J Mol Model (2006) 12: 577–589
DOI 10.1007/s00894-005-0065-z

ORIGINAL PAPER

Govindan Subramanian · Douglas B. Kitchen

Computational approaches for modeling human intestinal absorption and permeability

Received: 4 April 2005 / Accepted: 28 September 2005 / Published online: 1 April 2006

Abstract Human intestinal absorption (HIA) is an important roadblock in the formulation of new drug substances. Computational models are needed for the rapid estimation of this property. The measurements are determined via in vivo experiments or in vitro permeability studies. We present several computational models that are able to predict the absorption of drugs by the human intestine and the permeability through human Caco-2 cells. The training and prediction sets were derived from literature sources and carefully examined to eliminate compounds that are actively transported. We compare our results to models derived by other methods and find that the statistical quality is similar. We believe that models derived from both sources of experimental data would provide greater consistency in predictions. The performance of several QSPR models that we investigated to predict outside the training set for either experimental property clearly indicates that caution should be exercised while applying any of the models for quantitative predictions. However, we are able to show that the qualitative predictions can be obtained with close to a 70% success rate.

Dedicated to Professor Dr. Paul von Ragué Schleyer on the occasion of his 75th birthday.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00894-005-0065-z.

G. Subramanian · D. B. Kitchen (✉)
Computer-Aided Drug Discovery Department,
Albany Molecular Research, Inc.,
21 Corporate Circle,
P.O. Box 15098 Albany, NY 12212-5098, USA
e-mail: douglask@albmolecular.com
Tel.: +1-518-4640279
Fax: +1-518-4640289
e-mail: gsubramanian@ttpharma.com

Present address:
G. Subramanian
Transtech Pharma.,
4170 Mendenhall Oaks Parkway,
High Point, NC 27265, USA

Keywords Human intestinal absorption · Caco-2 permeability · QSPR · QSAR · ADME

Introduction

Understanding the physicochemical and pharmacokinetic properties of known drugs and potential drug candidates is a major bottleneck for the low success rate of compounds in clinical development [1]. Traditional structure–activity relationship (SAR) studies optimize the potency and efficacy of a congeneric compound series on a protein target. A rapid understanding of the absorption, distribution, metabolism, and excretion (ADME) characteristics of compounds still impedes significant progress in this area [2–4]. For instance, oral dosing is usually the most desirable way to administer drugs and therefore the therapeutic efficacy of a compound often involves the efficient transport or absorption of a drug to the blood stream. Therefore, the plasma solubility, membrane permeability, protein binding, transport properties, and the diffusion kinetics are some of the components influencing the overall bioavailability of the drug. Metabolism or elimination of a drug may also decrease the efficacy of a compound or increase toxicity or unwanted side effects.

The primary barrier towards good bioavailability is human intestinal absorption (HIA). Therefore, there is an increasing need to understand and measure the effect of physicochemical properties of the drug on the intestinal absorption process [5, 6]. It requires dissolution, passage through the gut and finally diffusion or transport into the blood stream. For example, the P-glycoprotein (P-gp) activity in the apical cell membrane may limit the bioavailability, while absorption through transcellular (membrane diffusion, carrier-mediated) or paracellular routes alter the pharmacokinetic profile of the compounds [7, 8]. The cost and time factors involved in in vivo and in situ experiments [9–11], make it difficult to collect sufficient data to analyze structural contributions to the rate of intestinal absorption. Recent advances in high-throughput screening and combinatorial chemistry synthesis have elevated the need to gain a priori information regarding the intestinal absorption on a variety of new chemical entities. Determining the amount of HIA requires expensive experiments and hence simpler in vitro models have been developed.

In order to obtain pharmacokinetic information earlier in drug discovery projects, in vitro Caco-2 (immortalized human colon adenocarcinoma cell line) monolayers are now used because they exhibit remarkable morphological and functional similarity to the small intestinal columnar epithelium [12–15]. The permeability measured from these experiments agrees, in general, with HIA and hence is useful for experimental modeling of in vivo absorption. The measured apparent permeabilities (P_{app}) depend on the cell culture, growth factors, and other experimental conditions, resulting in slightly different values reported by various groups (see supporting information) [16–23]. There is a need to develop quantitative and predictive mathematical models that relate various physicochemical properties to intestinal permeability and absorption so that poorly absorbed compounds are eliminated in the initial stages of drug discovery. The primary goal of this work is to develop computational models to describe the initial absorption of compounds from the gut into the blood stream.

Cellular permeability can be understood as a series of partitioning and associated diffusion of a molecule from one region to another in a lipid bilayer that surrounds the cell. Therefore, early physical and computational studies intended to explain intestinal absorption of drugs focused mainly on a single physicochemical property like octanol–water partition coefficient or the distribution coefficient [24–29]. Reasonable correlations were obtained for a homologous series of compounds, although structural diversity impeded the model's predictivity. Other properties such as the molecular weight, size, and shape, polar van der Waals surface area (PSA), and H-bonding capability are believed to be important for modeling intestinal permeability [30]. Hence, Lipinski proposed a scheme to classify the intestinal absorption using simpler yet powerful 1D-descriptors such as the count of polar atoms, log*P*, and the molecular weight [31]. Similarly, Clark used the computed PSA to categorize (good, medium, poor) the extent of intestinal absorption for structurally diverse compounds [32].

Table 1 Statistical results for $\log P_{app}$ (Eqs. 3, 8, 9, 10 and 11) and logHIA[a] (Eqs. 4, 5, 6, 7 and 12) for training and prediction (sub)set data and comparisons with QSPR models reported in literature

Models	Type[b]	N_{tr}[c]	r^2 (tr)	r^2 (cv)	N_{pc}[d]	rmse[e]	N_{pr}[f]	r^2 (pr)[g]	rmse[h]	N_{pr}[i]	r^2 (ss)[j]	rmse[k]
***logP**_{app}*												
Reference [23]	PSA	9	0.907			0.36						
Equation (8)	PSA+MW	17	0.693									
Reference [54]	PLS (9)	17	0.909	0.852	2	0.305						
Equation (9)	PSA+MW	17	0.771	0.654	2	0.458	50	0.455	0.734	42	0.672	0.533
Equation (10)	G/PLS (6)	17	0.972	0.549	1	0.161	50	0.010	1.526	35	0.632	0.493
Equation (11)	PSA+MW	22	0.610	0.437	2	0.561	37	0.622	0.570	35	0.691	0.520
Equation (3)	G/PLS (6)	22	0.892	0.535	1	0.295	37	0.629	0.840	31	0.780	0.530
logHIA &	***%HIA***											
Reference [40]	PLS (10)	20	0.916	0.798	3	8.1						
Reference [41]	Hashkeys	20	0.691			21.7						
Reference [37]	PSA	20	0.94			9.2						
Reference [32]	PSA	20	0.94			9.1						
Equation (12)[l]	PSA	20	0.886			12.9	62	0.463	29.1			
Equation (12)[m]	PSA	30	0.549			21.3	46	0.655	19.3			
Equation (4)	G/PLS (6)	30	0.907	0.808	3	9.0	46	0.166	31.7	37	0.452	21.0
Equation (5)	G/PLS (6)	30	0.908	0.814	3	7.6	46	0.191	31.2	38	0.462	19.6
Equation (6)	G/PLS (6)	30	0.814	0.617	1	13.4	46	0.304	28.5	37	0.539	17.9
Equation (7)	G/PLS (6)	30	0.668	0.343	3	17.5	46	0.367	26.4	39	0.559	17.9

[a] The rms error for Eqs. (4, 5, 6, 7) and that reported by Norinder are given in %HIA after the computed logHIA values are transformed using Eq. (2)

[b] Variables and methods used in fitting the experimental data. The values in parenthesis indicate the number of independent variables used in the multivariate analysis

[c] Number of molecules in the training set

[d] Number of PLS components

[e] Root mean square error for the training set

[f] Number of molecules in the prediction set

[g] square of the correlation coefficient for the prediction set

[h] Root mean square error for the prediction set

[i] Number of molecules in the prediction subset

[j] square of the correlation coefficient for the prediction subset (the outliers that were not included are underlined in Table 3)

[k] Root mean square error for the prediction subset

[l] PSA_{50}=83.987 and the slope is −15.246

[m] PSA_{50}=104.673 and the slope is −28.97

A number of quantitative structure–property relationship (QSPR) models have also been proposed in the literature for predicting logP_{app} and %HIA [33–35]. Most of the reported QSPR models utilize PSA as an indicator of intestinal permeability and absorption. For a diverse set of 17 compounds, van de Waterbeemd and Camenisch showed that the log of the apparent permeability (logP_{app}) measured from Caco-2 monolayers correlates well with the molecular weight and the PSA obtained for a single conformation [36]. By considering multiple conformations for the β-adrenoreceptor antagonists series, a sigmoidal relationship between the dynamic PSA and the fractional absorption (%HIA) was derived [37]. Similarly, the Boltzmann-averaged dynamic solvent-accessible surface area obtained from molecular dynamics simulations was related to the intestinal absorption [22].

Artursson and coworkers [38] extended the PSA model [17] by including the dynamic non-polar surface area (NPSA) component and showed a good correlation with the intestinal permeability of 19 oligopeptides. However, the PSA, NPSA and H-bond atom counts were not determined to be the critical elements responsible for the observed cellular permeability of the 21 peptide and peptidomimetic compounds considered by a different group [39]. This suggests that QSPR models based on logP, PSA or NPSA, while helpful for deriving meaningful correlations within a narrow structural class, may not extend universally. Including other physicochemical variables in the QSPR model would remove the over-dependence of P_{app} on PSA. Alternatively, descriptors replacing PSA can be mapped to HIA so that the intestinal absorption of both polar and apolar molecules can be modeled simultaneously. QSPR efforts along these directions reveal similar or better performances compared to models obtained using PSA as the response variable. In fact, the statistical results obtained by employing multivariate partial least square fitting (PLS) protocols with several molecular descriptors excluding PSA [40] and also by using molecular hash keys [41] was comparable to that derived using PSA [37] for the same set of compounds (Table 1). Very recently, a quantitative model for predicting the %HIA was proposed by combining a genetic algorithm and a neural network scoring function [42]. The significance of the model can be appreciated from the small (9.4%) root-mean-square error (rmse) obtained for a training set of 67 compounds. However, the major drawback rests in the compound selection since a single model was used to describe both passive and carrier-mediated absorption mechanisms. In addition, with 52/67 compounds exhibiting >75% HIA (strongly absorbed), the training set data primarily contains compounds with high absorption and therefore may result in a biased model.

One of the primary goals of this study is to develop QSPR models for estimating both logP_{app} and %HIA following the procedure outlined in Fig. 1. As a first step, experimental P_{app} measurements and %HIA data reported in literature were compiled. Subsequently, actively transported compounds were identified and separated from the dataset. The remaining molecules were then divided into training and prediction sets. The training set was composed such that the experimental value was distributed uniformly with respect to the measured P_{app} and %HIA. Using these evenly distributed datasets, QSPR models for computing logP_{app} and % HIA are derived independently. By deriving both simple and

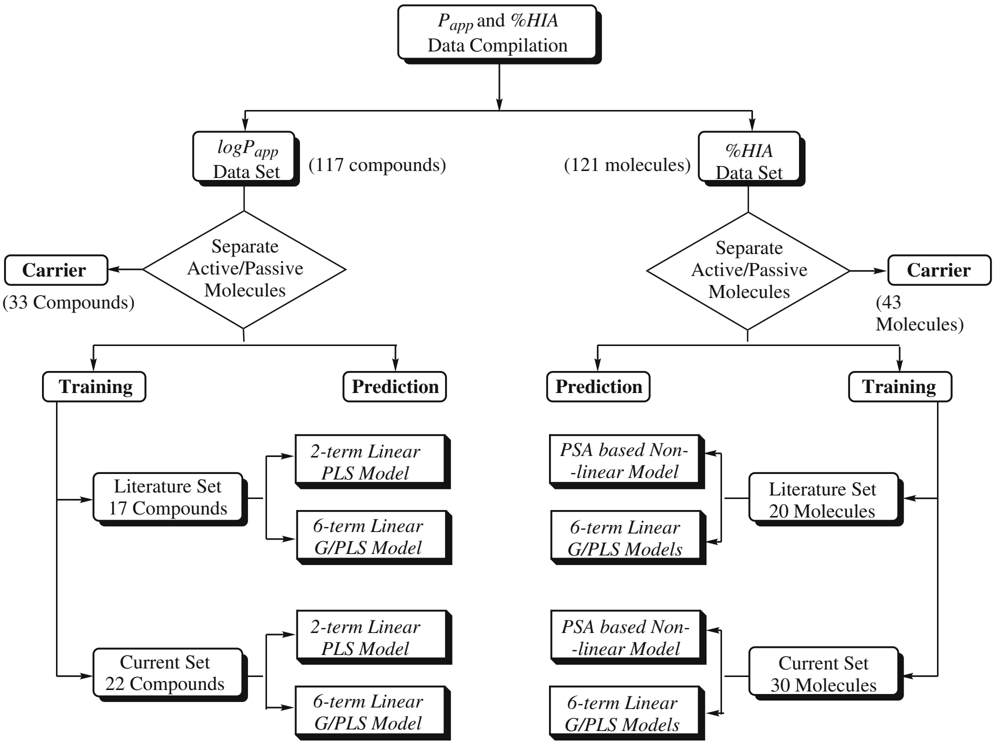

Fig. 1 Illustration of data compilation. The datasets were collected from the literature and compounds were divided into passively and actively transported sets. Further separations into training and prediction sets are described in the text. Regression analysis was only performed on passively absorbed compounds

complex QSPR equations, the influence of certain independent variables was assessed. The models were then extended to predict the logP_{app} and %HIA for an external validation set of passively absorbed compounds and compared to reported QSPR models [32, 36, 37, 40, 42]. Deriving similar mathematical relations for logP_{app} and %HIA and comparing the results obtained with the present training set data demonstrate the limitations with the oft-used training set compounds in the literature. As a next step, the logP_{app} and %HIA are predicted for molecules absorbed through active transport processes to see if these external influences are manifested in the molecular descriptors identified. This will also help determine if simple relationships exist between the active and passive absorption mechanisms. The quantitative logP_{app} and %HIA results are then extended to categorize the degree of intestinal permeability and absorption and compared to literature predictions [31, 32].

Dataset

Experimental P_{app} values determined using Caco-2 cell lines were pooled from several literature sources resulting in 117 compounds [16–23]. The data revealed that the P_{app} measurements reported by different laboratories for the same compounds varied significantly in magnitude. For example, the experimental logP_{app} for alprenolol ranges from −3.9 to −5.8, while that of atenolol varies between −5.7 and −7.5. For compounds where inter-laboratory experimental determinations were available, the values most consistent among the reported ones were used (see supporting information). The drawback of models derived based on the Caco-2 P_{app} values determined from a single laboratory is the potential bias in the results that limits extension to other compound sets reported by different groups. Alternatively, by using experimental P_{app} values that are consistent across different groups, realistic estimates of the experimental uncertainty in the various measurements can be ascertained for a larger set of compounds. This also provides a confidence limit for the experimental values reported for compounds from that group only. By following this procedure, some reported P_{app} values [43] were observed to deviate considerably from other literature reports for the same compounds and so were not included while generating the final training set. However, some of these compounds are used in the second logP_{app} prediction set but are not discussed in the text. The chemical structures used in this work are available from the authors upon request.

We believe that it is important to use compounds that permeate through passive absorption process in the dataset, since actively transported molecules are influenced by external environments that may not be modeled accurately. Prior to generating the final dataset, molecules reported to be substrates for various transport mechanisms or carrier-mediated processes (P-gp, peptide, nucleoside, etc.) were separated from those that are likely to undergo passive intestinal permeation. This division of compounds (Fig. 1) results in a dataset of only 59 compounds that most likely permeate through diffusion-controlled processes. A similar protocol was employed to gather the %HIA data from literature [19–21, 37, 42–45]. Although most of the fractional intestinal absorption data had been compiled [42], additional %HIA values reported in literature [43, 44] were also included, resulting in 121 compounds of which 76 molecules are probably passively absorbed. The training and prediction sets were chosen so that the logP_{app} and %HIA values span the whole range. Also, in the case of %HIA, compounds with values of 100% or 0% were used only in the prediction set.

Materials and methods

All the compounds were energy minimized using the universal force field [46] implemented in the Cerius^2 molecular modeling package (Accelrys, San Diego, CA) and imported into a Cerius^2 study table. Structural, constitutional, topological, and other calculated descriptors available through the Cerius^2 QSAR module were then included. In addition, the computed PSA [47] and the cube root of the gravitational index (GRAVIND) [42] were added. Since the passive transport of the compound across the intestinal epithelium is largely diffusion-controlled, and since the diffusion coefficient is inversely proportional to the square (SQINMW) or the cube root (CBINMW) of the molecular weight [48], descriptors representing such features were also included. By inspecting the independent variables in the study table, descriptors that did not have sufficient non-zero values or adequate variation were identified and removed.

A structurally heterogeneous logP_{app} dataset containing 22 compounds was used as the training set for predicting the apparent permeability of the remaining 19 molecules. Similarly, the training set for %HIA consisted of 30 compounds, leading to an external prediction set consisting of the remaining 46 molecules. The training set for modeling %HIA was selected such that molecules with 0 or 100 %HIA were not included. We believe that these limiting values do not correspond to the actual measurements, but reflect on the compound having crossed a threshold value. Therefore, using these compounds in the training set will affect the overall performance of the derived QSPR model. Hence, molecules with 0% and 100% HIA values are used in the prediction set and not in our training set. To have a direct comparison with reported QSPR models, additional equations were derived for the 17 and 20 compound training sets used in the literature for predicting logP_{app} and %HIA, respectively (Fig. 1) and the results are provided in the supporting information accompanying the paper.

Several studies in the literature have pointed out that P_{app} and %HIA are not likely to be linearly related to the computed descriptors [24–29, 36]. Therefore, the experimental P_{app} and the reported %HIA were transformed to logarithmic units. Since the %HIA has a closed scale with limiting values, fitting the data using linear approximations may lead to a statistically incorrect model. As in previous studies [40], a logit transformation was performed on the

Table 2 Experimental and computed logP_{app} for training set, prediction set, and compounds permeating through carrier-mediated processes[a]

Molecule	Expt	Equation (11)	Equation (3)
Training Set			
H216	−6.983	−5.929	−6.973
doxorubicin	−6.796	−7.182	−6.772
dnalaprilate	−6.208	−5.507	−5.661
benzylpenicillin	−5.708	−5.264	−5.398
practolol	−5.613	−4.811	−6.047
H95	−5.426	−4.807	−5.287
hydrocortisone	−4.914	−5.309	−4.595
imipramine	−4.851	−3.978	−4.601
desipramine	−4.666	−4.096	−4.627
naloxone	−4.550	−4.973	−4.885
clonidine	−4.521	−4.345	−4.094
pindolol	−4.460	−4.619	−4.752
propranolol	−4.378	−4.337	−4.010
betaxolol	−4.314	−4.627	−4.546
tenidap	−4.291	−5.073	−4.584
ibuprufen	−4.280	−4.153	−3.969
oxprenolol	−4.184	−4.512	−4.240
alprenolol	−4.125	−4.353	−4.234
caffeine	−4.074	−4.375	−4.035
betaxolol ester	−4.015	−4.842	−4.270
propranolol ester	−3.983	−4.596	−4.452
naproxen	−3.678	−4.329	−3.985
Prediction Set			
G6264	−6.854	−6.194	−7.530
G6700	−6.854	−6.114	−6.678
G6249	−6.745	−6.058	−6.681
G6191	−6.456	−6.297	−7.557
G6262	−6.367	−6.092	−6.848
G6703	−6.244	−6.228	−6.805
sulpiride	−6.160	−5.527	−5.356
atenolol	−5.936	−5.092	−6.494
acyclovir	−5.699	−5.486	−5.353
sumatriptan	−5.523	−4.933	−6.193
G6203	−5.509	−6.186	−6.072
H244	−5.220	−4.881	−5.470
ziprasidone	−4.910	−5.077	−2.741
tiacrilast	−4.900	−4.754	−3.245
mibefradil	−4.870	−5.382	−4.316
fleroxacin	−4.810	−5.033	−3.928
nitrendipine	−4.770	−5.502	−5.443
guanoxan	−4.710	−4.900	−5.597
chloramphenicol	−4.686	−5.559	−4.831
felodipine	−4.644	−4.914	−4.940
fluconazole	−4.526	−5.137	−3.518
trovaflaxicin	−4.520	−5.646	−4.134
warfarin	−4.417	−4.694	−3.980
timolol	−4.354	−5.068	−4.370
theophylline	−4.350	−4.603	−4.181
antipyrine	−4.310	−3.928	−3.745
diltiazem	−4.310	−4.945	−4.625
testosterone	−4.286	−4.382	−3.400
corticosterone	−4.263	−5.065	−4.273
lidocaine	−4.210	−4.184	−5.512
diazepam	−4.149	−4.310	−3.954
guanabenz	−6.000	−4.879	−3.908
coumarin	−4.110	−3.803	−1.792
timolol ester	−4.103	−5.331	−3.712
metoprolol	−4.036	−4.491	−4.978
oxprenolol ester	−4.012	−4.756	−4.369
alprenolol ester	−3.967	−4.610	−4.671
2ndPrediction Set			
olsalazine	−6.959	−5.859	−3.668
terbutaline	−6.420	−4.728	−4.394
epinephrine	−6.020	−4.659	−4.176
acetaminophen	−6.000	−4.203	−4.896
acrivastine	−7.721	−4.696	−4.390
bupropion	−5.824	−4.149	−4.572
cefuroxime	−8.420	−6.647	−5.477
chlorothiazide	−8.495	−5.650	−4.691
fluparoxan	−5.699	−4.132	−3.476
hydrochlorothiazide	−8.036	−5.705	−4.238
ketoprofen	−6.032	−4.515	−4.242
labetalol	−6.119	−5.421	−6.599
lamotrigine	−5.959	−5.233	−3.778
nadolol	−8.409	−5.109	−5.309
netivudine	−8.168	−5.558	−6.106
penicillin V	−8.770	−5.436	−5.290
phenytoin	−5.796	−4.680	−4.889
progesterone	−6.009	−4.398	−4.099
propylthiouracil	−6.018	−4.190	−4.608
sotalol	−7.377	−4.999	−4.850
trimethoprim	−6.061	−5.415	−5.601
Active Processes			
furosemide	−8.854	−5.688	−5.124
lisinopril	−8.658	−6.159	−6.689
loracarbef	−8.620	−5.650	−6.579
cefatrizine	−8.119	−7.055	−6.552
sulfasalazine	−6.886	−6.204	−3.679
methylprednisolone	−6.602	−5.375	−5.340
zidovudine	−6.553	−5.727	−4.000
amoxicillin	−6.481	−6.029	−6.380
L-leucine	−6.301	−4.320	−4.785
mannitol	−6.187	−5.253	−5.998
sucrose	−6.149	−6.412	−7.691
L-glutamine	−6.071	−5.065	−6.409
L-dopa	−6.000	−5.143	−5.474
azithromycin	−5.983	−7.163	−7.033
ondansetron	−5.959	−4.410	−4.817
D-mannitol	−5.932	−5.200	−5.173
methotrexate	−5.921	−7.552	−7.272
acetylsalicylate	−5.620	−4.300	−1.296
cephalexin	−5.570	−5.631	−6.384
cimetidine	−5.514	−5.284	−5.524
erythromycin	−5.428	−7.357	−6.789
taurocholate	−5.396	−6.500	−4.240

Table 2 (continued)

Molecule	Expt	Equation (11)	Equation (3)
acebutolol	−5.351	−5.206	−6.635
L-phenylalanine	−5.161	−4.478	−4.822
salicylic acid	−4.924	−4.228	−3.390
quinidine	−4.690	−4.678	−3.669
dexamethasone	−4.631	−5.423	−5.100
D-glucose	−4.602	−5.094	−3.323
verapamil	−4.580	−5.053	−4.566
acetylsalicylic acid	−4.513	−4.381	−3.678
taurocholic acid	−4.462	−6.487	−6.922
prazosin	−4.361	−5.536	−5.204
valproic acid	−4.319	−3.989	−4.595
glycine	−4.097	−4.236	−4.099
gabapentin	−4.046	−4.468	−5.231

[a] underlined values in the prediction set represent compounds that exhibited more than 1 log unit rms error and so were not included in the calculations of the prediction subset in Table 1

dependent variable (%HIA) using Eq. (1). The computed logitHIA was transformed back to %HIA using Eq. (2), so that direct comparisons with experimental values could be achieved.

$$\log\mathrm{it}(\mathrm{HIA}) = \ln\left[(\%\mathrm{HIA}+10)/(110-\%\mathrm{HIA})\right] \tag{1}$$

$$\%\mathrm{HIA} = \left[(110 * \exp(\mathrm{logitHIA}) - 10)/(1+\exp(\log\mathrm{itHIA}))\right] \tag{2}$$

Using Cerius^2, the training-set molecules were initially subjected to a Genetic/Partial Least Squares (G/PLS) statistical fitting procedure [49, 50] to fit the dependent function (logP_{app} or logitHIA) using 15 variables and up to four PLS components. In this way, a plausible subset of descriptors that best describe logP_{app} and logitHIA were obtained from the pool of independent variables. Subsequent G/PLS runs on the same training-set data used only the descriptor subset and seven terms in the regression equation. The initial random population of 100 equations was evolved for 20,000 generations, resulting in a final set of 100 model equations that characterizes the observed logP_{app} and logitHIA through the square of the correlation coefficient (r^2) fitness function. All the variables involved in the G/PLS regression were scaled and a maximum of three principal components were allowed. To determine the role of certain variables like PSA, Jurs-terms [51], electrotopological indices [52], and AlogP [53], four different models were obtained for logitHIA by including or removing some descriptors from the initial descriptor subset used in the G/PLS run. In addition, QSPR models with varying complexity and for the standard training set used in the literature were obtained for direct comparison with reported models (Table 1).

The internal cross-validated correlation coefficient, r^2(cv) (using the leave-one-out method) was used to ascertain the statistical quality of the derived functions. The accuracy of the estimations was evaluated by comparing the results obtained using the best model equation against the experimental logP_{app} and logitHIA values for each of the prediction sets. The quantitative logP_{app} and %HIA results were also used to categorize the molecules based on the extent of intestinal permeability and absorption using the following threshold values. Thus, compounds with computed logP_{app}<−6 and<−5 are classified as *poorly* and *moderately* permeable, respectively. All the remaining molecules are considered to possess *good* intestinal permeabilities. Similarly, a %HIA of <30 suggests poor intestinal absorption, while molecules exhibiting HIA >70% possess good intestinal absorption. Compounds with medium HIA range between 30 and 70%.

Results

Table 1 summarizes the performances of the various QSPR models derived using the experimental logP_{app} and %HIA for the passively absorbed compounds listed in Tables 2 and 3. The computed logP_{app} and %HIA are also given for molecules absorbed through various transport processes but were not used as part of the training set.

Statistical results for logP_{app}

The G/PLS analysis on the 22 training-set molecules (Table 2) resulted in Eq. (3) with an r^2 of 0.89 for logP_{app} and an r^2(cv) of 0.54. The 0.3 log unit root-mean-square error (rmse) for the training set is within experimental error limits.

$$\begin{aligned}\log\mathrm{P_{app}} = & -4.628 + 0.399 * \mathit{Hbond\ acceptor} + 0.450 \\ & * \mathit{AlogP} - 0.329 * \mathit{Hbond\ donor} + 0.698 \\ & * \mathit{Jurs\_RPCS} - 0.166 * \mathit{Kappa\_1} \\ & + 0.00161 * \mathit{Jurs{-}WNSA\_1}\end{aligned} \tag{3}$$

The computed r^2(pr) and rmse obtained for the 37 prediction-set compounds suggest a satisfactory performance of the model when applied to an external dataset, but with a significant rms error. However, closer examination reveals that the predicted logP_{app} is greater than one log unit when compared with the experimental values for six compounds identified in Table 2. Since the measured apparent permeabilities are available from one literature source only (see supporting information), it is difficult to ascertain if these compounds possess large uncertainty in the observed logP_{app} values or should be treated as the prediction-set outliers. Assuming the latter, significant improvement in the predicted r^2 and rmse is achieved for the remaining 31 molecules in the prediction subset

Table 3 Experimental and computed %HIA values for the training set, prediction set, and compounds absorbed through carrier-mediated mechanisms

Molecule	Expt	Equation (4)	Equation (5)	Equation (6)	Equation (7)	Equation (12)
Training Set						
cromolyn	0.5	−1.7	7.7	−1.2	3.4	23.2
olsalazine	2.3	0.1	8.4	6.7	19.4	27.5
cefuroxime	5.0	4.4	1.9	9.4	0.9	14.0
enalaprilat	10.0	18.1	12.5	12.6	50.3	58.1
benzyl penicillin	30.0	58.0	32.6	67.1	58.4	67.7
norfloxacin	35.0	33.3	22.5	67.7	78.6	77.0
Phenoxymethylpenecillinic acid	45.0	56.0	43.6	61.7	44.7	60.5
ziprasidone	60.0	85.1	64.7	69.0	88.1	85.7
hydorchlorothiazide	67.0	66.4	87.4	74.8	70.1	36.4
lamotrigine	70.0	74.6	79.8	61.4	67.2	56.4
guanabenz	75.0	67.5	82.9	73.0	82.8	71.7
propylthiouracil	75.0	86.7	80.5	81.0	76.8	89.3
sorivudine	82.0	80.3	66.3	47.2	57.4	46.6
bupropion	87.0	92.3	81.1	86.4	93.3	93.9
acrivastine	88.0	80.0	83.5	95.5	89.8	89.7
betaxolol	90.0	88.1	93.3	88.5	94.0	89.3
oxprenolol	90.0	93.8	96.9	86.9	88.1	87.9
pindolol	90.0	92.2	85.6	92.7	92.3	83.5
scopolamine	90.0	95.7	90.5	100.4	82.5	85.7
timolol	90.0	91.2	87.4	88.6	64.7	72.6
naloxone	91.0	89.1	83.0	83.8	87.5	80.4
progesterone	91.0	91.6	96.1	84.1	87.9	93.2
clonidine	95.0	92.7	91.2	87.7	95.6	89.8
imipramine	95.0	89.7	88.6	98.2	103.1	96.9
metoprolol	95.0	93.9	99.6	101.3	88.1	88.2
sotalol	95.0	89.2	85.8	90.2	85.5	73.3
trimethoprim	97.0	87.9	101.5	81.6	61.9	52.4
warfarin	98.0	80.9	98.7	91.2	87.8	87.2
diltiazem	99.0	94.4	86.5	79.2	71.8	88.1
nordiazepam	99.0	98.6	93.7	92.8	92.2	89.7
Prediction Set						
ceftriaxone	1.0	−4.2	−6.6	−7.6	−8.9	3.0
ganciclovir	3.8	62.9	62.0	45.0	27.4	25.3
doxorubicin	5.0	29.3	39.8	−4.0	5.6	8.3
chlorothiazide	13.0	69.6	83.5	76.8	58.2	39.2
netivudine	28.0	62.2	49.0	57.9	46.5	43.0
acyclovir	30.0	56.0	66.9	66.8	31.1	38.6
nadolol	34.5	95.4	87.8	84.6	86.3	71.6
sulpiride	36.0	64.2	75.4	74.2	59.8	55.0
atenolol	50.0	81.9	89.2	89.0	75.9	67.9
metolazone	64.0	75.3	77.4	60.5	84.6	64.6
ciprofloxacin	69.0	30.4	21.2	73.2	81.4	76.5
terbutaline	73.0	83.6	73.6	75.2	86.0	77.5
sumatriptan	75.0	93.4	83.6	97.9	92.9	82.5
acetaminophen	80.0	56.0	84.7	100.5	57.5	87.7
bromazepam	84.0	101.9	95.0	96.4	82.6	84.6
trovafloxacin	88.0	23.7	1.1	18.9	73.0	60.6
chloramphenicol	90.0	63.3	71.3	39.6	36.3	50.5
ephedrine	90.0	75.0	86.3	103.0	79.9	92.1
phenytoin	90.0	60.6	92.6	94.1	87.1	82.8
propranolol	90.0	93.7	85.3	91.4	97.5	90.8
tenidap	90.0	62.5	53.7	75.8	94.8	75.2

Table 3 (continued)

Molecule	Expt	Equation (4)	Equation (5)	Equation (6)	Equation (7)	Equation (12)
hydrocortisone	91.0	88.6	86.1	48.3	71.6	68.7
terazosin	91.0	89.4	95.9	77.8	38.3	58.2
alprenolol	93.0	89.4	89.9	96.4	92.8	90.8
fluconazole	95.0	95.3	67.7	84.3	43.1	67.3
labetalol	95.0	64.6	63.6	56.0	92.2	63.8
oxazepam	97.0	87.6	86.4	81.7	87.5	81.7
theophylline	98.0	46.6	40.1	82.6	35.2	77.7
prednisolone	98.8	87.5	75.6	55.6	72.4	68.7
naproxen	99.0	66.1	88.5	95.4	85.3	90.1
tiacrilast	99.0	59.1	49.7	78.4	63.9	79.3
antipyrine	100.0	99.6	100.2	102.4	84.1	94.5
bumetanide	100.0	28.6	79.1	47.5	63.3	45.4
caffeine	100.0	64.8	42.2	87.9	48.5	86.6
corticosterone	100.0	90.8	95.3	74.6	73.9	78.7
coumarin	100.0	49.5	87.3	101.2	67.0	94.6
desipramine	100.0	84.0	96.0	104.9	102.6	95.1
diazepam	100.0	102.4	96.4	96.7	93.3	93.2
felodipine	100.0	104.8	104.1	79.8	80.0	85.9
fluparoxan	100.0	87.4	81.1	89.7	89.8	91.9
fluvastatin	100.0	74.1	64.4	42.0	81.7	74.2
ibuprofen	100.0	45.8	75.0	91.7	76.5	92.0
ketoprofen	100.0	52.2	75.7	90.0	78.8	87.6
lormetazepam	100.0	94.2	85.2	75.2	88.8	87.4
practolol	100.0	92.8	92.6	90.0	84.1	79.9
testosterone	100.0	90.6	100.6	99.5	94.1	92.1
Active Processess						
gentamicin	0.0	102.5	105.7	107.8	83.7	10.2
raffinose	0.3	99.5	79.8	−8.2	−1.2	1.1
lactulose	0.6	80.1	54.0	2.4	16.6	7.3
sulfasalazine	13.0	18.5	31.3	17.4	12.5	25.0
mannitol	15.0	83.0	79.1	55.8	45.3	46.4
foscarnet	17.0	−9.2	54.5	−8.8	−4.1	68.4
lisinopril	25.0	7.5	21.1	−2.6	31.6	31.6
pravastatin	34.0	73.4	79.6	10.4	30.0	44.3
erythromycin	35.0	56.3	104.0	3.6	−9.1	13.0
azithromycin	35.0	82.8	105.9	4.1	−8.6	18.1
defuroxime axetil	36.0	2.0	12.8	8.3	−6.9	12.7
sucrose	42.0	98.4	71.2	69.2	34.2	9.7
etoposide	50.0	92.7	108.9	21.7	−2.3	23.6
ranitidine	50.0	89.9	70.4	76.2	55.7	70.8
gabapentin	50.0	21.9	34.2	81.0	55.0	81.3
tranexamic acid	55.0	20.2	27.1	75.0	46.5	81.0
L-glutamine	60.0	6.6	25.7	51.5	14.2	49.0
taurocholic acid	60.0	50.7	51.7	38.4	25.0	29.2
furosemide	61.0	44.9	56.5	58.4	70.1	44.7
cimetidine	85.0	82.4	67.4	66.2	34.2	53.8
captopril	67.0	57.1	42.2	76.8	75.5	86.0
cefatrizine	76.0	35.9	12.6	6.0	2.6	7.3
quinidine	80.0	95.0	89.6	99.7	89.2	88.4
methylprednisolone	82.0	88.4	79.0	29.6	65.5	66.4
acebutolol	89.5	86.9	94.3	77.5	64.1	70.0
amoxicillin	93.5	37.4	14.2	22.4	45.5	30.4
verapamil	95.0	92.4	104.8	85.0	46.8	86.9
cephalexin	98.0	65.6	30.6	54.9	66.3	49.6

Table 3 (continued)

Molecule	Expt	Equation (4)	Equation (5)	Equation (6)	Equation (7)	Equation (12)
acetylsalicylic acid	100.0	24.1	63.0	87.5	48.0	85.4
D-glucose	100.0	79.6	52.2	52.1	49.0	54.1
D-Phe-L-Pro	100.0	48.7	42.5	80.8	71.7	71.9
dexamethasone	100.0	85.6	70.7	55.7	72.2	68.6
glycine	100.0	−7.1	2.4	65.5	−6.9	80.3
L-dopa	100.0	29.8	37.1	42.8	54.9	53.0
L-leucine	100.0	13.6	37.9	67.4	25.4	82.7
L-phenylalanine	100.0	30.3	52.9	88.2	58.6	80.9
loracarbef	100.0	47.4	16.0	43.5	60.8	48.4
methotrexate	100.0	−0.6	6.7	−6.9	−6.7	2.7
ondansetron	100.0	95.5	81.9	98.9	90.5	92.0
prazosin	100.0	89.9	85.1	78.2	48.5	60.7
salicylic acid	100.0	14.6	62.4	92.0	43.4	86.3
valproic acid	100.0	13.0	57.0	73.6	32.1	91.8
zidovudine	100.0	38.6	65.6	75.0	−4.5	22.0

[a] The underlined %HIA values correspond to compounds removed in the statistical analysis of the prediction subset (Table 1)

(Table 1). Clearly, there are many other interpretations available for the outliers in this case, including structural diversity outside the training set or alternate absorption mechanisms in these compounds.

Statistical results for logitHIA

The statistical quality of logitHIA (Eqs. 4, 5, 6, 7) and the quantitative estimations of %HIA for the 30 training-set compounds (Table 3) show that good QSPR models are achieved for %HIA. Among the four models, the best results are obtained through Eq. (4) with an r^2(tr) of 0.907 while the logitHIA estimations obtained using Eq. (7) account for only 67% of the variance in the training set (Table 1). This is also reflected in the larger rmse obtained using Eq. (7) compared to the 9% error in HIA associated with Eq. (4). The subtle changes in the r^2(tr) obtained through Eqs. (4, 5, 6 and 7) are attributed to the effect of certain independent variables like PSA, Jurs term, etc. in modeling the intestinal absorption.

$$\begin{aligned}\mathrm{logitHIA} =& 10.342 + 0.783 * S\_dssC - 104.493 \\ & * SQINMW - 0.0262 * S\_dO \\ & - 0.022 * PSA - 0.164 * S\_ssCH2 \\ & - 0.366 * \mathrm{Alog}P\end{aligned} \tag{4}$$

$$\begin{aligned}\log \mathrm{itHIA} =& 3.849 + 10.991 * Jurs\_RNCG \\ & + 0.686 * S\_dssC + 0.114 * S\_ssO \\ & + 0.883 * JX - 95.220 * SQINMW \\ & - 0.232 * Hbond\ acceptor\end{aligned} \tag{5}$$

$$\begin{aligned}\log \mathrm{itHIA} =& 11.324 - 0.015 * PSA - 55.141 * CBINMW \\ & + 1.032 * Jurs\_FNSA\_2 + 0.0234 \\ & * Jurs\_WNSA\_3 + 11.329 * Jurs\_RNCG \\ & - 0.171 * A \log P\end{aligned} \tag{6}$$

$$\begin{aligned}\log \mathrm{itHIA} =& 8.029 - 0.00379 * Area + 0.735 \\ & * GRAVIND - 0.0577 * PSA \\ & - 61.408 * CBINMW + 0.724 \\ & * Hbond\ donor - 0.0142 * MW\end{aligned} \tag{7}$$

The computed r^2(pr) and rmse associated with logitHIA for the 46 prediction-set compounds suggest a weaker performance of the above models for quantitative logitHIA predictions. In striking contrast to the training set results, the predictions obtained through Eqs. (6) and (7) are the best among the four G/PLS models derived here for estimating %HIA. The poor performance of Eqs. (4) and (5) is attributed to the use of certain electrotopological indices in these models that are not represented in some of the validation-set compounds. For example, the absence of methylene carbons (>CH_2) in ketoprofen, ibuprofen, and coumarin is partially responsible for the >40% rmse observed for these molecules by using Eq. (4) (Table 3). Similarly, the lack of ether-like oxygen (–O–) in ticrilast, caffeine, and ciprofloxacin contributes to the smaller r^2 observed by using Eq. (5), while compounds lacking a carbonyl carbon (nadolol, chlorothiazide, etc.) are predicted poorly by Eqs. (4) and (5). When molecules predicted with large rmse (>40%) are excluded from the validation set, the correlation coefficient for the validation subset is improved significantly along with a substantial decrease in the rmse (~10%). Considering the large deviations in the reported inter-laboratory %HIA values for certain compounds (see supporting information), the predictions are still quantitative and within acceptable experimental error limits.

Discussion

The G/PLS results presented above reveal that reliable models have been obtained for estimating logP_{app} and %HIA across a heterogeneous dataset (Tables 2 and 3). However, the choice of the training-set data and the complexity of the regression equations (Eqs. 3, 4, 5, 6, 7) obtained in this study do not allow a direct comparison of the present results with reported QSPR models. Consequently, additional models were derived for the standard 17 and 20 training-set compounds used in the literature for estimating $\log\mathrm{P}_{\mathrm{app}}$ and %HIA and compared with the existing models (Table 1) derived using 22 and 30 compound sets, respectively (see supporting information).

Comparison with reported logP_{app} models

Palm et al. reported one of the earliest $\log\mathrm{P}_{\mathrm{app}}$ models by correlating the dynamic PSA for six β-adrenoreceptor-blocking agents [17, 23] and subsequently extended to model nine homologues [23] with considerable success. Our initial attempts to obtain a linear relationship between the computed PSA and logP_{app} for the 22 training-set compounds yielded an r^2 of 0.583, considerably lower than the results obtained with Eq. (3) (Table 1). This reveals the limitation on the use of a single variable for quantitative logP_{app} predictions across a diverse dataset. However, a statistically significant correlation coefficient (0.833) was obtained for the literature training set of 17 compounds when logP_{app} was modeled using molecular weight and PSA (Eq. 8) [32]. This equation also agrees with the general notion that intestinal permeability is enhanced by increasing the molecular weight and by decreasing PSA. Since Eq. (8) was not validated through predictions for an external dataset, the generality of the model could not be assessed. Therefore, we refit a model using these terms to verify its extension for all the molecules considered in this study (Table 2). The comparable statistical quality (Table 1) and the coefficient on the variables obtained through Eq. (9) for the same 17 training-set compounds reflect the similarity of our model with Eq. (8). The marginal difference in the coefficients in Eqs. (8) and (9) is attributed to the use of slightly different PSA values and also to the use of consensus logP_{app} values (see supporting information), rather than the experimental values reported in the previous work [9].

$$\begin{aligned}\log\mathrm{P}_{\mathrm{app}} &= 0.008 * MW - 0.043 * PSA \\ &\quad - 5.165 (n = 17)\end{aligned} \tag{8}$$

$$\begin{aligned}\log\mathrm{P}_{\mathrm{app}} &= 0.00443 * MW - 0.0288 * PSA \\ &\quad - 4.459 (n = 17)\end{aligned} \tag{9}$$

$$\begin{aligned}\log\mathrm{P}_{\mathrm{app}} &= -3.238 - 0.695 * CHI\_3\_C \\ &\quad + 0.00129 * Jurs\_PPSA\_2 \\ &\quad - 0.309 * Kappa\_3 - 0.00730 \\ &\quad * Jurs\_TPSA - 0.149 * Hbond\ donor \\ &\quad + 0.0655 * Rotlbonds\end{aligned} \tag{10}$$

$$\begin{aligned}\log\mathrm{P}_{\mathrm{app}} &= -0.00291 * MW - 0.0150 * PSA \\ &\quad - 3.052 (n = 22)\end{aligned} \tag{11}$$

The simplicity and the easily computable variables suggest Eqs. (8) or (9) to be versatile in modeling the apparent intestinal permeability of virtual compound-libraries. Applying Eq. (9) to a validation set of 50 molecules resulted in an inferior performance with an r^2(pr) of 0.455 and an rmse of 0.73 log units. To determine if the predictions are affected by the choice of descriptors, additional regressions were performed and compared with a reported PLS model [54] (r^2=0.91 versus logP_{app} and using nine variables). The G/PLS model represented by Eq. (10) for the same 17 compounds demonstrates a much better performance with fewer variables (Table 1). In contrast, the correlation coefficient obtained for the 50 validation-set molecules clearly invalidates the extension of Eq. (10) for logP_{app} predictions on unknowns.

In comparing previous QSPR models of logP_{app} and Eq. (10), it is clear that neither the choice of descriptors nor the number of terms in the QSPR equation is a major contributor for the weaker correlations of the models. One of the key reasons for the failure of Eqs. (9) and (10) derived using the standard 17 compounds in the literature arises due to the lack of structural diversity in the training set. Molecules like alprenolol, atenolol, practolol, and metaprolol fall within a homologous series, as do the steroids corticosterone, hydrocortisone, and testosterone. In addition, recent pharmacokinetic studies show some of the literature training-set compounds, like dexamethasone and sulfasalazine, to be transported through efflux mechanisms. [43] Since our attempt was to derive a QSPR model for passively permeable compounds, a training set that spans more of the diversity space was essential. The 22 training-set compounds (Table 2) used for deriving the logP_{app} model in the present study overcome some of the contaminations in the dataset and represent diverse compounds that appear to permeate predominantly through passive processes.

Equation (11) derived from 22 training-set compounds yielded weaker correlations with an increased rmse compared to equations obtained with the 17-compound training set (Table 1). However, the contrasting interpretations of the effect of molecular weight (compare the coefficients in Eqs. (9) and (11) suggest the sensitivity of the independent variables to the number and choice of the compounds in the dataset. On the other hand, the comparable statistical quality of Eq. (3) with the reported PLS model [54] (Table 1) clearly demonstrates that the lack of predictive ability of Eq. (11) is not due to the different selection of the training set molecules but caused by the incomplete representation of the descriptors. This incomplete representation is further substantiated by the fact that models of HIV-protease-inhibitor uptake by Caco-2 monolayers required multivariate regressions with proper choice of independent variables [55].

Comparison with reported %HIA models

Similar to the logP_{app} models, QSPR efforts in modeling %HIA involved a standard dataset of 20 molecules in the literature (see supporting information). Using only PSA as the independent variable, a simple non-linear model was obtained with an r^2 of 0.90 [23, 29]. Similarly, a linear model was derived for logitHIA by using multivariate statistics [40]. In spite of a good correlation [r^2(tr)=0.916] for the 20 training-set compounds, neither of these models has been used for quantitative %HIA prediction on an external validation set. Our attempts to model the %HIA using the Boltzmann sigmoidal curve function (Eq. 12) and PSA for the same 20 compounds yielded a correlation coefficient of 0.941, in close agreement to reported QSPR models (Table 1).

$$\%\mathrm{HIA} = 100/[1 + \exp((\mathrm{PSA}_{50} - \mathrm{PSA})/\mathrm{slope})\ \text{where}\ \mathrm{PSA}_{50}\ \text{is the PSA value at which HIA} = 50\% \qquad (12)$$

However, the use of both actively transported and passively diffused compounds in the literature %HIA and logitHIA models can affect the predictive ability. Although stringent conditions were used in obtaining the 20 training-set compounds in the literature [23], pharmacokinetic studies reveal foscarnet and sulfsalazine to be absorbed through various transport mechanisms [7, 8, 43]. Consequently, Eq. (12) was fitted against the experimental %HIA values for the 30 training-set compounds used in this study (Table 3) and compared with the G/PLS results obtained through Eqs. (4, 5, 6 and 7). In spite of explaining only 55% of the variance in the training-set data, this single descriptor non-linear model outperforms all the G/PLS models in its predictive ability. The reversal in the r^2 and rmse trends between the training and prediction sets (Table 1) for the %HIA models also demonstrates that the G/PLS procedure perhaps overfits the 30 training-set compounds considerably and more so when electrotopological descriptors are used (compare Eqs. 4 and 5 with 6 and 7). Given the similar r^2(ss) and rmse obtained through Eqs. (6) and (7) with that of Eq. (12), it is tempting to use the PSA model for the sake of simplicity. However, this non-linear model breaks down when apolar molecules are considered and while predicting the %HIA of a congeneric series where hydrophobic substitutions (–CH_3, –C_2H_5, –Ph, etc.) are made. In both these cases, no change in the computed %HIA will be necessarily observed, although the intestinal absorption process could be affected significantly. Again, by virtue of the sigmoidal relationship of PSA with %HIA, the PSA_{50} and the slope values in Eq. (12) are highly sensitive to the number of compounds used in the training set (Table 1).

Interpretation of the physicochemical descriptors

Several studies have demonstrated that the intestinal absorption and permeability are governed by a number of factors including the lipophilicity, molecular size and shape, and hydrogen-bonding capabilities [27]. The G/PLS models derived here (Eqs. 3, 4, 5, 6, 7 and 10) utilize all these variables in the regression equations and illustrate that random descriptors are not included in explaining the dependent property. For example, three out of the five variables considered by Lipinski [31] as features consistent with drug-like compounds for classifying the intestinal absorption process are manifested through Equation (3). The diversity in the physicochemical variables describing

Table 4 Total and percent correct classification (%, in parenthesis) of HIA based on Eqs. (4, 5, 6, 7, and 12) for the training set, prediction set, and the molecules considered to be absorbed through carrier-mediated processes

Dataset	N[a]	Eq. (3)	Eq. (11)	N[a]	Eq. (4)	Eq. (5)	Eq. (6)	Eq. (7)	Eq. (12)
Training	22	20 (90.9)	16 (72.7)	30	28 (93.3)	27 (90.0)	27 (90.0)	22 (73.3)	24 (80.0)
Prediction	19	17 (89.5)	10 (52.6)	46	26 (56.5)	29 (63.0)	32 (69.6)	31 (67.4)	34 (73.9)
Transport	34	15 (44.1)	15 (44.1)	43	15 (34.9)	13 (30.2)	21 (48.8)	14 (32.6)	19 (44.2)

[a] Number of molecules in the dataset

Table 5 Total and percent correct consensus prediction and success rate (%, in parenthesis) based on Eqs. (4, 5, 6, 7 and 12) for the training set, prediction set, and the molecules considered to be absorbed through other mechanisms (Transport Set)

Confidence level	N[a]	100%	80%	60%	40%	20%	0%
Equation's (4, 5, 6, and 7, 12)							
Training Set	30	21(70.0)	2(6.7)	2(6.7)	4(13.3)	1(3.3)	0(0.0)
Prediction Set	46	18(39.1)	5(10.9)	8(17.4)	6(13.0)	6(13.0)	3(6.6)
Transport Set	43	2(4.6)	5(11.6)	8(18.6)	10(23.3)	8(18.6)	10(23.3)

[a] Number of molecules in the dataset

the logP_{app} also shows that the H-bonding and the Jurs' terms in Eq. (3) capture the effects of PSA in Eq. (11). Eq. (3) further reveals that H-bond acceptor atoms and lipophilic substitutions improve the intestinal permeability while H-bond donors decrease permeability.

The multiple QSPR models for logitHIA substitute the Jurs descriptors (compare Eq. 4 with Eq. 5) and the substitution of electrotopological indices through components of Jurs-terms, gravitational index, and area (compare Eqs. 4 and 5 with Eqs. 6 and 7). The use of molecular weight in all the G/PLS equations agrees with the general notion that molecular weight and diffusion are interrelated.

Qualitative predictions of intestinal permeability and absorption

Although quantitative estimates of logP_{app} and %HIA are effective for comparisons across a homologous series, such accuracies are not essential for screening large compound libraries. Furthermore, given the considerable uncertainty in the experimental %HIA measurements and the sizeable errors in the computed values, it would be beneficial if the QSPR models reported in this study reproduce qualitative features correctly. The coarse filters (poor, medium, good) used to define the extent of intestinal permeability (see Materials and methods) demonstrate that reasonable hit rates are observed when the logP_{app} values obtained from multivariate analysis are used (Table 4; Eq. 3). The results also show that the 2-term logP_{app} model is approximately 20% less accurate in qualitatively classifying the apparent permeability of compounds in the training and the prediction sets. In spite of the relatively poor r^2(pr) observed for the four %HIA models (Eqs. 4, 5, 6 and 7), the consistent classification of the degree of intestinal absorption adds confidence in extending these equations for qualitative %HIA predictions of molecules that have not been considered in the prediction set. Although the classifications obtained through the non-linear %HIA model (Eq. 12) perform better for the prediction-set compounds, Eqs. (6) and (7) perform well when outliers are removed. Also, Eq. (12) may not perform well for a homologous series where the analogues differ by their hydrophobic substitutions ($–CH_3$, $–C_2H_5$, $–CH(CH_3)_2$, etc.) while Eqs. (6) and (7) should predict trends in these series more consistently. The <50% hit rates obtained in correctly classifying compounds that are carrier-mediated (Table 4) suggests that Eqs. (6), (7) and (12) discriminate molecules transported through different transport mechanisms from those that are passively diffused. In addition, the QSPR models derived here indirectly suggest that the intestinal permeability and absorption of actively transported compounds are modeled reasonably, except for the inability in accounting for the factors contributed by the solute environment and the kinetics of the mediators.

Since the five QSPR models derived for predicting the %HIA perform only moderately well for the prediction-set data, a consensus approach was designed to minimize the number of false positives in the classification scheme. Alternatively, comparing the results obtained using all five %HIA models provides an indirect mechanism of assessing the reliability of computed predictions. The results in Table 5 show that 21 of the 32 compounds in the training-set compounds are classified correctly by all the models, resulting in a 100% confidence in the computed values. The predictions are 80% accurate if four out of the five models classify a molecule similarly. Alternatively, borderline cases are identified when three out of the five models predict similarly. These results show an improved performance when the results of all the five models are used collectively (Table 5) over the %HIA models considered individually (Table 4) for qualitative classification.

Conclusions

To derive the logP_{app} and logitHIA models, linear relationships were assumed to exist with the descriptors investigated. However, a non-linear behavior is illustrative when %HIA is modeled using PSA. In comparing the QSPR models derived for logP_{app} and %HIA, we believe that the multivariate-statistics approach performs the best for estimating the P_{app} while a single variable like PSA describes the human intestinal absorption process satisfactorily. The logP_{app} models derived in the present study also demonstrate that the intestinal permeability is governed by several parameters. The varied dataset used in the reported QSPR models restricts a direct comparison on the performance of the present models to that proposed in literature for modeling logP_{app} and HIA. However, QSPR models derived using the standard training set used in the literature resulted in weaker correlations when extended to an external prediction set. This is attributed to the lack of structural diversity and the inclusion of actively transported compounds in the literature training set and to the use of non-passively absorbed molecules. In contrast, the use of training-set data containing more passively absorbed compounds resulted in models with good predictive ability. In the absence of a quantitative %HIA comparison for the models derived here with reported QSPR results, qualitative features on the extent of intestinal absorption can be used as an index to assess the performance.

We believe that Eqs. (4, 5, 6, 7) describe %HIA and Eqs. (9, 10 and 11) describe logP_{app} well. The complexity of both experiments implies that careful use of any models for these properties is necessary. We have attempted to eliminate actively transported compounds from the data. Further work on computational models will be required as additional measurements are reported. We do find that we can obtain predictive models without large numbers of variables and that these models are predictive outside the training set.

The fact that coefficients on common variables in the two equation sets are not always consistent is problematic. P_{app} is often used as a surrogate test for HIA, yet some of our equations imply the physical processes involved in the two experiments are not the same. Because the correlations and validations are better for P_{app} and that it is a simpler experiment, we believe that the P_{app} models are more likely to extend to new compounds.

References

1. Smith DA, van de Waterbeemd H (1999) Curr Opin Chem Biol 3:373–378
2. Clark DE, Pickett SD (2000) Drug Discovery Today 5:49–58
3. Fecik RA, Frank KE, Gentry EJ, Menon SR, Mitscher LA, Telikepalli H (1998) Med Res Rev 18:149–185
4. Tarbit MH, Berman J (1998) Curr Opin Chem Biol 2:411–416
5. Navia MA, Chaturvedi PR (1996) Drug Discovery Today 1:179–189
6. Chan OH, Stewart BH (1996) Drug Discovery Today 1:461–473
7. Tsuji A, Tamai I (1996) Pharm Res 13:963–977
8. Lin JH, Lu AYH (1997) Pharmacol Rev 49:403–449
9. Stewart BH, Chan OH, Lu RH, Reyner EL, Schmid HL, Hamilton HW, Steinbaugh BA, Taylor MD (1995) Pharm Res 12:693–699
10. Stewart BH, Chan OH, Jezyk N, Fleisher D (1997) Advanced Drug Delivery Reviews 23:27–45
11. Barthe L, Woodley J, Houin G (1999) Fundam Clin Pharm 13:154–168
12. Pinto M, Robine-Leon S, Appay M-D, Kedinger M, Triadou N, Dussaulx E, LaCroix B, Simon-Assmann P, Haffen K, Fogh J, Zweibaum A (1983) Biol Cell 47:323–330
13. Hidalgo IJ, Raub TJ, Borchardt RT (1989) Gastroenterology 96:736–749
14. Wilson G, Hassan IF, Dix CJ, Williamson I, Shah R, Mackay M (1990) J Controlled Release 11:25–40
15. Delie F, Rubas WA (1997) Crit Rev Ther Drug Carrier Syst 14:221–286
16. Artursson P, Karlsson J (1991) Biochem Biophy Res Commun 175:880–885
17. Palm K, Luthman K, Ungell A-L, Strandlund G, Artursson P (1996) J Pharm Sci 85:32–39
18. Rubas W, Cromwell MEM (1997) Advanced Drug Delivery Rev 23:157–162
19. Yee S (1997) Pharm Res 14:763–766
20. Lennernäs H (1998) J Pharm Sci 87:403–410
21. Grès M-C, Julian B, Bourrié M, Meunier V, Roques C, Berger M, Boulenc X, Berger Y, Fabre G (1998) Pharm Res 15:726–733
22. Krarup LH, Christensen IT, Hovgaard L, Frokjaer S (1998) Pharm Res 15:972–978
23. Palm K, Luthman K, Ungell A-L, Strandlund G, Beigi F, Lundahl P, Artursson P (1998) J Med Chem 41:5382–5392
24. Martin YC (1981) J Med Chem 24:229–237
25. Nook T, Doelker E, Buri P (1988) Int J Pharmaceut 43:119–129
26. Merino V, Freixas J, Val Bermejo MD, Garrigues TM, Moreno J, Plá-Delfina JM (1995) J Pharm Sci 84:777–782
27. Camenisch G, Folkers G, van de Waterbeemd H (1996) Pharmaceutica Acta Helvetiae 71:309–327
28. Testa B, Carrupt P-A, Gaillard P, Billois F, Weber P (1996) Pharm Res 11:335–343
29. Camenisch G, Folkers G, van de Waterbeemd H (1998) Eur J Pharm Sci 6:325–333
30. Camenisch G, Alsenz J, van de Waterbeemd H, Folkers G (1998) Europ J Pharm Sci 6:313–319
31. Lipinski CA, Christopher AL (1997) Advanced Drug Delivery Rev 23:3–25
32. Clark DE (1999) J Pharmaceutical Sci 88:807–814
33. Sugawara M, Takekuma Y, Yamada H, Kobayashi M, Iseki K, Miyazaki K (1998) J Pharmaceutical Sci 87:960–966
34. Winiwarter S, Bonham NM, Ax F, Hallberg A, Lennernäs H, Karlén A (1998) J Med Chem 41:4939–4949
35. Bermejo M, Merino V, Garrigues TM, Plá-Delfina JM, Mulet A, Vizet P, Trouiller G, Mercier C (1999) J Pharm Sci 88:398–405
36. van de Waterbeemd H, Camenisch G (1996) Quant Struct Act Relat 15:480–490
37. Palm K, Stenberg P, Luthman K, Artursson P (1997) Pharm Res 14:568–571
38. Stenberg P, Luthman K, Artursson P (1999) Pharm Res 16:205–212
39. Goodwin JT, Mao B, Vidmar TJ, Conradi RA, Burton PS (1999) J Peptide Res 53:355–369
40. Norinder U, Österberg T, Artursson P (1999) European J Pharm Sci 8:49–56
41. Ghuloum AM, Sage CR, Jain AN (1999) J Med Chem 42:1739–1748
42. Wessel MD, Jurs PC, Tolan JW, Muskal SM (1998) J Chem Inf Comp Sci 38:726–735
43. Irvine JD, Takahashi L, Lockhart K, Cheong J, Tolan JW, Selick HE, Grove JR (1999) J Pharm Sci 88:28–33
44. Kansy M, Senner F, Gubernator K (1998) J Med Chem 41:1007–1010
45. Pade V, Stavchansky S (1998) J Pharmaceutical Sci 87:1604–1607
46. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM (1992) J Am Chem Soc 114:10024–10035
47. The PSA arising from the oxygen, nitrogen, and halogen atoms, and the –OH, –NH2 groups was calculated using the solvent accessible surface area in Cerius2 with the van der Waals radii reported in reference 37.
48. Herman RA, Veng-Pedersen P (1994) J Pharm Sci 83:423–428
49. Rogers D, Hopfinger AJ (1994) J Chem Inf Comp Sci 34:854–866
50. Genetic partial least squares in QSAR: Rogers D (1996) In: Devillers J (ed) Genetic algorithms in molecular modeling. Academic, London
51. Stanton DT, Jurs PC (1990) Anal Chem 62:2323–2329
52. Hall LH, Kier LB, Brown BB (1995) J Chem Inf Comp Sci 35:1074–1080
53. Viswanadham VN, Ghose AK, Revankar GR, Robins RK (1989) J Chem Inf Comput Sci 29:163–172
54. Norinder U, Österberg T, Artursson P (1997) Pharm Res 14:1786–1791
55. Stewart BH, Chung FY, Tait B, Blankley CJ, Chan OH, John C (1998) Pharm Res 15:1401–1406