Full Paper

DOI: 10.1002/minf.201000118

In Silico Prediction of Caco-2 Cell Permeability by a Classification QSAR Approach

Hai Pham The,[a] Isabel González-Álvarez,[b] Marival Bermejo,[b] Victor Mangas Sanjuan,[b] Inmaculada Centelles,[c] Teresa M. Garrigues,[c] and Miguel Ángel Cabrera-Pérez*[a]

Presented at the 18th European Symposium on Quantitative Structure Activity Relationships, EuroQSAR 2010, Rhodes, Greece

Abstract: In the present study, 21 validated QSAR models that discriminate compounds with high Caco-2 permeability ($P_{app} \geq 8\times10^{-6}$ cm/s) from those with moderate-poor permeability ($P_{app} < 8\times10^{-6}$ cm/s) were developed on a novel large dataset of 674 compounds. 20 DRAGON descriptor families were used. The global accuracies of obtained models were ranking between 78–82%. A general model combining all types of molecular descriptors was developed and it classified correctly 81.56% and 83.94% for training and test sets, respectively. An external set of 10 compounds was predicted and 80% was correctly assessed by in vitro Caco-2 assays. The potential use of the final model was evaluated by a virtual screening of a human intestinal absorption database of 269 compounds. The model predicted 121 compounds with high Caco-2 permeability and the 90% of them had high values of human intestinal absorption ($HIA \geq 80$). This study provides the most comprehensive database of Caco-2 permeability and evidenced the utility of the combined methodology (in silico + in vitro) in the prediction of Caco-2 permeability. It suggests that the present methodology can be used in the design of large libraries of compounds with appropriate values of permeability and to perform virtual screening in the early stages of drug development.

Keywords: Caco-2 · Intestinal permeability · Quantitative structure-activity relationship · Classification model · In vitro permeability assay

1 Introduction

Approximately 90% of all the marketed drugs are administered orally and about 10% of oral drugs fail in development due to poor pharmacokinetic properties.[1–3] One of the most important challenges facing an oral drug is its movement across the intestinal epithelial barrier that determines the rate and extent of human absorption and ultimately affects its bioavailability. In addition, the permeability property per se is a multifactorial process which separates into different mechanisms, like paracellular and transcellular passive diffusion, active uptake and active secretion.[4]

Several in vitro permeability assays that mimic the relevant characteristics of in vivo absorption have been developed. The parallel artificial membrane permeability assay (PAMPA),[5] the human colon adenocarcinoma (Caco-2) cells assay,[6] the Madin-Darby canine kidney (MDCK) cell assay,[7] the rat duodenal immortalized cell line (2/4/A1 cell),[8] and the rat everted gut assay[9] are routinely used for the assessment of drug permeability. Their remarkable increased throughput and lower cost made them a good election to study drug permeability compared with in situ and/or in vivo animal studies. Particularly, the Caco-2 monolayer cell culture model is the "gold standard" for drug permeability and is widely used in drug discovery for the prediction of human intestinal permeability.[3] As a matter of fact, this in vitro model has been recommended by the US Food and Drug Administration (FDA) for determination of permeability of compounds to be classified according to the Biopharmaceutics Classification System (BCS).[10]

Nowadays, as the number of compounds that can be generated has increased dramatically, the in vitro methods can not longer match the demand in throughput. Even with the advances and improvements of the well defined in vitro assays,[3] the experimental measurement of ADME properties is still expensive and time consuming. As a result there has been a strong motivation to develop alter-

[a] *H. Pham The, M. Á. Cabrera-Pérez*
Molecular Simulation & Drug Design Group. Centre of Chemical Bioactive. Central University of Las Villas. Santa Clara 54830, Villa Clara, Cuba
phone/fax: 53-42-281192/53-42-281130
**e-mail: macabrera@uclv.edu.cu*
migue@gammu.com

[b] *I. González-Álvarez, M. Bermejo, V. Mangas Sanjuan*
Department of Engineering, Area of Pharmacy and Pharmaceutical Technology, Miguel Hernández University, 03550 Sant Joan d'Alacant, Alicante, Spain

[c] *I. Centelles, T. M. Garrigues*
Department of Pharmacy and Pharmaceutical Technology, University of Valencia, Burjassot 46100, Valencia, Spain

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201000118.

native ways to inexpensively and rapidly measure, model or predict ADME properties.[11] On the other hand a rising need for introducing selection filters prior to design and purchase of large compound libraries has turned the interest towards in silico predictions for ADME related properties.

Several researchers have explored Quantitative Structure Activity/Property Relationships (QSAR/QSPR) involving Caco-2 permeability.[12] In these studies, various types of molecular descriptors have been introduced to the QSPR modeling including physicochemical properties (polar surface area, hydrogen-bonding and size descriptors),[13,14] 1D, 2D and 3D indices. In all cases, QSPR models predicted Caco-2 permeability with a reasonable degree of accuracy. However, these models might not be practical because the size of permeability data sets is rather small. When the data are limited, statistical models often fail due to an over-fitting problem, resulting in a limitation on their use. Inter- and intralaboratory variability associated with Caco-2 permeability measurements remains a problem to be solved before a widely applicable QSPR model can be constructed.[15]

Taking into consideration the mentioned issues, the main goals of the present paper are to collect a large database of Caco-2 permeability; to obtain reliable models able to classify compounds into high permeability or moderate-poor permeability by means of Linear Discriminant Analysis (LDA); to validate the best classification model statistically and experimentally and finally, to evaluate the absorption predictive capacity of the best classification model in a virtual screening of human intestinal absorption database.

2 Materials and Methods

2.1 Databases and Molecular Descriptors

A heterogeneous database composed by 674 organic compounds was randomly assembled from more than 250 published articles. Several compounds were excluded attending to the following criteria: extremely high molecular weight; dose-limited and dose dependent absorption; endogenous substances (such as hormones and neurotransmitters); prodrugs and molecules containing a permanently charged group. The structurally heterogeneous database used covers a relatively wide range of molecular size, polarizability, hydrophilicity, lipophilicity and molecular charge.

Taking into consideration the large number of protocols to determine Caco-2 permeability, the collection procedure considered the similarity of the experimental assays in those important factors that affect the intestinal permeability (P_{app}) values.[16] Additionally, a classification scheme by categories (high and moderate-poor permeability) was applied, reducing the variability associated with measurement from different laboratories, assays, cell types and experimental conditions. Consequently, the drugs were divided into two classes according to the Caco-2 permeability cut-off values. A "high" permeability value was defined for $P_{app} \geq 8\times10^{-6}$ cm/s, whereas "moderate-poor" permeability was defined for compounds with $P_{app} < 8\times10^{-6}$ cm/s. This classification criterion is lower than the permeability value reported for Metoprolol ($P_{app} = 20\times10^{-6}$ cm/s), the standard compound considered by the FDA as high permeability class boundary.[17] Discussion on the acceptability of FDA cut-off permeability value is ongoing since many drugs with lower P_{app} are generally considered completed absorbed. Our experience in previous studies suggests that 8×10^{-6} cm/s as permeability cut-off value shows a reasonable ratio between compounds with a fraction of dose absorbed (FA) greater than 85% and compounds with lower values of FA.[18]

The available database was divided into training and test sets, using *k*-means cluster analysis (*k*-MCA). The *k*-MCA was carried out with the STATISTICA software 8.0.[19] In order to guarantee acceptable statistical quality of data cluster, the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible) were considered. Additionally, the standard deviation between and within cluster, the respective Fisher ratio and *p*-level of significance ($p < 0.05$) were inspected.

Compounds of the training and test sets were randomly collected from the previous clusters. This procedure permitted us to select, in a representative way and in all level of the linking distance (*Y*-axis), compounds for both sets.

Finally, the training and test sets were composed by 537 and 137 compounds (around 20% of the complete database), respectively. Compounds belonging to the test set were never used in the development of the discriminant functions and were set aside to assess the obtained discrimination models.

The permeability values were also predicted for an external set of 10 compounds belonging to different groups of the Biopharmaceutical Classification System (BCS).[20] The in silico results were validated by a Caco-2 cell assay.

Considering the relationship between in vitro Caco-2 permeability and human intestinal absorption,[21] a virtual screening to evaluate the potentiality of the in silico permeability model in identifying compounds with high human intestinal absorption (HIA) was carried out on a dataset of 269 compounds with *HIA* reported.

In the present study the parameters corresponding to 0D-1D, 2D and 3D molecular descriptors were estimated by DRAGON 5.4 software.[22] Descriptors 0D-2D were calculated using the SMILES (simplified molecular input line entry specification) code of each compound.[23] The calculation of 3D descriptors was carried out following a preliminary MM+ geometry optimization for each compound and then the (*x*,*y*,*z*)-atomic coordinates of the minimal energy conformations for each compound were determined using the quantum chemical semi-empirical method AM1[24] included in MOPAC 6.0 computer software.[25] In summary, more than 1400 descriptors were calculated.

2.2 Chemometric Methods

During the process of drug discovery, the accurate prediction of any pharmacokinetic or biopharmaceutical property sometimes is not necessary and the simple classification of compounds, in ranges of the property, is enough. The classification model has some advantages over the linear correlation models. First, the nonlinear effects of this kind of biological properties are implicitly accounted for in the classification. Second, the statistical classification methods can discriminate different biological mechanisms. Finally, experimental values are not usually necessary for classification, because classification deals with binary data, accepting moderate property variabilities.

In this sense, LDA was a convenient multivariate exploratory technique to differentiate compounds with high Caco-2 permeability from those with moderate-poor values.

For the classification models, the compounds were firstly clustered in two groups according to their permeability values (P_{app}). The discriminant function (Equation 1) which best describes Caco-2 permeability as a linear combination of the predictor *X*-variables (molecular descriptors) weighted by the a_n coefficients, was obtained by means of the General Linear Discriminant Analysis Module (GLDA) implemented in STATISTICA software 8.0.[19]

$$CLASS = a_1A_1 + a_2A_2 + a_3A_3 \ldots + a_nA_n + a_0 \quad (1)$$

In developing these classification functions, *CLASS* values of $+1$ and -1 were assigned to compounds with high and moderate-poor permeabilities, respectively. In addition, the compounds were considered unclassified (U) by the model when the differences in the percentage of classification between two groups did not differ by more than 5%.

The "*best subset*" technique, using the Wilks Lambda (λ) value as the criterion for choosing the best subset of predictor effects, was applied to select the molecular descriptors with the highest influence on the dependent variable. Standard statistical parameters such as the square of Mahalanobis distance (D^2), the Fisher ratio (F) and the corresponding p-level (p) were also considered. The validity of the preadopted assumptions such as normality, homoscedasticity, noncollinearity and linearity of the model was also checked.

The principle of maximal parsimony (Occam's razor) was taken into account as a strategy for model selection. Therefore, we selected the model with highest statistical signification having as few parameters (a_k) as possible.

The general performance of each classification model was assessed in terms of its Cooper Statistics.[26] Percentages of global good classification (accuracy) and Matthews' correlation coefficients (MCC) also permitted the validation of the model. The MCC is always between -1 and 1. A values of -1 indicates total disagreement (all-false predictions), and 1, total agreement (perfect predictions). The MCC is 0 for completely random classification.

The Receiver Operating Characteristic curve (ROC) was used to evaluate the accuracy of the discriminant function. ROC curve is the representation of sensitivity versus 1-specificity. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. Accuracy is measured by the area under the ROC curve (AUC). An area of 1 represents a perfect test; an area of 0.5 represents a worthless test.[27]

A rigorous procedure to assess the "realistic" predictive power of the model and the generalizability of the QSAR model to predict Caco-2 permeability for new chemical compounds was evaluated using a test dataset. This type of model validation is important, if we take into consideration that the predictive ability of any QSAR model can only be estimated using a set of compounds that was not used for building the model.[28] Therefore, it is important to ensure that the prediction algorithms are able to perform well on novel data belonging to the same applicability domain (AD) of training set compounds.

In this study a more demanding evaluation was performed. The predictive power of the model was assessed with an external set of 10 compounds belonging to different groups of the BCS. The computational predictions were experimentally corroborated by in vitro Caco-2 cell assays.

The chemical domain of the studied compounds in the models and the identification of potential *response outliers* (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, $>3\sigma$) were verified by the leverage approach.[29] A visual evaluation of the possible *outliers* as well as chemicals outside the AD of the model was carried out using the well known William's plot.[30]The reliability of the predicted data by the final model was verified using the leverage values. Only the predicted data for compounds belonging to the chemical domain of the training set were considered. In fact, leverage can be used as a quantitative measure of the model AD suitable for evaluating the degree of extrapolation. Prediction must be considered unreliable for compounds with a high leverage value ($h > h^*$, the critical value being $h^* = 3p/n$, where p is the number of model variables plus one, and n is the number of the objects used to calculate the model). These compounds are structurally influential chemicals in a model.

2.3 In Vitro Transport Study in Caco-2 Cells

Caco-2 cell line was provided by Dr. Ming Hu (Washington State University, Pullman). Cell monolayers were grown in Dulbecco's modified Eagle's media and were prepared by seeding 2.5×10^5 cells per insert on MILLICEL-PCF transwells (surface area 4.2 cm^2, 3 μm pore size).[31] Cell culture was maintained at 37 °C under 90% humidity and 5% CO_2.

The in vitro study was developed when confluence was reached, 19–22 days postseeding Caco-2 cells. The integrity

of each cell monolayer was checked by measuring its transepithelial electrical resistance (*TEER*) value before an experiment. *TEER* values were typically 500–750 Ω cm^2. Hank's balanced salt solution (HBSS) (9.8 g/L) supplemented with $NaHCO_3$ (0.37 g/L), HEPES (5.96 g/L) and glucose (3.5 g/L) at pH 7.00 was used for preparing the dosing solution and for the receptor chamber. The drug solution was loaded into the donor side and buffer was added to the receiver side of each cell monolayer. The six-well plate containing the cell monolayers was put into an orbital environmental shaker, which was maintained at a constant temperature (37 °C) and agitation rate (50 rpm) for the duration of the transport experiments. Four samples of 200 μL were taken at 30 min intervals, from the receiver side, and replaced by fresh buffer, for the permeation profile; moreover, two samples of 200 μL were taken from the donor side, at the beginning and the end of the assay, for the mass balance calculation.

Transport studies were performed from apical-to-basolateral (A-to-B) sides. Two different donor drug concentrations were used for Cimetidine, Propranolol, Norfloxacin, Vinblastine, Fexofenadine, Atenolol, Theophylline, Carbamazepine, Metoprolol and Verapamil. The apparent permeability in cell lines (P_{app}) was calculated following the equation:

$$(dQ/dt)/(S \cdot C) = P_{app} \quad (2)$$

where dQ/dt is the apparent appearance rate of drug in the receiver side, S the surface area of the monolayer (4.2 cm^2), and C the concentration in the donor side. The flux term dQ/dt was calculated from the linear regression of amounts in the receiver chamber versus time.

Samples of drugs were analyzed by HPLC with fluorescence and UV detection. Although the analytical techniques for these drugs are described in the literature,the analytical methods were previously validated over the concentration range of the samples used.

3 Results and Discussion

3.1 LDA-QSAR Models: Development and Validation

In this work some LDA-based models were developed to classify molecules according to Caco-2 permeability value. Twenty one classification models based on indices generated from 20 descriptor families were built and the best model by family of descriptors was selected considering the statistical parameters and percentage of good classifications for training and test data sets. The global accuracies of all models ranked between 78–82 %. A general model based on all molecular descriptors was also developed and it classified correctly 81.56 % and 83.94 % for training and test set, respectively. These results are listed in Table 1. As can be appreciated in this table the best statistical parameters were obtained for the last model, where all families of descriptors were used.

The best discriminant function to classify the Caco-2 permeability is given below along with their statistical parameters:

$$\begin{aligned} CLASS = {} & -0.60 - 1.08 \cdot Ms + 3.28 \cdot DELS - 0.67 \cdot BELp3 - \\ & 0.24 \cdot E1e - 1.07 \cdot H3e - 0.44 \cdot nArCONHR + \\ & 0.33 \cdot nArNH_2 - 4.48 \cdot TPSA(NO) + \\ & 0.40 \cdot GVWAI\text{-}50 \end{aligned} \quad (3)$$

$$N = 537 \quad \lambda = 0.58 \quad D^2 = 2.86 \quad F = 41.5 \quad p < 0.001$$

The meaning of the variables included in the final model (Equation 3) appears in Table 2.

As can be seen in the Equation 3 and in Table 2, nine variables are present in the final classification model.

The large F index and the small p value are indicative of the model statistical significance. In addition, the values of the Wilks λ statistic (λ can take values from cero, perfect discrimination, to one, no discrimination) and the Mahalanobis distance (a measure of the separation between two groups) show that the model has an adequate discriminative capacity.

A better threshold for a priori classification probability can be estimated by means of ROC curve. As Figure 1 shows, the optimal threshold for predicting the best permeability values with the present model is 0.83. As can be seen in the figure, the model is not random since the area under the ROC curve (0.89 u[2)] is significantly higher than the area under the random classifier curve (diagonal line). This high value is a measure of the excellent accuracy of the discriminant function selected.

Table 1. Statistical parameters of the best QSAR models obtained using different molecular descriptors to predict the Caco-2 permeability. Tr: training, Ts: test.

Descriptor family[a]	λ[b]	MCC[c] (Tr/Ts)	D^2 [d]	F[e]	Accuracy (Tr/Ts)	Sensitivity (Tr/Ts)	Specificity (Tr/Ts)
Constitutional	0.67	0.54/0.55	2.01	23.8	77.3/78.1	78.5/82.4	80.1/78.2
Charge & molecular properties	0.68	0.54/0.62	1.89	41.2	77.1/81.0	78.8/85.1	79.6/80.8
2D Autocorrelation	0.67	0.53/0.55	1.94	42.5	77.1/77.3	81.8/77.0	77.9/80.3
Getaway	0.61	0.62/0.59	2.62	31.0	81.4/79.6	83.8/79.7	82.7/81.4
All	**0.58**	**0.62/0.68**	**2.86**	**41.5**	**81. 6/83.9**	**82.5/83.8**	**83.9/86.1**

[a] Appears only the best classification model by family of descriptors; [b] Wilk's lambda; [c] Mathew correlation coefficient (Training/Test); [d] Mahalanobis distance; [e] Fisher ratio.

Table 2. Molecular descriptors of the best QSAR classification model reported in this study.

Descriptors	Meaning
Ms	Mean electrotopological state
DELS	Molecular electrotopological variation
BELp3	Lowest eigenvalue n. 3 of Burden matrix/weighted by atomic polarizabilities
E1e	1st component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities
H3e	H autocorrelation of lag 3/weighted by atomic Sanderson electronegativities
nArCONHR	Number of secondary amides (aromatic)
nArNH2	Number of primary amines (aromatic)
TPSA(NO)	Topological polar surface area using N,O polar contributions
GVWAI-50	Ghose–Viswanadhan–Wendoloski drug-like index at 50%

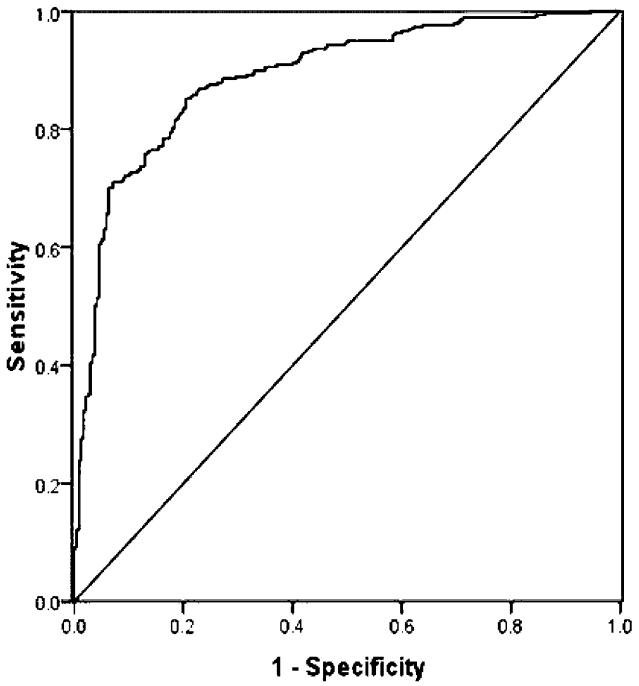

Figure 1. Receiver Operating Characteristic (ROC) curve for the classification model (Equation 3).

3.2 Interpretation of Molecular Descriptors

The electrotopological state indices (*Ms* and *DELS*) are numerical values associated with each atom in a molecule that encode information about the topological and electronic environment of that atom resulting from all the other atoms in the molecule. The constitutional descriptor *Ms* is related with electronic and steric attributes of atoms in molecules.[32] This descriptor is calculated as follow: $Ms = Ss/nsK$, where *Ss* is the sum of the Kier–Hall electrotopological states and *nsK* is the number of non-hydrogen atoms. This means that compounds with high electronic density and few polar heteroatoms should have good permeability properties. Meanwhile those compounds with more polar heteroatoms and less aromatic atoms will have higher hydrogen bonding capacity, reducing the permeability value. This result is in agreement with previously derived models where the hydrogen bonding was detrimental for permeability phenomena.[33,34]

DELS is a topological descriptor that is calculated by the following equation: $DELS = \Sigma\triangle I_i$, where $\triangle I_i = \Sigma(I_i - I_j)/(d_{ij}+1)^2$.[35] In this equation the sum runs over all the other atoms in the molecular graph, *I* is the atomic intrinsic state and *d* the topological distance between the two considered atoms. The intrinsic state of an atom is calculated as the ratio between Kier–Hall atomic electronegativity and the vertex degree, encoding information related to both partial charges of atoms and their topological position relative to the whole molecule. This descriptor reports about the relative availability of electrons accessible to intermolecular interactions.[36] Compounds with high electronic density will increase binding affinity, decreasing the rate of permeability.[37]

BELp3 is a Burden-CAS- University of Texas eigenvalue descriptor (BCUT)[38] that gives information about molecular complexity and polarizability. The variables *E1e* and *H3e*, a WHIM,[39] and GETAWAY descriptors[40] weighted by atomic Sanderson electronegativity, give information on atomic distribution and electronic density along the main direction of the molecule, respectively. Both descriptors contemporaneously characterize molecular size and polarizability. It is interesting to note that all these variables have a negative contribution to permeability, remarking the role of electrostatic interaction on low drug permeability values.

The structural fragment *nArCONHR* is a functional group count descriptor with a negative contribution to permeability. Secondary aromatic amides contribute significantly to increase the intermolecular interactions by strong H-bonds, The dependence of permeation on the polar surface area is thus consistent with the limiting role of amide hydration on membrane permeation.[41]

The functional group count descriptor $nArNH_2$ has a positive contribution to permeability. This fact is explained because of primary aromatic amines have their ion pair electrons conjugated into the benzene ring, decreasing their hydrogen bonding capacity and consequently the water solubility.

It has been widely proven that polar surface area is a very significant descriptor for drug transport properties such as human intestinal permeation and blood-brain barri-

er penetration.[42, 42b] *TPSA(NO)* is defined as the part of the surface area of the molecules associated with N and O and the H bonded to any of these atoms, and is related to the hydrogen bonding ability of the molecule. In this study the permeability values decrease with the increase of *TPSA(NO)*. These results are in concordance with other published studies,[43–45] where it has been suggested that PSA would describe the desolvation of a compound as it moves from aqueous to lipid environment, affecting the permeability value.[46]

Finally, the GVWAI50 is a drug-like index with information related to chemical structure similar to drugs.[47] In this case, a positive value of this descriptor means better permeability and, consequently, better absorption and pharmacokinetic properties.

3.3 Applicability Domain

The applicability of the final QSAR model was verified considering chemical domain, in order to obtain predicted data reliable only for structurally similar chemicals. In fact, in the case of structurally dissimilar molecules, the data predicted by the model must be considered as extrapolations. The applicability chemical domain of the reported model was verified by an analysis of the Williams plot (see Figure 2), in which the standardized residuals versus the leverage value (h) were plotted.

For the training set fifteen *influential* chemicals, with leverage values higher than the warning leverage (h^*) value of 0.056 were detected. Frequently, this kind of compounds is assumed to be erroneous data, or observations that cannot be explained although they are commonly valid observations and one of the most interesting parts of the dataset. A deeper exploration of the data composition reveals that among these compounds appears Amiloride, Foscarnet, Lamotrigine, Clodronate and Urea. The absorption process of these small hydrophilic compounds is mediated by carrier or follows several transport mechanisms. Other compound such as Leuprolide, a charged hydrophilic molecule with a high molecular weight, follows a passive diffusion transport throught paracellular pathway.[48] In the case of Paclitaxel, lipofilic molecule with high molecular weight, its permeability is affected by efflux mechanisms.[49]

It has been discussed that such compounds with high leverage and small residuals can be considered as "good leverage points" and can make the model more precise.[50] For this reason we decided to keep them inside the training data and hence to preserve the information they encode. These compounds in the training set reinforce the model. For the test set, only two compounds (Trimethoprim and Ribaudioside A) had high leverage values. The permeability for both molecules is mediated by carrier.[51,52] Any *response outlier* (>3 s) for the test set can be observed from Figure 2.

3.4 External Validation of the Classification Model and Virtual Screening

The most important criterion for the quality of the discriminant model is based on the statistics for the external set. In

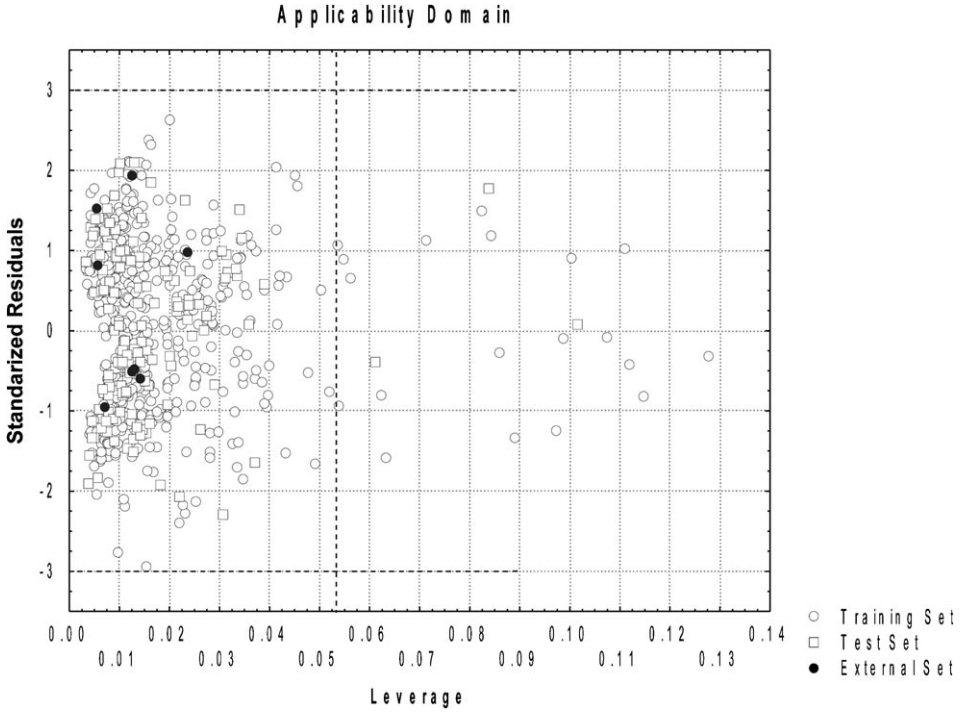

Figure 2. William's plot based on general classification model, for the training, test and external sets.

this sense, 10 compounds belonging to different groups of the BCS were predicted by the best LDA model. As can be seen in Figure 3 the permeability predictions for these compounds are reliable as all of them are inside the AD of the training set.

The classification results for the external set are shown in Table 3. The results evidenced an 80% of good classification. Only Norfloxacin and Vinblastine were classified wrongly. Norfloxacin active transephitelial transport and efflux have been demonstrated in Caco-2 cell.[53] Meanwhile Vinblastine permeability is affected by P-glycoprotein.[54]

Although a good relationship between the passive drug transport across Caco-2 cells and the absorbed fraction after oral administration to humans was obtained,[55] the oral drug absorption is influenced by many factors besides drug permeability, such as drug solubility, dissolution, active transport and, in some cases, presystemic metabolism.[56]

Taking into consideration these factors, a virtual search was simulated in the present study to evaluate the potentiality of our model to identify compounds with high human intestinal absorption ($HIA \geq 80$) from those molecules predicted with high permeability values ($P_{app} \geq 8 \times 10^{-6}$ cm/s). For this screening a dataset of 269 compounds with human intestinal absorption reported was used. The reliability of the prediction was validated checking if the dataset is inside the model's domain of applicability. The William's graph is shown in Figure 3.

Figure 3 shows that the glycopeptide Teicoplanin has a very large leverage value. Aminoglycosides and the antidiabetic Acarbose are *response outliers* as well as outside of the domain of applicability (>3 s and $h > h^*$). The permeability predictions for these very hydrophilic compounds, with a high molecular weight, are not reliable because these molecules are not located within the chemical space region occupied by the training set compounds.

The in silico model (Equation 3) predicted 121 compounds with high permeability in Caco-2 cells (See Supporting Information). From these compounds, 108 were well classified for a 90% of correspondence with the human reported data ($HIA \geq 80$).

Taking into consideration that absorption processes are strongly affected by molecular weight (*MW*) it is important to demonstrate the discriminant capacity of permeability models for molecules with $MW > 400$ due to high lipophilic molecules can become poorly permeable. In this sense a deeper analysis of our dataset evidenced that 225 compounds in training set had $MW > 400$. From those, 220 compounds were classified by the model and 5 unclassified. The general accuracy for training set was of 86% (189/220). In the test set 55 compounds had $MW > 400$ and 53 were classified. In this case our model classified 44 accurately (83%). These results evidenced the discriminant power of our classification model for compounds with high molecular weight.

3.5 Comparison with Other In Silico Classification Models

Although several computational models have been carried out to predict Caco-2 permeability, few of them are classification models. A direct comparison among previous reported studies is inappropriate because of differences in the

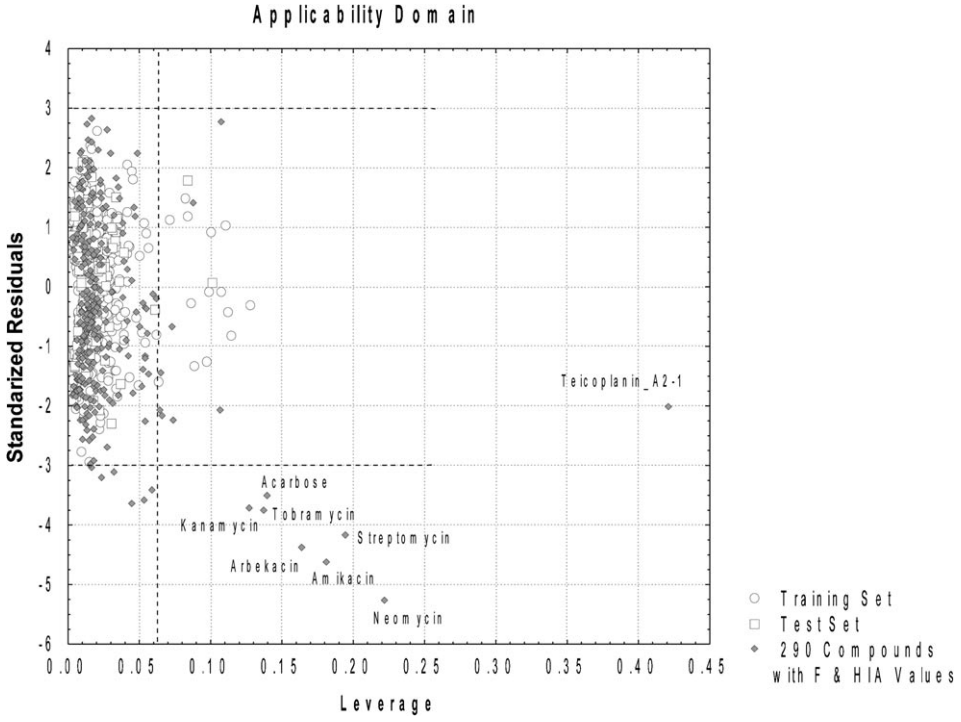

Figure 3. William's plot based on general classification model, for the training, test and virtual screening sets.

Table 3. In silico prediction of Caco-2 permeability and in vitro assessment for compounds of the external set. M-L: moderate-low permeability; H: high permeability.

Compound	Concentration (μM)	In silico pred[a]	In vitro P_{app} ($\times 10^{-6}$ cm/s)[b]	In vitro class[c]
Cimetidine	100	M-L	1.6	M-L
	1000		3.5	
Propranolol	100	H	25.2	H
	1000		28.1	
Norfloxacin	100	H	1.9	M-L
	1000		2.1	
Vinblastine	100	H	2.3	M-L
	1000		5.1	
Fexofenadine	100	M-L	0.1	M-L
	1000		1.9	
Atenolol	100	M-L	2.6	M-L
	1000		2.8	
Theophylline	10	H	9.4	H
	1000		24.9	
Carbamazepine	–	H	–	H
	1000		23.0	
Metoprolol	100	H	21.3	H
	1000		23.6	
Verapamil	100	H	21.0	H
	1000		35.0	

[a] Prediction by Equation 3; [b] Experimental permeability in Caco-2 cells; [c] Classification according to cut-off value selected for permeability;

data set, molecular descriptors and statistical methods. Nevertheless, a tentative comparison may point out advantages of the use of our model with respect to other studies.

Table 4 presents a comparison among these classification models and our approach. For example, models proposed by Marrero et al.[57,18] and Castillo-Garit et al.[58] have shown relative good statistical significance. However, their predictive capacity still remained questionable due to the small dataset used to obtain the models and the lack of a complete theoretical and experimental validation. Moreover to the complex mathematical nature of the topological descriptors used (based on graph theory), makes the interpretation of results the main challenge. On the other hand, Refsgaard et al.[59] proposed a methodology based on a large number of in vitro permeability data obtained in the same laboratory and used nine simple molecular descriptors. They also validated the obtained models with a quite large experimental set (obtained in MDCK cell assay). However, as they used an unpublished in-house dataset, the results remained questionable. In addition, the morphological differences among Caco-2 and MDCK cell lines should be taken into consideration when implementing the model validation. On the other hand, in our opinion, the selected Caco-2 P_{app} cut off value of 4×10^{-6} cm/s is not a reliable to predict human oral absorption through permeability models.

In our opinion, the most remarkable differences of our models, compared with other classification models, are the following: a) the models were developed based on the largest public data of Caco-2 permeability, b) the models have statistical significances as good as the previous approaches, c) the selected molecular descriptors are interpretable and d) the selected cut-off evidenced the screening capacity of the model to predict human intestinal absorption.

4 Conclusions

The in silico prediction of Caco-2 permeability has been always a challenge for drug design researchers. During the last decade many models have been developed but different predictive results have been achieved. Almost all the models developed are based on small datasets taken from literature. This has limited the structural variability of compounds included in the training sets, with a direct impact in the predictive capacity of the final models. Others models have been developed using extensive and reliable datasets belonging to large pharmaceutical companies, but this kind of data is not public for the scientific community, which limit the possibility to obtain more and better predictive permeability models.

In the present study the most comprehensive database of Caco-2 permeability is provided; the computational models show good results and the utility of the combined methodology (in silico + in vitro) to the classification and prediction of Caco-2 permeability was evidenced.

The results of this study suggest that the present methodology allows obtaining models with strong predictive ability that can be used in the design of large libraries of compounds with appropriate values of permeability and to

Table 4. Comparison with previously published Caco-2 permeability classifications models.

Reference	Method	Descriptors	Software	Data set (Tr/Ts)	Performance
Marrero-Ponce et al.[57]	LDA	Atom-based quadratic indices	TOMOCOMD-CARDD	33 (Training set) 18 (Test set)	$Q_{Training} = 87.9\,\%$ $Q_{Test} = 88.9\,\%$
Refsgaard et al.[59]	Nearest-neighbor classification	*m*Log*P*,[a] *MW*,[b] HD,[c] HA,[d] ROT,[e] *c*Log*P*,[f] VOL,[g] SURF,[h] PSA[i]	SYBYL6[j] SAVol 3.7	380 (Class 0) 332 (Class 1) 112 (External set)	$Q_{Class0} = 75.5\,\%$ $Q_{Class1} = 88.5\,\%$ $Q_{External\ set} = 84.6\,\%$
Marrero-Ponce et al.[18]	LDA	Atom-based quadratic indices	TOMOCOMD-CARDD	134 (Training set) 12 (Prediction set)	$Q_{Training} = 90.3\,\%$ $Q_{Predict.\ set} = 83.3\,\%$
Castillo-Garit et al.[58]	LDA	Atom-based nonstochastic linear indices	TOMOCOMD-CARDD DRAGON V2.1	138	$Q_{Training} = 90.6\,\%$ (Eq. 2) $Q_{Test} = 84.2\,\%$ (Eq. 2)
		Atom-based stochastic linear indices		138	$Q_{Training} = 90.3\,\%$ (Eq. 3) $Q_{Test} = 84.2\,\%$ (Eq. 3)
Current work	LDA	DRAGON descriptors	DRAGON V5.4	537 137 10 (External set)	$Q_{Training} = 81.6\,\%$ (Eq. 3) $Q_{Test} = 83.9\,\%$ (Eq. 3) $Q_{External\ set} = 80.0\,\%$

[a] Moriguchi partition coefficient; [b] Molecular weight; [c] Number of hydrogen bond donors; [d] Number of hydrogen bond acceptors; [e] Number of rotatable bonds in molecule; [f] Pomona College log*P* (water/octanol partition coefficient); [g] Total molecular volume ($Å^3$) after regularizing the molecular geometry (in 3-D); [h] Total molecular surface area ($Å^2$); [i] Polar surface area ($Å^2$).

perform virtual screening in the early stages of drug development. Nevertheless, other challenges should be later considered like the development of similar classification models using as a cut-off value the permeability of Metoprolol (20×10^{-6} cm/s) in order to approximate this in silico approach to the BCS. The combination of classification and regression statistical techniques could be considered as a good option to improve the accuracy of the final models.

Acknowledgements

The authors acknowledge financial support of *Agencia Española de Cooperación Iberoamericana para el Desarrollo* (*AECID*) to the project D/024153/09: Montaje de un laboratorio de Química Computacional, con fines académicos y científicos, para el diseño de potenciales candidatos a fármacos, en enfermedades de alto impacto social. M. A. C. P. also thanks to the program: *Estancias de movilidad de profesores e investigadores extranjeros en centros españoles* (SAB2009-0106), developed at Miguel Hernández University.

References

[1] R. A. Prentis, Y. Lis, S. R. Walker, *Br. J. Clin. Pharmacol.* **1988**, *25*, 387.
[2] T. Kennedy, *Drug Discov. Today* **1997**, *2*, 436.
[3] G. W. Caldwell, Z. Yan, W. Tang, M. Dasgupta, B. Hasting, *Curr. Top Med. Chem.* **2009**, *9*, 965.
[4] A. L. Ungell, *Drug Discov. Today Technol.* **2004**, *1*, 423.
[5] A. Avdeef, S. Bendels, L. Di, B. Faller, M. Kansy, K. Sugano, Y. Yamauchi, *J. Pharm. Sci.* **2007**, *96*, 2893.
[6] P. Artursson, K. Palm, K. Luthman, *Adv. Drug Deliv. Rev.* **2001**, *46*, 27.
[7] J. D. Irvine, L. Takahashi, K. Lockhart, J. Cheong, J. W. Tolan, H. E. Selick, J. R. Grove, *J. Pharm. Sci.* **1999**, *88*, 28.
[8] S. Tavelin, J. Taipalensuu, F. Hallbook, K. S. Vellonen, V. Moore, P. Artursson, *Pharm. Res.* **2003**, *20*, 373.
[9] H. Bohets, P. Annaert, G. Mannens, L. Van Beijsterveldt, K. Anciaux, P. Verboven, W. Meuldermans, K. Lavrijsen, *Curr. Top Med. Chem.* **2001**, *1*, 367.
[10] M. S. Ku, *AAPS J.* **2008**, *10*, 208.
[11] D. S. Wishart, *Drugs R. D.* **2007**, *8*, 349.
[12] C. A. Bergstrom, *Expert. Opin. Drug Metab. Toxicol.* **2005**, *1*, 613.
[13] K. Palm, K. Luthman, A. L. Ungell, G. Strandlund, P. Artursson, *J. Pharm. Sci.* **1996**, *85*, 32.
[14] H. van de Waterbeemd, G. Camenisch, G. Folkers, O. A. Raevsky, *Quan. Struct. Activ. Relat.* **1996**, *15*, 480.
[15] F. Yamashita, S. Fujiwara, M. Hashida, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 408.
[16] I. Hubatsch, E. G. Ragnarsson, P. Artursson, *Nat. Protoc.* **2007**, *2*, 2111.
[17] A. Dahan, J. M. Miller, J. M. Hilfinger, S. Yamashita, L. X. Yu, H. Lennernas, G. L. Amidon, *Mol. Pharm.* **2010**, *7*, 1388.
[18] Y. Marrero-Ponce, M. A. Cabrera, V. Romero, M. Bermejo, D. Siverio, F. Torrens, *Internet. Electron. J. Mol. Des.* **2005**, *4*, 124.
[19] *STATISTICA* (data analysis software system), Version 8, StatSoft, Inc., www.statsoft.com, **2007**.
[20] M. V. Varma, S. Khandavilli, Y. Ashokraj, A. Jain, A. Dhanikula, A. Sood, N. S. Thomas, O. Pillai, P. Sharma, R. Gandhi, S. Agrawal, V. Nair, R. Panchagnula, *Curr. Drug Metab.* **2004**, *5*, 375.
[21] M. Kansy, H. Fischer, K. Kratzat, F. Senner, B. Wagner, I. Parrilla, in *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies* (Eds: B. Testa, H. van de Waterbeemd, G. Folkers, R. H. Guy), Verlag Helvetica Chimica Acta, Zürich, Switzerland, **2001**, pp. 447.
[22] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, *DRAGON*, 5.4 ed., Talete srl, Milano, **2006**.

[23] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.
[24] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902.
[25] J. Frank, *MOPAC* 6.0 ed., Seiler Research Laboratory. U.S. Air Force Academy, **1993**.
[26] J. A. Cooper, R. Saracci, P. Cole, *Brit. J. Cancer* **1979**, *39*, 87.
[27] C. E. Metz, *Sem. Nuc. Med.* **1978**, *8*, 283.
[28] A. Golbraikh, A. Tropsha, *J. Mol. Graph. Model.* **2002**, *20*, 269.
[29] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **2003**, *111*, 1361.
[30] a) L. Saiz-Urra, M. P. Gonzalez, Y. Fall, G. Gomez, *Eur. J. Med. Chem.* **2007**, *42*, 64; b) L. Saiz-Urra, M. P. Gonzalez, I. G. Collado, R. Hernandez-Galan, *J. Mol. Graph. Model.* **2007**, *25*, 680.
[31] a) M. Hu, J. Chen, D. Tran, Y. Zhu, G. Leonardo, *J. Drug Target* **1994**, *2*, 79; b) M. Hu, J. Chen, Y. Zhu, A. H. Dantzig, R. E. J. Stratford, M. T. Kuhfeld, *Pharm. Res.* **1994**, *11*, 1405.
[32] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley VCH, Weinheim, Germany, **2000**.
[33] U. Norinder, T. Osterberg, P. Artursson, *Pharm. Res.* **1997**, *14*, 1786.
[34] U. Norinder, T. Osterberg, *J. Pharm. Sci.* **2001**, *90*, 1076.
[35] L. B. Kier, L. H. Hall, J. W. Frazer, *J. Math. Chem.* **1991**, *7*, 229.
[36] L. B. Kier, L. H. Hall, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 548.
[37] S. Agatonovic-Kustrin, R. Beresford, A. P. Yusof, *J. Pharm. Biomed. Anal.* **2001**, *25*, 227.
[38] F. R. Burden, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225.
[39] R. Todeschini, P. Gramatica, *Perspect Drug Discov. Des.* **1998**, *9/10/11*, 355.
[40] V. Consonni, R. Todeschini, M. Pavan, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
[41] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615.
[42] a) T. J. Hou, X. J. Xu, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137; b) T. Hou, J. Wang, W. Zhang, X. Xu, *J. Chem. Inf. Model.* **2007**, *47*, 208.
[43] H. van de Waterbeemd, G. Camenisch, G. Folkers, J. R. Chretien, O. A. Raevsky, *J. Drug Target* **1998**, *6*, 151.
[44] S. Winiwarter, F. Ax, H. Lennernas, A. Hallberg, C. Pettersson, A. Karlen, *J. Mol. Graph. Model.* **2003**, *21*, 273.
[45] J. Kelder, P. D. Grootenhuis, D. M. Bayada, L. P. Delbressine, J. P. Ploemen, *Pharm. Res.* **1999**, *16*, 1514.
[46] M. V. Varma, R. S. Obach, C. Rotter, H. R. Miller, G. Chang, S. J. Steyn, A. El-Kattan, M. D. Troutman, *J. Med. Chem.* **2010**, *53*, 1098.
[47] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Comb. Chem.* **1999**, *1*, 55.
[48] J. Guo, Q. Ping, G. Jiang, J. Dong, S. Qi, L. Feng, Z. Li, C. Li, *Int. J. Pharm.* **2004**, *278*, 415.
[49] U. K. Walle, T. Walle, *Drug Metab. Dispos.* **1998**, *26*, 343.
[50] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *Altern. Lab. Anim.* **2005**, *33*, 445.
[51] V. E. Thiel-Demby, J. E. Humphreys, L. A. St John Williams, H. M. Ellens, N. Shah, A. D. Ayrton, J. W. Polli, *Mol. Pharm.* **2009**, *6*, 11.
[52] J. M. Geuns, P. Augustijns, R. Mols, J. G. Buyse, B. Driessen, *Food Chem. Toxicol.* **2003**, *41*, 1599.
[53] N. M. Griffiths, B. H. Hirst, N. L. Simmons, *J. Pharmacol. Exp. Ther.* **1994**, *269*, 496.
[54] D. A. Laska, J. O. Houchins, S. E. Pratt, J. Horn, X. Xia, B. R. Hanssen, D. C. Williams, A. H. Dantzig, T. Lindstrom, *In Vitro Cell Dev. Biol. Anim.* **2002**, *38*, 401.
[55] P. Artursson, J. Karlsson, *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880.
[56] M. Rowland, T. N. Tozer, *Clinical Pharmacokinetics: Concepts and Applications*, 3rd ed., Lea & Febiger, Philadephia, **1995**, p. 601.
[57] Y. Marrero-Ponce, M. A. Cabrera Perez, V. Romero Zaldivar, H. Gonzalez Diaz, F. Torrens, *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
[58] J. A. Castillo-Garit, Y. Marrero-Ponce, F. Torrens, R. Garcia-Domenech, *J. Pharm. Sci.* **2008**, *97*, 1946.
[59] H. H. Refsgaard, B. F. Jensen, P. B. Brockhoff, S. B. Padkjaer, M. Guldbrandt, M. S. Christensen, *J. Med. Chem.* **2005**, *48*, 805.

Received: October 12, 2010
Accepted: January 16, 2011
Published online: March 31, 2011