Published in final edited form as:
Adv Drug Deliv Rev. 2015 June 23; 86: 11–16. doi:10.1016/j.addr.2015.01.001.

Advances in Computationally Modeling Human Oral Bioavailability

Junmei Wang[a,*] and **Tingjun Hou**[b,c]

[a]Department of Biochemistry, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd. Dallas, TX 75390

[b]Institute of Functional Nano and Soft Materials (FUNSOM), Jiangsu Key Laboratory for Carbon-Based Functional Materials and Devices and Collaborative Innovation Center of Suzhou Nano Science and Technology, Soochow University, Suzhou, Jiangsu 215123, P. R. China

[c]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

Abstract

Although significant progress has been made in experimental high throughput screening (HTS) of ADME (absorption, distribution, metabolism, excretion) and pharmacokinetic properties, the ADME and Toxicity (ADME-Tox) *in silico* modeling is still indispensable in drug discovery as it can guide us to wisely select drug candidates prior to expensive ADME screenings and clinical trials. Compared to other ADME-Tox properties, human oral bioavailability (HOBA) is particularly important but extremely difficult to predict. In this paper, the advances in human oral bioavailability modeling will be reviewed. Moreover, our deep insight on how to construct more accurate and reliable HOBA QSAR and classification models will also discussed.

Keywords

Human Oral Bioavailability (HOBA); Human Intestinal Absorption (HIA); ADME-Tox; QSAR; *In Silico* Modeling Computer-Aided Drug Design

1. Introduction

It is estimated that the entire chemical space exceeds 10^{60} molecules, and it is impossible to synthesize all of them given the fact that the total weight of earth is only about 6.0×10^{27} grams. As a matter of fact, there are only 27 million compounds have been registered.[1] Even though the synthesized compounds only occupy a tiny fraction of the entire chemical space, it is much larger than the biological chemical space due to the fact that there are a few thousands of small molecules within our own bodies. As the biological chemical space only represent an amazingly small fraction of the entire chemical space, it is understandable that

[*]**Corresponding author**: Junmei Wang, junmei.wang@utsouthwestern.edu, **Phone**: +1-214-6484146.

to discover small molecules that efficiently interact with protein targets is a very difficult task. Although numerous new technologies, such as combinatorial chemistry, high throughput screening and computer-aided drug design have been applied to facilitate the discovery of new drugs, the number of new molecular entities approved annually by FDA (U.S. Food and Drug Administration) has not changed significantly in the last two decades. What are the major reasons that cause the attrition of drug candidates during clinical trials? The lack of efficiency and poor ADME-Tox (absorption, distribution, metabolism, excretion, and toxicity) and pharmacokinetics are responsible for most of the drug attrition.[2]

Among the many ADME-Tox/PK properties, bioavailability is particularly important for the orally administered drugs. Today, high throughput screenings of human oral bioavailability (HOBA) are routinely conducted in pharmaceutical companies. However, the *in vitro* and *in vivo* assays are much time consuming and costly. Only a tiny fraction of synthesized and screened compounds are selected to do the analysis. *In silico* HOBA modeling, on the other hand, is much more efficient and can deal with large screening libraries. Moreover, *in silico* HOBA models can serve as drug likeness filters to prioritize screening libraries. Those filters typically have better discriminative power than the conventionally used drug likeness filters, like Lipinski's 'Rule of Five'.[3] It is a trend that *in silico* ADME-tox models, particularly HOBA, are incorporated into the paradigm of drug lead identification and optimization procedures.[4–16]

1.1 ADME-Tox

As one of the hot fields in computer-aided drug design (CADD), numerous reviews have been published recently on the progress of ADME-Tox modeling,[17–21] here in this paper we only focus on the latest advances of *in silico* modeling of human oral bioavailability. ADME-Tox properties which can be broadly classified into two categories, namely, the "physicochemical" and "physiological". The physicochemical properties, including aqueous solubility, logarithm of octanol – water partition coefficient (*logP*), logarithm of octanol – water distribution coefficient (*logD*) and pKa, etc., are governed by simple physicochemical laws. The physiological ADME-Tox properties can be further grouped into in vitro ADME-Tox properties (such as Caco-2 permeability and MDCK permeability, liver microsomes, etc.) and in vivo pharmacokinetic properties (such as oral bioavailability, human intestinal absorption – HIA, plasma protein binding – PPB, urinary excretion, area under the plasma concentration – time curve (AUC), total body clearance (Cl), volume of distribution, elimination half time (t1/2), etc.). As physiological ADME-Tox properties, particularly oral bioavailability, are governed by many factors, it is a very challenging task to adequately model and accurately predict the physiological ADME-Tox properties.

1.2 human oral bioavailability

Oral bioavailability (OBA) is one of the most important pharmacokinetic properties in drug discovery. As the oral form is the most convenient way to administrate a drug, it is not a surprise that about 80% of the dosage forms in the worldwide market are administrated orally.[22] OBA represents the percentage of an oral dose that is available to produce pharmacological actions. In practice, OBA is defined as the fraction of the oral does that reaches the system circulation in an active form and measured by the ratio of the dosecorrected AUC (area under curve) of the oral route to that of the intravenous route. For an oral drug, the amount of the active form that reaches the system circulation is reduced due to incomplete absorption in gastrointestinal track and the first-pass metabolism. Therefore, oral bioavailability is ranged from 0 to 100%.

It is a very challenging task to adequately model and accurately predict the oral bioavailability of a drug because this physiological property is a complex function of many biological and physicochemical properties, which include the aqueous solubility of the drug in the gastrointestinal tract, the intestinal membrane permeability, and the extend of the first-pass metabolism occurred in liver, gut and intestine, and even the dosage form of the drug. Moreover, the measurement of oral bioavailability of a drug is affected by other factors like whether the drug is taken with or without food, whether other drugs are taken concurrently, as well as the disease states, etc. Those factors may alter the drug absorption, the liver metabolism. For example, the oral bioavailability of patients with liver disease may be increased due to the reduced liver metabolism. The above factors may vary from patient to patient and from time to time for the same patient. This complicate picture explains why the measurement errors of oral bioavailability are very large. According to the survey of 367 drugs conducted by Wang et al.,[23], the average unsigned error and root-mean square error of the experimental measurements are 12.1 and 14.5%, respectively.

In order to develop an oral drug with high bioavailability, medicinal chemists apply a simple rule to select drug candidates: those having high aqueous solubility and high membrane permeability tend to have high OBA; those having low aqueous solubility and low permeability tend to have poor OBA; and the others might need careful formulation to improve their dissolution or absorption rate. This simple rule is based on the fact that drug dissolution and permeability control the rate and extent of drug absorption in GI track. Certainly, a drug with high oral bioavailability should also be largely free from fast-pass metabolism.

1.3 *In silico* models of HOBA prediction

Attempts have been made to predict HOBA back to Year 2000 by Andrews, Bennett and Xu,[24] and Yoshida and Topliss.[25,26] Later on, numerous models were published[25,27,28] and reviewed by ourselves,[29] and others.[30] The following is a brief summary of HOBA models developed prior to 2008: most models were developed using relatively small data sets (n < 600) and they merely make reliable prediction for the compounds in the screening libraries. For the classification models developed before 2008, the rates of the correct assignment are usually lower than 70%; for the QSAR models, the RMSE are ranged from 24 to 30%. In the following, we will present reviews on the latest HOBA models.

2. Recent Advances in HOBA Modeling

In 2008, a classifier was developed by Ma et al. with GA (genetic algorithm)–CG (conjugated gradient)–SVM (support vector machine) method for 866 compounds that have human oral bioavailability data.[31] GA was applied to select descriptors that were calculated using Cerius 2 software package (http://www.accelyrs.com), while SVM was used to construct classification model and CG was applied to optimize the parameters of kernel functions of SVM. The predict accuracy, 80% for the training set (690 compounds) and 86% for the test set (76 compounds) is encouraging. However, the classifier has poor performance for the "negative" class: the prediction accuracy is only 44% and the false positive (FP) is even larger than true negative (TN). This phenomenon can be explained by that fact that a very small cutoff of 20% was applied to assign 'positive' and 'negative' classes. Even the prediction accuracy is good, it cannot be used to further discriminate the compounds belong to the "positive" class.

In 2009, a set of predictive models for human bioavailability were developed by Imawaka et al. using the human oral administration data and animal pharmacokinetic data as descriptors.[32]

$$OBA=\frac{AUC_{po}/Dose_{po}}{AUC_{iv}/Does_{iv}}=\frac{CL_{tot}}{CL_{po}}=\frac{\beta\times Vd_{\beta}}{CL_{po}} \quad (1)$$

Where AUC_{po} and AUC_{iv} are the areas under the plasma concentration-time curves for the oral administration and intravenous administration, respectively; $Dose_{po}$ and $Dose_{iv}$ are the doses of the two types of administrations accordingly; Vd_{β} and β are the distribution volume of the terminal phase and the elimination rate constant, respectively. Vd_{β} is estimated with the animal data using Eq. (2).

$$Vd_{\beta}\approx Vd_{ss}=V_B\times R_B+f_P\times V_E+(1-f_P)\times R_{E/I}\times V_B\times(1-H_C)+V_T\times f_P/f_T \quad (2)$$

Where f_P and f_T are the free fraction of drug in plasma and tissue; V_B and V_E, the volumes of blood and the extracellular fluid, are assumed to be 80 ml/kg and 260 ml/kg, respectively; R_b, the blood to plasma concentration ratio is assumed to be 0.945 if not available; V_T represents the volume the tissue into which the drug is distributed; $R_{E/I}$, which is set to 1.4, is the ratio of the amounts of binding protein in the extracellular fluid to that in plasma; H_c, the hematocrit value, is set to 0.42; After taking the above assumed constants into Eq. 2, Eq. 2 becomes Eq. 3.

$$Vd_{\beta}=f_P\left[\frac{V_T}{f_T}\right]+80R_B+195f_P+65 \quad (3)$$

The authors proposed three methods to estimate $[V_T/f_T]$ using the animal values of Vd_{β}, R_B and f_P. In Method 1a, the $[V_T/f_T]$ of human is assumed equal to that of animal; in Method 1b, Eq. 4 is used to calculate the $[V_T/f_T]$ of human:

$$\left[\frac{V_T}{f_T}\right]_{human}=0.6628\left[\frac{V_T}{f_T}\right]_{rat}^{1.096} \quad (4)$$

Method 2 is similar to Method 1a except that the $[V_T/f_T]_{\mathrm{animal}}$ values come from not only rat but also other animals. Once $[V_T/f_T]_{\mathrm{human}}$ is obtained, then human Vd_{β} can be calculated with Eq. 3 using human $[V_T/f_T]$, f_P and R_B.

The performance of the three calculation protocols was summarized in Table 1. The *AUE* and *RMSE* of Protocol 3 (using Method 2 to calculate $[V_T/f_T]_{\text{human}}$) is 10.7 and 15.4%, respectively. Unfortunately, the statistic parameters (*AUE* and *RMSE*) were calculated using a small set of data and the reliability may be questionable. The RMSE of the Protocol 1 (using Method 1a to calculate $[V_T/f_T]_{\text{human}}$) and Protocol 2 (using Method 1b to calculate $[V_T/f_T]_{\text{human}}$) are 20.9 and 23.2, respectively.

In 2011, we have performed a systematic study on understanding how various molecular descriptors correlate to HOBA and developed a set of classification and QSAR models to predict HOBA using a large database of 1014 compounds.[33] We could not find any property-based rule which has sufficient discriminative power to serve as a predictor for HOBA. We then constructed a set of multiple linear regression models using genetic function approximation. The best model has achieved a very encouraging performance in modeling HOBA: R = 0.79, Q_{LOO} = 0.72, *RMSE* = 22.3 % for the training set; R_{test} = 0.71 and $RMSE_{test}$ = 23.6% for the test set.

In 2012, a set of HOBA models were developed by Paixão, et al.[34] using a set of *in vitro* and *in silico* physiological properties as descriptors, including absorption and solubility at the gastrointestinal pH range 1.5–7.5, apparent permeability - P_{app}, and intrinsic clearance - Cl_{int}.[34] The authors divided the whole data set into four sets according to the availability of *in vitro* data of P_{app} and Cl_{int}. For the 49 drugs for which both *in vitro* P_{app} and Cl_{int} are available, a computer model with an excellent predictive ability was constructed: *RMSE*=16.0%, 84% of data within ±20% and 96% within ±35% error margins. Unfortunately, when only *in silico* data of P_{app} and Cl_{int} were applied, the performance of the computer model is much worse: RMSE=34.6.0%, 53% of data within ±20% and 74% within ±35% error margins. As suggested in Table 1, apparent permeability is a more important descriptor than intrinsic clearance for modeling HOBA.

Recently, Xu et al.[35] constructed a set of QSAR models using our HOBA database.[36] The descriptors of those models, which include constitutional and topological descriptors, walk and path counts, connectivity indices, etc., were calculated using Dragon.[37] The best-performed model, which was constructed by SVM-regression has R^2_{test} and $\text{RMSE}_{\text{test}}$ of 0.80 and 31%, respectively. Apparently, the predictability of this model is not strong. Other linear models constructed by multiple linear regression and partial-least-square fitting perform even worse.

In 2012, a set of classifiers were developed by Ahmed and Ramakrishnan for a large dataset of HOBA.[38] To get the optimal descriptors, the authors first constructed classifiers for both HIA and Caco-2 permeability using both the physiochemical and structural properties as descriptors. The underlying descriptors that are effective in discriminating between distinct classes were then identified by partial least squares discriminant analysis. 47 descriptors which are common for both HIA and Caco-2 were used to construct classifiers for HOBA. The best performed classifier, the logistic classifier, achieves a classification accuracy of 71%. However, the authors didn’t further evaluate the classifier using external test set.

In 2014, Olivares-Morales et al. developed a set of classification models for HOBA using animal oral bioavailability data.[39] The HOBA data were classified into two groups (high and low) using a cutoff of 50%. Optimal OBA_{animal} thresholds were identified for mouse (67%), rat (22%), dog (58%) and non-human primates (35%) through ROC (receiver operating characteristics) analysis. The performance of the classification models are summarized in Table 2. It is not a surprise that the model based on the oral bioavailability of non-human primate has a very high predictability.

In summary, there is some progress on the *in silico* modeling of human oral bioavailability in recent years: (1) the database has been expanded from ~600 to about ~1000, more high quality data make it possible to develop more reliable models; (2) more experimental pharmacokinetic data, either from human or animals were used to construct models for HOBA; (3) for QSAR models using descriptors purely based on molecular structures, the best QSAR model could predict HOBA for the external dataset with an RMSE of 23–24%, marginally better than models developed before 2008; (4) as to the classification models using common molecular structure-based descriptors, the best successful rate is about 70%, not much difference from the models constructed before 2008. The representative HOBA QSAR models and classifiers were summarized in Table 1.

3. Discussion

In the last section, several representative HOBA models developed in last six years were reviewed. It is obvious that more effort is needed to develop more accurate and predicative models for this important ADME/PK property. What are the strategies to achieve the goal of developing high quality HOBA models?

3.1 Database construction

First of all, more high quality human oral bioavailability data is needed. It is not an easy task to collect a large number of HOBA database since one needs AUC for both the oral and intravenous administrations to measure HOBA (Eq. 1). However, the intravenous does is not always available due to safety and solubility reason. This situation becomes even severer by the fact that the HOBA data which are owned by the pharmaceutical companies industry, usually do not exist in public domain.

A big advantage of in-house database owned by pharmaceutical companies is that the data are consistent and have less variation. The following are some publications on HOBA using in-house data sets: an analysis of bioavailability performed by Veber et al. in with a dataset containing over 1100 compounds owned by GSK;[40] a study of oral bioavailability for 591 structures from the GlaxoWellcome's internal database by Andrew et al.[24] Unfortunately, these in-house data are usually not available for the public scientific community. On the other hand, the HOBA data in the public domain may lack of self-consistency as the HOBA data usually come from more than one source and significant variability might exist between different sources.

To successfully model HOBA, it is critical to collect a large amount of reliable and consistent experimental data. We have put a lot of effort to construct the HOBA database. In

2006, we collected HOBA data for 577 drugs from Goodman and Gilman's the Pharmacological Basis of Therapeutics, both the 9th and 10th editions. In 2007, the database was expanded to 805 and some experimental data were updated using the latest literatures.[36] Now the third version of the HOBA database has collected human oral bioavailability for more than 1000 drugs. The latest database is accessible from http://cadd.suda.edu.cn/admet/. To achieve high reliability and consistency of the database, a great effort was put to verify the newly collected data. Suspicious data were recognized when there are large root-mean-square deviations of the HOBA data for the same drugs. We also pay extra attention to entries that have large prediction errors by our HOBA models. This strategy was successfully used by ourselves to verify the aqueous solubility data in the Beilstein dataset.[15]

3.2. Descriptors

To successfully model human oral bioavailability, another key factor is to select suitable molecular descriptors. There are thousands of descriptors available, not only for 1D, but also for 2D and 3D. Lots of descriptors are redundant and should be eliminated using covariance analysis. The widely used molecular descriptors include Abraham,[41] Volsurf,[42] Molsurf,[43] Adriana.Code,[44] FAF-Drugs2,[45,46] Dragon,[37] Mold,[47] Cerius 2 (www.accelrys.com), Molconn-Z (www.tripos.com), and so on. To wisely select proper molecular descriptors to model a property, one needs know some basic knowlege on that property. As to HOBA, it is a function of f_a, the fraction of drug does that is absorbed in the gastrointestinal tract and F_G, the fraction of drug dose that escape first pass metabolism in the gastrointestinal tract and F_H, the fraction of drug does that escape first pass metabolism in the liver.

$$OBA = f_a \times F_G \times F_H \quad (5)$$

We recently studied the relationship between HOBA and HIA (human intestinal absorption) for 510 compounds that have both values measured. The correlation coefficient between HOBA and HIA is 0.63. In these 510 compounds, 182 show a significant difference between HOBA and HIA (HIA – HOBA ≥20). Those compounds are considered to have a significant first-pass metabolism rate while the others do not. The relationship between HOBA and HIA is shown in Fig. 1. It is obvious that most compounds have HIA larger than or equal to HOBA (on or above the red line). Similar result was found in our previous publication.[33] Given the fact that HIA is highly correlated to HOBA and it defines the upper limit of HOBA, HIA data from *in vitro/in vivo* analysis can be a good descriptor for HOBA. Molecular descriptors that well describe HIA, such as *logP* and *logD* which are measures of lipophilicity, aqueous solubility, should also be good descriptors for HOBA.

As HOBA is also a function of F_G and F_H, the molecular descriptors that describe the substructures directly participating the metabolism reactions should be good descriptors for modeling HOBA. Indeed, many of the best-performed models were constructed using counts of molecular fragments, molecular fingerprints as descriptors.[23,33] Moda et al. found that the use of all atoms, bonds, connections and chirality to define molecular fragments led to a set of encouraging models in their study.[28] Although not been used, we expect that pharmacophore fingerprints, such as those produced by the Tuplets model of Sybyl,[48] could be good descriptors for HOBA.

Other experimental data of pharmacokinetic properties, such as bioavailability of animals, volume of distribution, *AUC* concentration – time curve, C_{max}, $t_{1/2}$, clearance, plasma protein binding, etc. can be valuable descriptors for HOBA. The only problem of this type of descriptors is that they are available only for a very limited set of molecules.

3.3. Model construction

In the next step, QSAR or classification models are constructed using the calculated descriptors. In order to develop statistically significant and robust models, the whole dataset is divided into a training set, a cross-validating set and a test set. The predictability of a model is objectively evaluated by the test set. In our work of modeling aqueous solubility,[15] we performed 85/15 cross-validation test (85% of data randomly selected into training set and 15% into test set) 10,000 times and we found out that Q^2, *AUE* and *RMSE* had Gaussian distributions. For example, the mean, maximum and minimum of Q^2 were 0.832, 0.884 and 0.762, respectively for the ASM-ATC-LOGP model. As to *RMSE*, the mean, maximum and minimum were 0.841, 0.959 and 0.732, respectively for the ASM-ATC-LOGP model. It is obvious that the Q^2 and $RMSE_{test}$ depends on how a test set is formed. Obrezanova et al. suggested to split a dataset into training, test and cross-validation sets through classification analysis of the 2D path-based fingerprints of molecular structures.[49] First of all, the cluster centroids and singletons automatically enter the training set; the others were sorted by the Y-values and assigned to the raining set, cross-validation set and test set randomly with the userdefined probabilities.

As listed in Table 1, a variety of statistical tools have been applied to construct QSAR models for HOBA. The following are the most commonly-used tools: multiple linear regressions (*MLR*), partial least square fitting (*PLS*), genetic algorithms (*GA*), artificial neural network (*ANN*), support vector machines-regression (*SVMR*), Gaussian processes (*GP*), etc. When a great number of descriptors are applied to model HOBA, hybrid methods, such as *GA/MLR* or *GA/PLS* or *GA/SVMR* are recommended since the hybrid methods can effectively avoid the over-fitting problem by selecting a subset of descriptors to construct QSAR models. In the fitness functions of GA, not only the performance of the model (R^2, Q^2, $RMSE_{train}$, $RMSE_{test}$) matter, but also the number of descriptors selected. The more descriptors selected, the larger the penalty. Gaussian processes,[49,50] which is able to model non-linear relationships and select important descriptors from a large pool, could have a great application on HOBA modeling. More importantly, GP is inherently to avoid overtraining and provide a way to objectively estimating the uncertainty of prediction.

As to classification models of HOBA, support vector machine, recursive partitioning, *k*-nearest neighbors, etc. are the commonly used methods. The performance of a classifier strongly depends on how to define '+' and '−' or multiple classes. There is still no consensus on which cutoffs should be used to form the classes. The following cutoffs have ever been used to construct two-class classifiers: 20% by Ma et al. (if HOBA>=20%, then the molecule belongs to the '+' class, otherwise, it is in '−' class),[31] 30% by ourselves,[36] and 50% by Olivares-Morales et al.[39] Few classifiers were published for more than two classes.[26] To construct a successful classifier, the number of data in each classes should be balanced, otherwise, one may have good sensitivity but poor specificity or vice versa. Classification sensitivity and specificity are described by numbers of true positives (*TP*), true negatives (*TN*), false positives (*FP*) and false negatives (*FN*) using the following equations: sensitivity = *TP/(TP+FN)*, specificity = *TN/(TN+FP)*.

How can we improve the prediction accuracy of HOBA for a new compound outside of the training set? The first approach is based on the concept of 'the domain of applicability' proposed by Konovalov et al.[51] and Weaver et al.[52] In this approach, the entries that have the similar descriptor profile to that of the compound to be predicted are selected and put into the training set. The models constructed using this specific training set are then used to predict the HOBA of the compound. Certainly, this approach may not be suitable in database screening if the model construction procedure takes a long time.

The second possible approach to improve prediction accuracy of HOBA is to develop consensus models by combining two or more models together. We constructed a set of consensus models to model bioavailability (30 models), plasma protein bonding (20 models) and urinary excretion (30 models).[53] All those models were constructed by *MLR* using *GA* to select descriptors, the count of molecular fragments. More reliable prediction was achieved with the consensus models for all the three ADME-Tox properties. The concept of "consensus modeling" was also applied by other researchers in ADME prediction. For example, Abshear *et al.* evaluated the performance of four QSAR models in KnowItAll (http://www.bio-rad.com) on predicting the intrinsic solubility of 113 diverse organic compounds.[54] The absolute average errors (AAE) of the four models are 0.314, 0.422, 0.327 and 0.324 log units, respectively. However, the prediction error is reduced to 0.257 log unit by using a consensus model combined with the four individual models. With more and more HOBA models published, it is possible to achieve more reliable prediction utilizing the strategy of consensus modeling.

3.4. Other Development

Besides human oral bioavailability, a set of models were constructed for predicting bioavailability administrated through other routes. For example, a spreadsheet-based model to estimate bioavailability of chemicals from dermal exposure was constructed by Dancik et al. recently.[55]

Some models were developed to predict animal bioavailability: Nomoto et al. constructed a computational model for rate bioavailability using in vitro intestinal parameters as descriptors.[56] De Buck et al. developed computer models for predicting a set of physiological properties including bioavailability for 50 structurally diverse compounds in rat.[57] Grabowski and Jaroszewski developed models to predict bioavailability of veterinary drugs.[58]

Computer models for modeling HOBA of specific types of drugs were also reported recently. Artemenko et al. constructed PLS models for a set of 362 antiviral drugs. The performance of the model with $R^2_{test} > 0.6$ is satisfactory.[59] QSPR models were developed for modeling and predicting both MEK inhibitory activity and oral bioavailability of novel isothiazole-4-carboxamidines.[60]

4. Conclusions

In this paper, the latest progress on the *in silico* modeling of human oral bioavailability is reviewed. Due to the complexity of this physiological property and the lack of the high quality and self-consistent experimental data, there is still a long way to go to reliably predict human oral bioavailability for arbitrary compounds in the screening libraries.

Acknowledgments

We are grateful to acknowledge the research support from NIH (R01GM79383, Y. Duan, P.I.) and Natural Science Foundation of China (No. 20973121, T. Hou, P.I.), and the TeraGrid for the computational time (TG-CHE090098, J. Wang, P.I.).

Abbreviations

ADME-Tox	Absorption, Distribution, Metabolism, Excretion, and Toxicity
HOBA	Human oral bioavailability
HIA	Human intestinal absorption
QSAR	Quantitative structure – activity relationship
MLR	Multiple linear regressions
GA	Genetic algorithm
LOO	Leave-one-out
$\mathbf{R^2}$	Regression coefficient
$\mathbf{Q^2}$	Cross-validation regression coefficient
AUE	Average unsigned error
RMSE	Root-mean-square error
$\mathbf{AUE_{test}}$	Average unsigned error of the test set
$\mathbf{RMSE_{test}}$	Root-mean-square error of the test set
N	Number of data points
$\mathbf{N_{test}}$	Number of data points in a test set
LogP	Logarithm of octanol – water partition coefficient
LogD	Logarithm of water distribution coefficient

References

1. Dobson CM. Nature. 2004; 432:824. [PubMed: 15602547]
2. Kola I, Landis J. Nature Reviews Drug Discovery. 2004; 3:711.
3. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Advanced Drug Delivery Reviews. 1997; 23:3.

4. Zheng MQ, Zhang XY, Zhao M, Chang HW, Wang W, Wang YJ, Peng SQ. Bioorganic & Medicinal Chemistry. 2008; 16:9574. [PubMed: 18835178]

5. Frecer V, Berti F, Benedetti F, Miertus S. Journal of Molecular Graphics & Modelling. 2008; 27:376. [PubMed: 18678515]

6. Antunes JE, Freitas MP, da Cunha EFF, Ramalho TC, Rittner R. Bioorganic & Medicinal Chemistry. 2008; 16:7599. [PubMed: 18656371]

7. Liao CZ, Karki RG, Marchand C, Pommier Y, Nicklaus MC. Bioorganic & Medicinal Chemistry Letters. 2007; 17:5361. [PubMed: 17719223]

8. Deng JX, Sanchez T, Al-Mawsawi LQ, Dayam R, Yunes RA, Garofalo A, Bolger MB, Neamati N. Bioorganic & Medicinal Chemistry. 2007; 15:4985. [PubMed: 17502148]

9. Sivakumar PM, Babu SKG, Mukesh D. Chemical & Pharmaceutical Bulletin. 2007; 55:44. [PubMed: 17202700]

10. Sengupta D, Verma D, Naik PK. In Silico Biology. 2008; 8:275. [PubMed: 19032162]

11. Sengupta D, Verma D, Naik PK. J Biosci. 2007; 32:1307. [PubMed: 18202455]

12. Da Silva VB, Andrioli WJ, Carvalho I, Taft CA, Silva C. Journal of Theoretical & Computational Chemistry. 2007; 6:811.

13. Ekins S, Mestres J, Testa B. British Journal of Pharmacology. 2007; 152:9. [PubMed: 17549047]

14. Ji HT, Stanton BZ, Igarashi J, Li HY, Martasek P, Roman LJ, Poulos TL, Silverman RB. Journal of the American Chemical Society. 2008; 130:3900. [PubMed: 18321097]

15. Wang J, Hou T, Xu X. J Chem Inf Model. 2009; 49:571. [PubMed: 19226181]

16. Wang J, Krudy G, Hou T, Zhang W, Holland G, Xu X. J Chem Inf Model. 2007; 47:1395. [PubMed: 17569522]

17. Yengi LG, Leung L, Kao J. Pharmaceutical Research. 2007; 24:842. [PubMed: 17333392]

18. Wunberg T, Hendrix M, Hillisch A, Lobell M, Meier H, Schmeck C, Wild H, Hinzen B. Drug discovery today. 2006; 11:175. [PubMed: 16533716]

19. Wishart DS. Drugs R D. 2007; 8:349. [PubMed: 17963426]

20. Hou T, Wang J. Expert Opinion on Drug Metabolism & Toxicology. 2008; 4:759. [PubMed: 18611116]

21. Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH. Journal of Pharmacological and Toxicological Methods. 2000; 44:251. [PubMed: 11274894]

22. Morishita M, Peppas NA. Advanced Drug Delivery Reviews. 2012; 64:479. [PubMed: 22388003]

23. Wang J, Krudy G, Xie X-Q, Wu C, Holland G. J Chem Inf Model. 2006; 46:2674. [PubMed: 17125207]

24. Andrews CW, Bennett L, Yu LX. Pharmaceutical Research. 2000; 17:639. [PubMed: 10955834]

25. Yoshida F, Topliss JG. Journal of Medicinal Chemistry. 2000; 43:2575. [PubMed: 10891117]

26. Yoshida F, Topliss JG. Journal of Medicinal Chemistry. 2000; 43:4723.

27. Wang Z, Yan AX, Yuan QP, Gasteiger J. European Journal of Medicinal Chemistry. 2008; 43:2442. [PubMed: 18603330]

28. Moda TL, Montanari CA, Andricopulo AD. Bioorganic & Medicinal Chemistry. 2007; 15:7738. [PubMed: 17870541]

29. Wang J, Hou T. Annual Report Computational Chemistry. 2000

30. Han, V.; Bernard, T. Weinheim: Wiley-vch; 1998. p. 435

31. Ma CY, Yang SY, Zhang H, Xiang ML, Huang Q, Wei YQ. J Pharmaceut Biomed. 2008; 47:677.

32. Imawaka H, Ito K, Kitamura Y, Sugiyama K, Sugiyama Y. Pharm Res. 2009; 26:1881. [PubMed: 19418207]

33. Tian S, Li Y, Wang J, Zhang J, Hou T. Molecular Pharmaceutics. 2011; 8:841. [PubMed: 21548635]

34. Paixao P, Gouveia LF, Morais JAG. Int J Pharmaceut. 2012; 429:84.

35. Xu X, Zhang W, Huang C, Li Y, Yu H, Wang Y, Duan J, Ling Y. International journal of molecular sciences. 2012; 13:6964. [PubMed: 22837674]

36. Hou T, Wang J, Zhang W, Xu X. J Chem Inf Model. 2007; 47:460. [PubMed: 17381169]

37. Helguera AM, Combes RD, Gonzalez MP, Cordeiro M. Current Topics in Medicinal Chemistry. 2008; 8:1628. [PubMed: 19075771]

38. Ahmed SS, Ramakrishnan V. PLoS One. 2012; 7:e40654. [PubMed: 22815781]

39. Olivares-Morales A, Hatley OJD, Turner D, Galetin A, Aarons L, Rostami-Hodjegan A. Pharmaceutical Research. 2014; 31:720. [PubMed: 24072264]

40. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Journal of Medicinal Chemistry. 2002; 45:2615. [PubMed: 12036371]

41. Abraham MH, Ibrahim A, Zhao Y, Acree WE. J Pharm Sci. 2006; 95:2091. [PubMed: 16886177]

42. Cruciani C, Crivori P, Carrupt PA, Testa B. J Mol Struc-Theochem. 2000; 503:17.

43. Norinder U, Osterberg T, Artursson P. Eur J Pharm Sci. 1999; 8:49. [PubMed: 10072478]

44. Yan AX, Wang Z, Cai ZY. International journal of molecular sciences. 2008; 9:1961. [PubMed: 19325729]

45. Miteva MA, Violas S, Montes M, Gomez D, Tuffery P, Villoutreix BO. Nucleic Acids Research. 2006; 34:W738. [PubMed: 16845110]

46. Lagorce D, Sperandio O, Galons H, Miteva MA, Villoutreix BO. Bmc Bioinformatics. 2008; 9

47. Hong HX, Xie Q, Ge WG, Qian F, Fang H, Shi LM, Su ZQ, Perkins R, Tong WD. J Chem Inf Model. 2008; 48:1337. [PubMed: 18564836]

48. Abrahamian E, Fox PC, Naerum L, Christensen IT, Thogersen H, Clark RD. J Chem Inf Comput Sci. 2003; 43:458. [PubMed: 12653509]

49. Obrezanova O, Gola JMR, Champness EJ, Segall MD. J Comput Aided Mol Des. 2008; 22:431. [PubMed: 18273554]

50. Obrezanova O, Csanyi G, Gola JMR, Segall MD. J Chem Inf Model. 2007; 47:1847. [PubMed: 17602549]

51. Konovalov DA, Coomans D, Deconinck E, Vander Heyden Y. J Chem Inf Model. 2007; 47:1648. [PubMed: 17602606]

52. Weaver S, Gleeson NP. Journal of Molecular Graphics & Modelling. 2008; 26:1315. [PubMed: 18328754]

53. Wang JM, Krudy G, Xie XQ, Wu CD, Holland G. J Chem Inf Model. 2006; 46:2674. [PubMed: 17125207]

54. Abshear T, Banik GM, D'Souza ML, Nedwed K, Peng C. SAR and QSAR in environmental research. 2006; 17:311. [PubMed: 16815770]

55. Dancik Y, Miller MA, Jaworska J, Kasting GB. Advanced Drug Delivery Reviews. 2013; 65:221. [PubMed: 22285584]

56. Nomoto M, Tatebayashi T, Morita J, Suzuki H, Aizawa K, Kurosawa T, Komiya I. J Pharm Sci. 2009; 98:1532. [PubMed: 18683862]

57. De Buck SS, Sinha VK, Fenu LA, Gilissen RA, Mackie CE, Nijsen MJ. Drug Metab Dispos. 2007; 35:649. [PubMed: 17267621]

58. Grabowski T, Jaroszewski JJ. Journal of veterinary pharmacology and therapeutics. 2009; 32:249. [PubMed: 19646089]

59. Artemenko A, Muratov E, Kuz'min V, Kulinskij M, Borisuk I, Golovenko N, Tropsha A. Antivir Res. 2009; 82:A56.

60. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Koutentis PA, Kollias G. Chem Biol Drug Des. 2010; 76:397. [PubMed: 20925691]

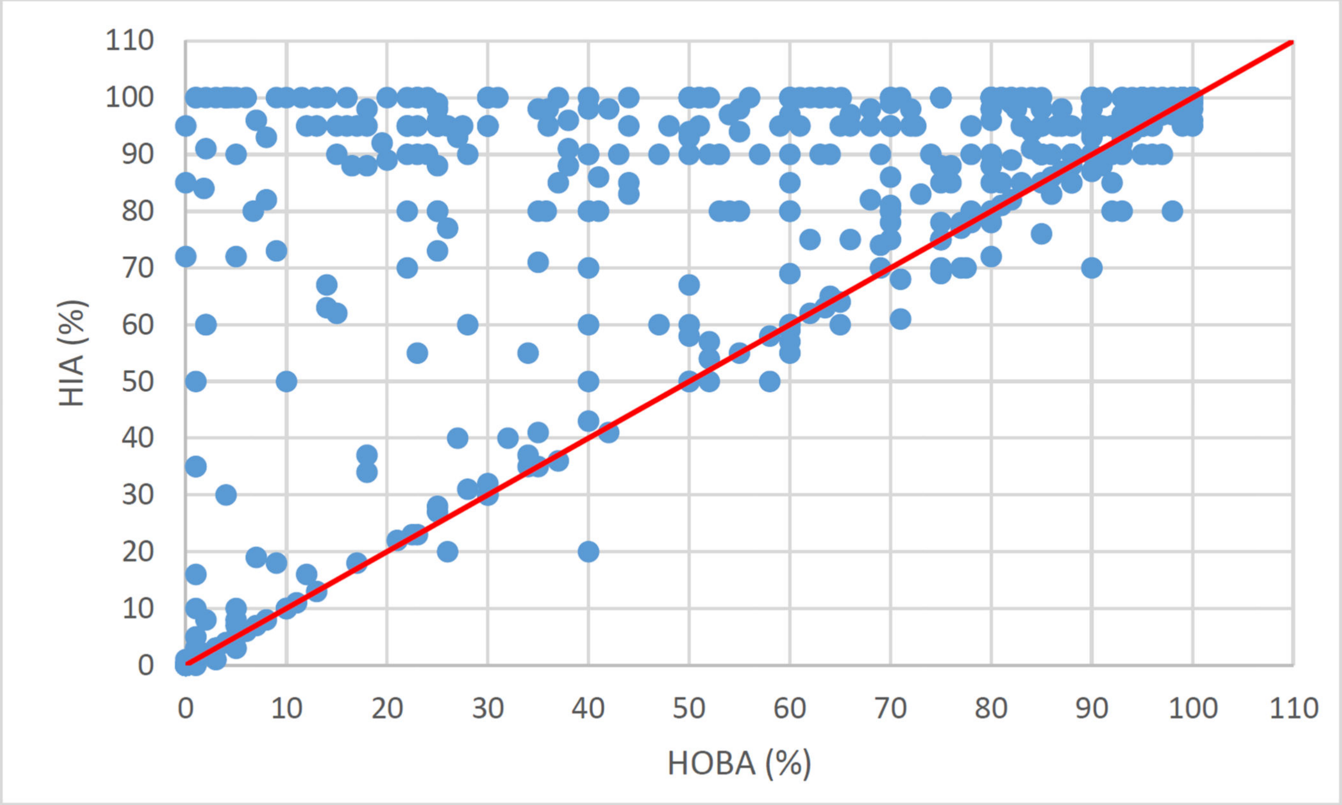

Figure 1.

Relationship between human oral bioavailability and human intestinal absorption for 510 drugs. The red line is the diagonal line where the X-axis and Y-axis values are equal, the correlation coefficient R is 0.63.

Table 1

Summary of the latest computational models for predicting human oral bioavailability

#	Authors	Model Type	Descriptors	Modeling Tools	Performance
1	Ma et al.	Classification	Cerius 2 descriptors	GA-GC-SVM	N_{train} = 690, N_{test} = 76, for the test set: TP = 58, FN = 2, TN=7, FP =9, R_{suc} = 86%
2	Imawaka et al.	Direct calculation	Animal pharmacokinetic data	Direct calculation, Vb_{β} estimated using Method 1a	N=61, AUE=15.0, RMSE=20.9
3	Imawaka et al.	Direct calculation	Animal pharmacokinetic data	Direct calculation, Vb_{β} estimated using Method 1b	N=61, AUE=17.0, RMSE=23.2
4	Imawaka et al.	Direct calculation	Animal pharmacokinetic data	Direct calculation, Vb_{β} estimated using Method 2	N=28, AUE=10.7, RMSE=15.4
5	Tian et al.	QSAR	Basic molecular properties and structural fingerprint	Genetic function approximation and multiple liner regression	N=996, R=0.79, q_{LOO}=0.72, RMSE=22.3%, R_{test}=0.71 $RMSE_{test}$=23.6%
6	Paixão, et al.	QSAR	*In vitro* P_{app} and Cl_{int}, etc.	Regression	N=49, RMSE=16%
7	Paixão, et al.	QSAR	*In silico* P_{app}, *in vitro* Cl_{int}, etc.	Regression	N=25, RMSE=19.8%
8	Paixão, et al.	QSAR	*In vitro* P_{app}, *in silico* Cl_{int}, etc.	Regression	N=22, RMSE=31.9%
9	Paixão, et al.	QSAR	*In silico* P_{app} and Cl_{int}, etc.	Regression	N=68, RMSE=34.6%
10	Ahmed, et al.	Classification	physiochemical and structural properties	logistic classifier	N_{train} = 969, R_{suc} = 71%
11	Xu et al.	QSAR	Dragon descriptors	SVM-regression	N_{train} = 156, N_{test} = 36, R^2 = 0.8, Q^2 = 0.72, RMSE = 31%, $RMSE_{test}$ = 22%
12	Xu et al.	QSAR	Dragon descriptors	SVM-regression	N_{train} = 122, N_{test} = 27, R^2 = 0.75, Q^2 = 0.63, RMSE = 28%, $RMSE_{test}$ = 77%
13	Xu et al.	QSAR	Dragon descriptors	SVM-regression	N_{train} = 180, N_{test} = 44, R^2 = 0.78, Q^2 = 0.80, RMSE = 36%, $RMSE_{test}$ = 36%
14	Xu et al.	QSAR	Dragon descriptors	SVM-regression	N_{train} = 197, N_{test} = 43, R^2 = 0.69, Q^2 = 0.68, RMSE = 42%, $RMSE_{test}$ = 46%
15	Olivares-Morales et al.	Classification	Rat oral bioavailability	Receiver operating characteristics	N=123, R_{suc}=69 %
16	Olivares-Morales et al.	Classification	Dog oral bioavailability	Receiver operating characteristics (ROC)	N=126, R_{suc}=75 %
17	Olivares-Morales et al.	Classification	NHP oral bioavailability	Receiver operating characteristics (ROC)	N=40, R_{suc}=90 %

Table 2

The performance of classification models predicting HOBA from animal data.[*]

Species	Threshold	TP	FP	TN	FN	Sensitivity	Specificity	Successful Rate
Rat	22%	54	17	31	21	0.72	0.65	69%
Dog	58%	51	20	44	11	0.82	0.69	75%
NHP	35%	20	4	16	0	1.00	0.80	90%

[*] TP: true positive, FP: false positive, TN: true negative, FN: false negative, sensitivity: TP/(TP+FN), specificity: TN/(TN+FP) and successful rates: (TP+TN)/(TP+TN+FP+FN) were calculated according to Table SII of Ref. The performance of the classification