N, Li J, Bourgoyne AR, Bui H, Keller KM, Knowles V, Li W, Stavros FD, Holland GW, Brock TA, Dixon RAF. Acyl substitution at the ortho position of anilides enhances oral bioavailability of thiophene sulfonamides: TBC3214, an ET_A selective endothelin antagonist. J Med Chem. 2001; 44:1211–1216. [PubMed: 11312921]

47. Meunier B, de Visser SP, Shaik S. Mechanism of oxidation reactions catalyzed by cytochrome P450 enzymes. Chem Rev. 2004; 104:3947–3980. [PubMed: 15352783]
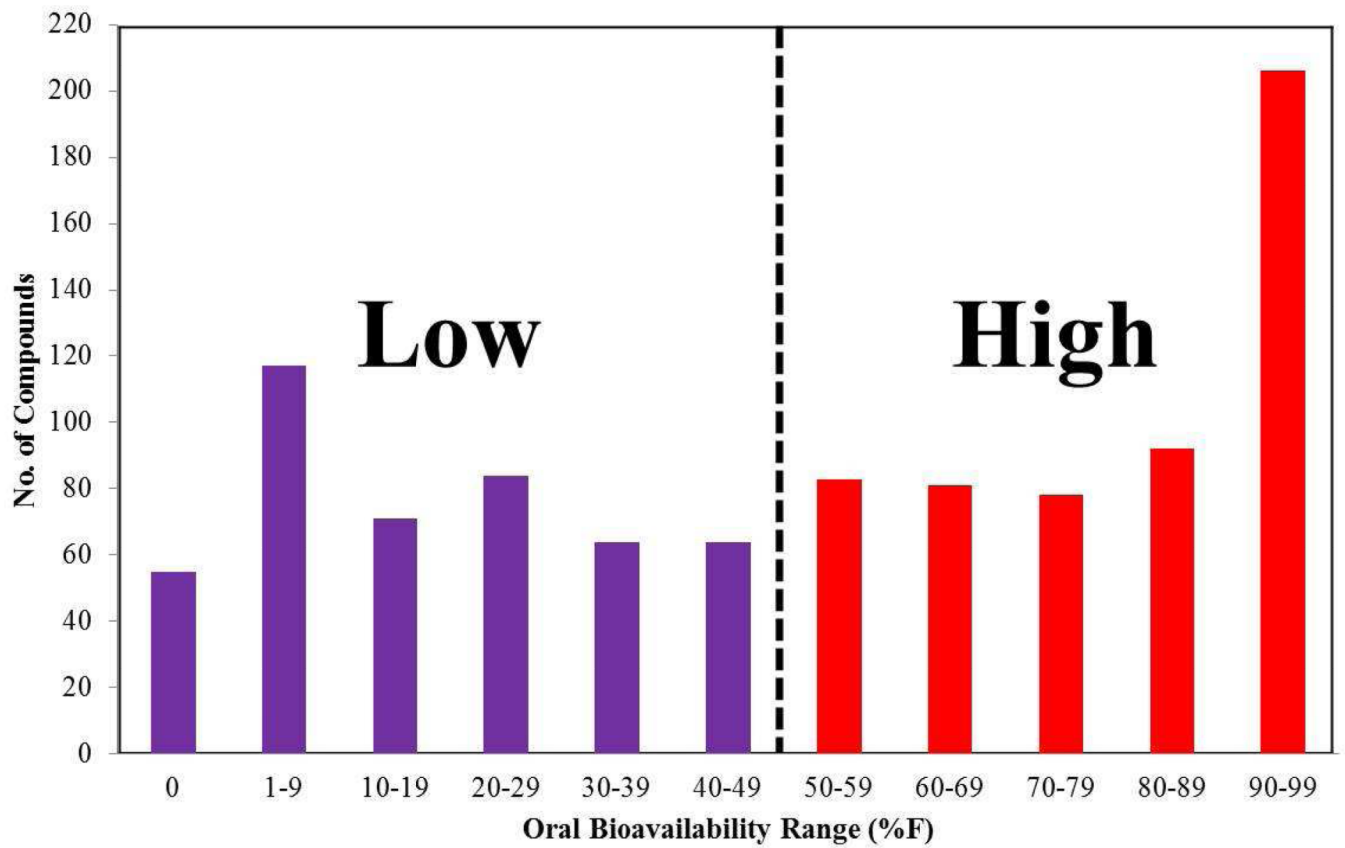
Figure 1.
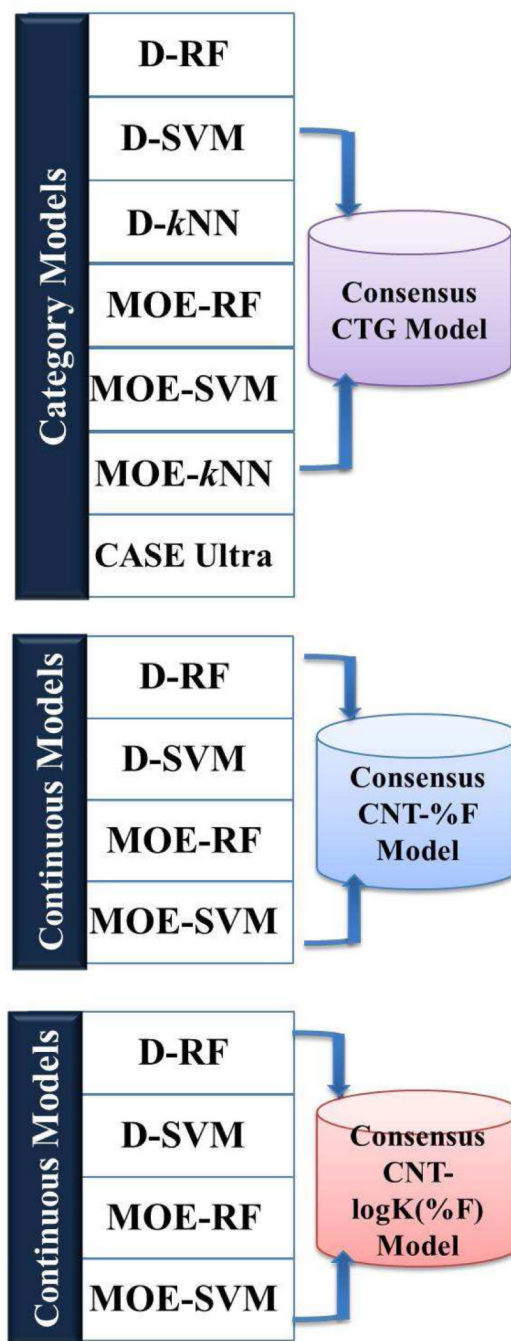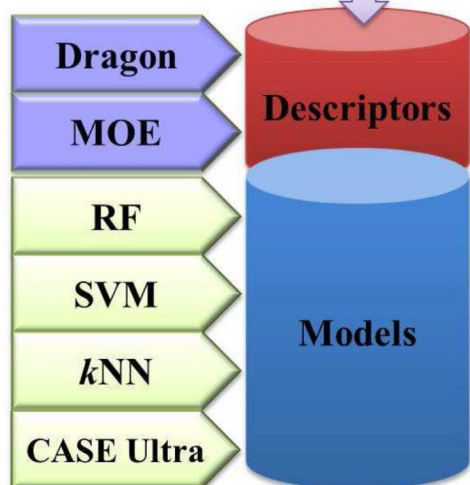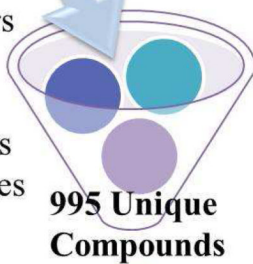Distribution of compounds by various %F ranges

Figure 2.
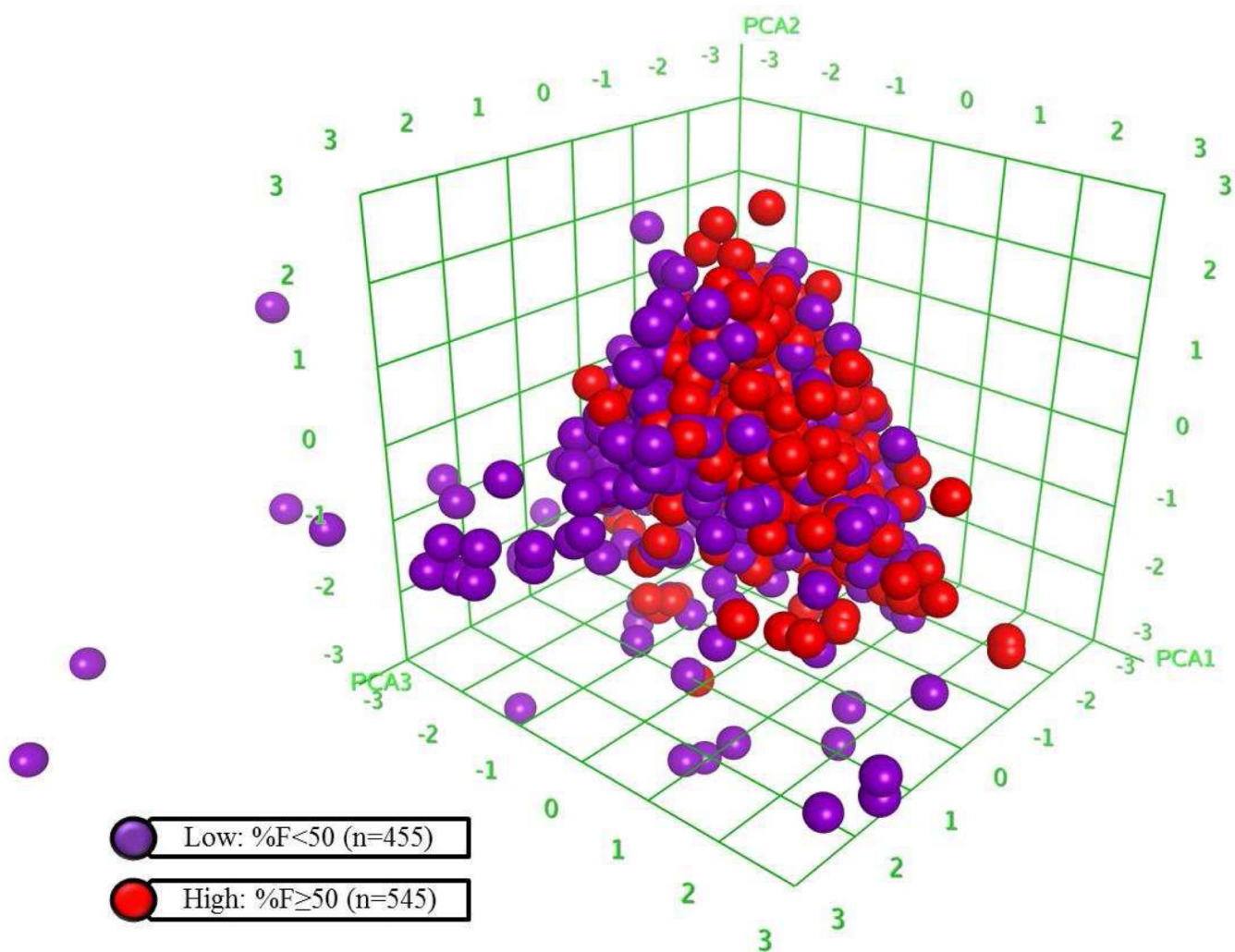Combinatorial QSAR modeling workflow

Figure 3.
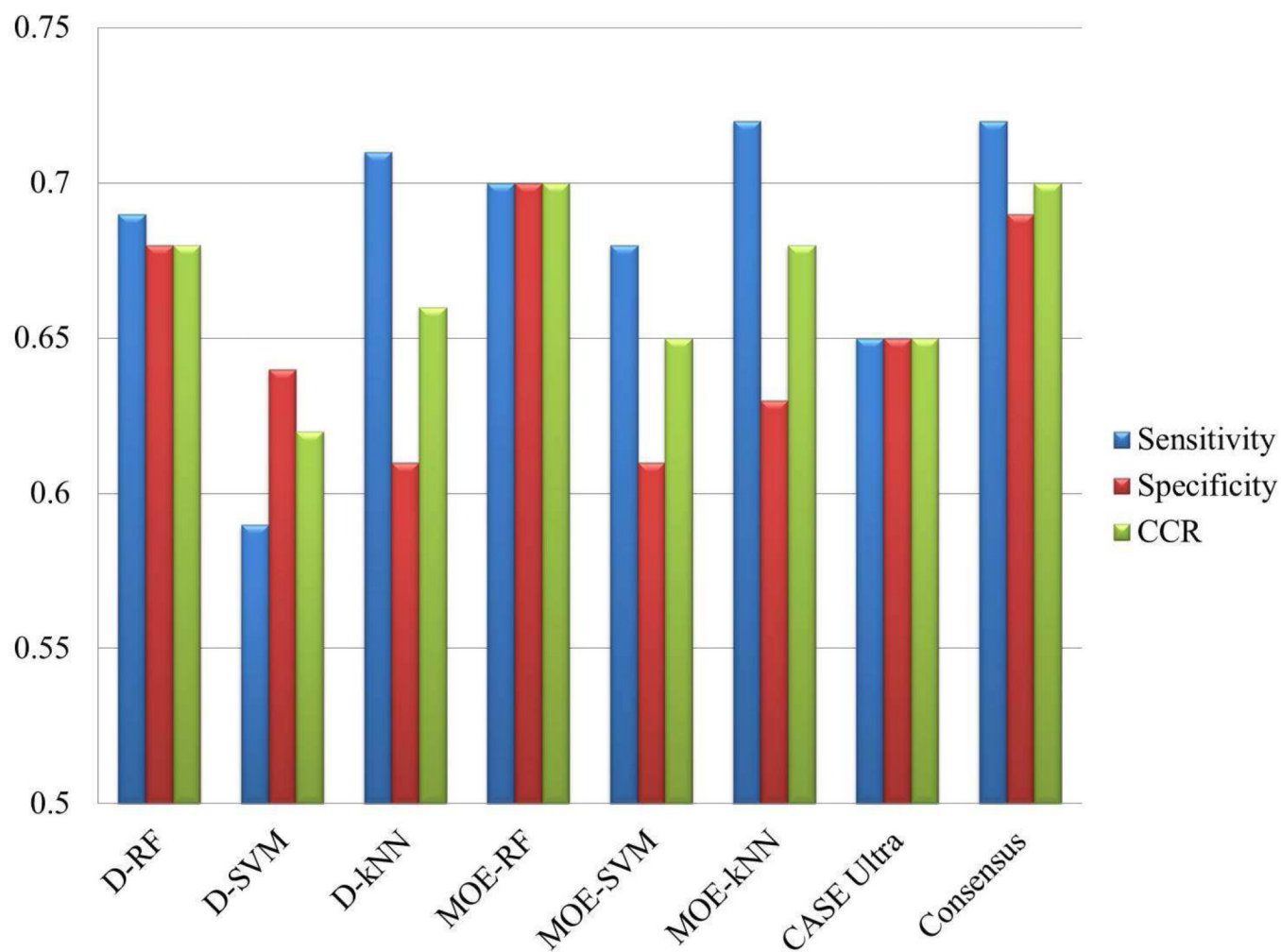Chemical space of human %F database (n=955) using top 3 principal components of MOE
descriptors

Figure 4.
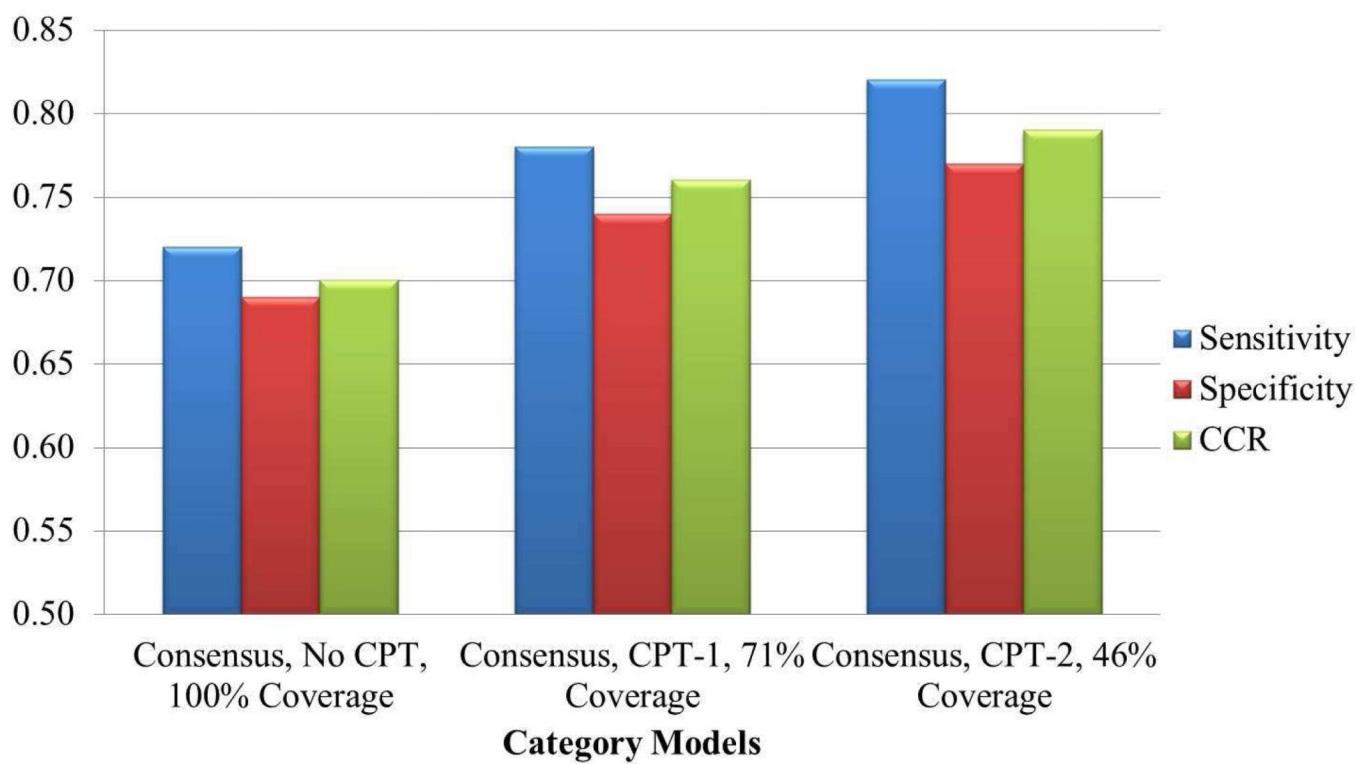Performance of CTG QSAR models using five-fold cross-validation

Figure 5.
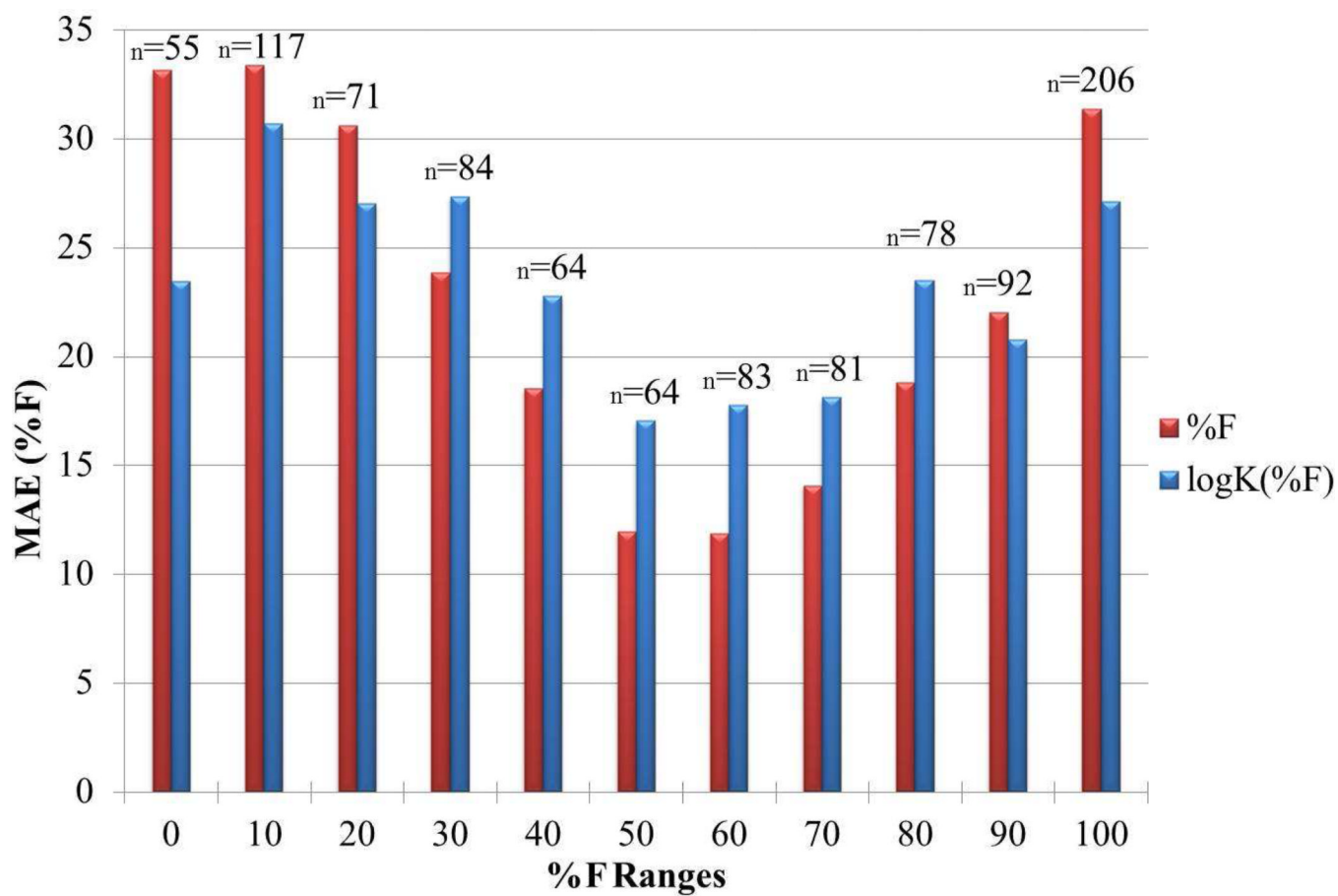Predictivity of consensus CTG model with different consensus prediction thresholds (CPT)

Figure 6.
Distribution of prediction errors (as MAE) relative to experimental %F. Red and blue bars represent consensus CNT-%F and CNT-logK(%F) models respectively.
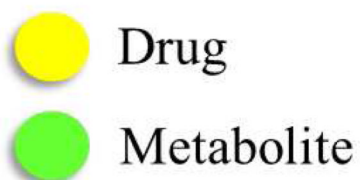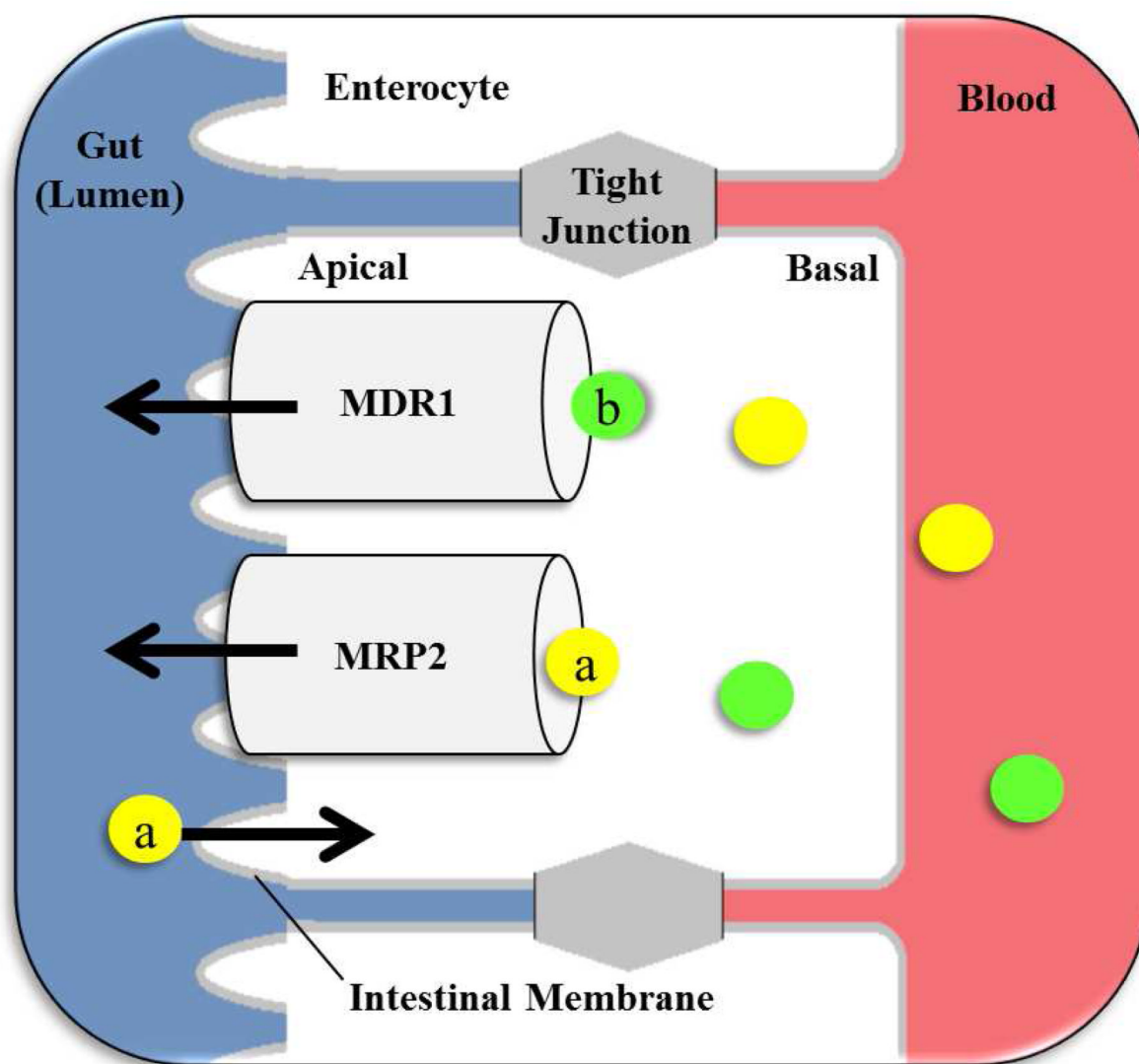
Figure 7.
Drug efflux by intestinal transporters MDR1 and MRP2 in an enterocyte; **a)** drug passively diffusing through the intestinal membrane; **b)** drug and metabolite transported out of the enterocyte.
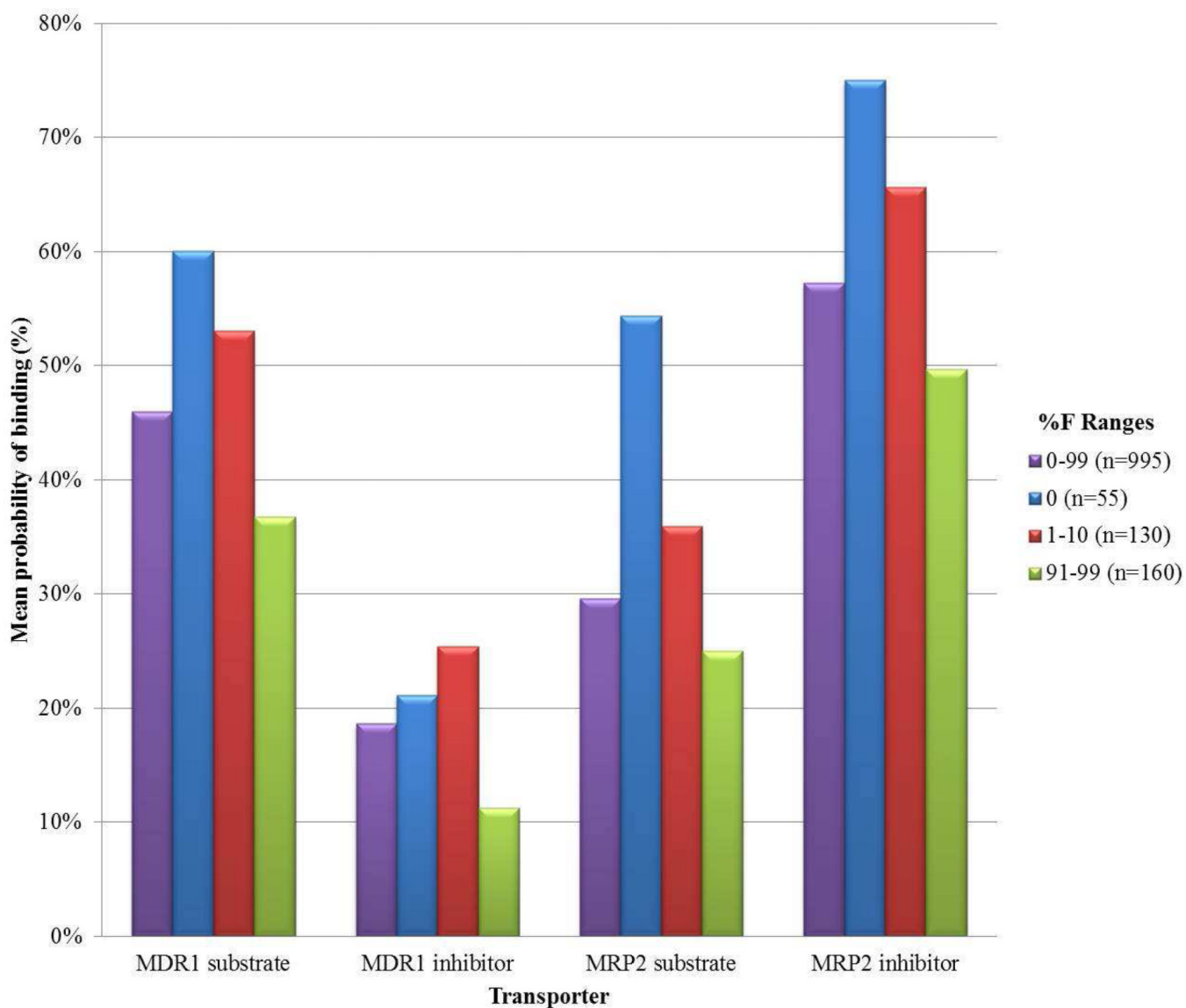
Figure 8.
Mean probability of interaction (MPOI) for compounds in specified %F ranges
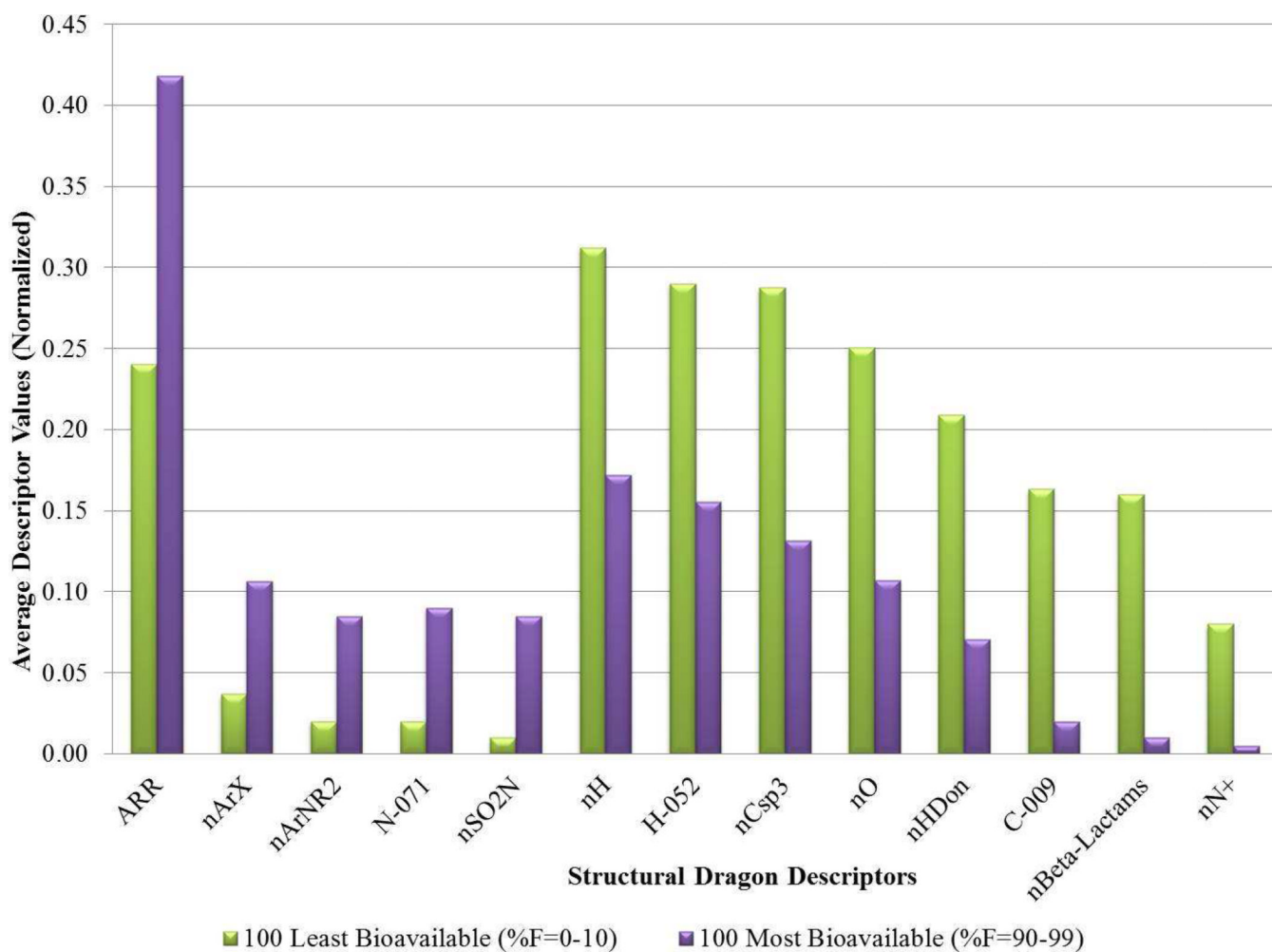
Figure 9.
Chemical structure analysis for the 100 least and 100 most bioavailable compounds

Table I

Brief description of previous QSAR oral bioavailability models

Source	Description	Performance	Train/Test Set Sizes
3	Hologram QSAR, CNT, modeling %F	$q^2 = 0.35-0.70 / R_{ext}^2 = 0.85$	250/52 (mostly highly bioavailable drugs)
5	Combinatorial QSAR, CNT, modeling %F	$R_{ext}^2 = 0.50$ (after removing outliers)	916/80(?)
2	Stepwise Regression; CNT: modeling %F	$R_{ext}^2 = 0.58$	473/118
4	Combinatorial QSAR, CTG, modeling: positive (%F 20), negative (%F<20)	CCR_{Train} (5-fold CV) =62%/ CCR_{Test}= 59–71%	690/76 (mostly highly bioavailable drugs)

q^2 - Cross validated correlation coefficient; R^2 - Coefficient of determination;
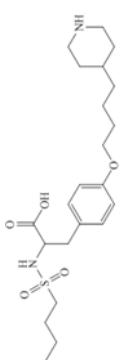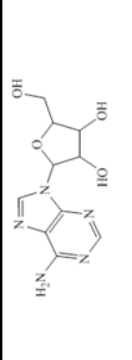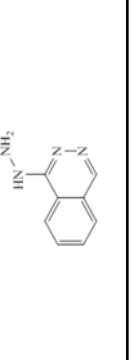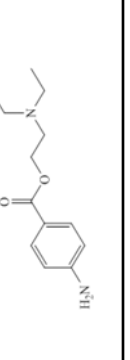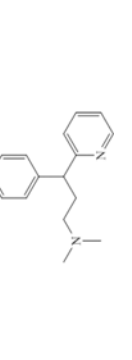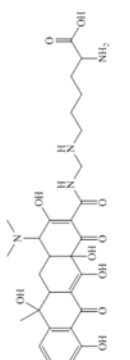
Table II

Performance of individual and consensus CNT models using a five-fold cross-validation (n=995)

Models		CNT-%F					CNT-logK(%F)				
		D-RF	**D-SVM**	**MOE-RF**	**MOE-SVM**	**Consensus**	**D-RF**	**D-SVM**	**MOE-RF**	**MOE-SVM**	**Consensus**
R²		0.30	0.13	0.29	0.19	0.28	0.30	0.11	0.28	0.14	0.25
Error Analysis	**Mean Absolute Error (%F)**	51.33	52.65	23.88	52.66	24.05	22.93	28.38	23.09	28.28	24.02
	Standard Deviation (%F)	16.51	26.02	17.16	25.37	18.46	22.77	31.66	23.79	32.73	26.4
Two Way ANOVA Test α=95%	$F_{significance}$ **(p-value)**	1.00×10^{-78}	2.33×10^{-31}	5.33×10^{-76}	2.48×10^{-48}	6.53×10^{-71}	1.81×10^{-80}	3.60×10^{-27}	2.48×10^{-74}	3.71×10^{-34}	3.57×10^{-65}
	$F_{Calculated}$	423.96	145.48	406.2	238.21	373.65	435.44	123.73	395.44	160.25	337.87

Table III

Examples of compounds with high prediction errors that were successfully (No. 1–3) or not successfully (No. 4–6) removed after combining the interaction properties of MDR1-s, MDR1-i, MRP2-s, and MRP2-i.

No.	Compounds	Transporter Score[a]				%F	Pred %F	Structure
		MDR1-s	MDR1-i	MRP2-s	MRP2-i			
1	Tirofiban	1	1	1	1	0	54	
2	Vidarabine	1	1	1	1	0	36	
3	Hydralazine	1	1	1	1	23	74	
4	Procaine	0	0	0	0	0	66	
5	Pheniramine	0	0	0	0	99	34	
6	Lymecycline	1	1	1	1	99	7	

[a]Transporter score: 0 (POI<MPOI) and 1 (POI>MPOI)

Table IV

Performance of consensus CTN-%F model with and without integrating HIT POI

Models		Consensus	Combinations of Consensus with HIT POI				
			MDR1-s	MRP2-s	MRP2-s & MRP2-i	MRP2-i & MDR1-s	MDR1-s, MDR1-i MRP2-s, & MRP2-i
R		0.28	0.31	0.26	0.36	0.40	0.40
n		995	558	758	450	362	304
Coverage		100%	56%	76%	45%	36%	30%
Error Analysis	Mean Absolute Error	24.05	22.97	23.83	21.54	21.61	21.00
	Standard Deviation	16.05	16.35	16.38	15.71	16.13	16.11
Two-Way ANOVA Test $\alpha=95\%$	$F_{significance}$ (p-value)	6.53×10^{-71}	3.42×10^{-48}	2.74×10^{-57}	6.67×10^{-45}	6.66×10^{-41}	6.55×10^{-35}
	$F_{Calculated}$	373.65	259.48	259.50	248.83	232.89	197.85
Bootstrap Non-Parametric Permutation N=10,000 $\alpha=95\%$	Spearman (p value)	N/A	0.33	0.74	0.08	0.06	0.20
	Pearson (p value)	N/A	0.17	0.17	0.04	0.01	0.02

Table V

Examples of compounds with the arylsulfonamide structural feature found using the CASE Ultra model