[g009]

Estimation of ADME Properties in Drug Discovery: Predicting Caco-2 Cell Permeability Using Atom-Based Stochastic and Non-Stochastic Linear Indices

Juan A. Castillo-Garit,[1,2,3*] Yovani Marrero-Ponce,[2,3,4,5] Francisco Torrens,[4] and Ramón García-Domenech[5]

[1]*Applied Chemistry Research Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.* **Phone:** 53-42-281192, **Fax:** 53-42-281130, **e-mail:** jacgarit@yahoo.es, juancg.22@gmail.com or juancg@uclv.edu.cu
[2]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Department of Pharmacy, Faculty of Chemistry-Pharmacy, Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*
[3]*Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*
[4]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P. O. Box 22085, 46071 Valencia, Spain*
[5]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, València, Spain.*

Abstract

The *in vitro* determination of the permeability through cultured Caco-2 cells is the most often-used *in vitro* model for drug absorption. In this report, we use the largest data set of measured $\mathbf{P}_{\mathbf{Caco\text{-}2}}$, consisting of 157 structurally diverse compounds. Linear discriminant analysis (LDA) was used to obtain quantitative models that discriminate higher absorption compounds from those with moderate-poorer absorption. The best LDA model has an accuracy of 90.58% and 84.21% for training and test set. The percentage of good correlation, in the virtual screening of 241 drugs with the reported values of the percentage of human intestinal absorption (HIA), was greater than 81%. In addition, multiple linear regression models were developed to predict Caco-2 permeability with determination coefficients of 0.71 and 0.72. Our method compares favorably with other approaches implemented in the Dragon software, as well as other methods from the international literature. These results suggest that the proposed method is a good tool for studying the oral absorption of drug candidates.

Keywords: Caco-2 cells, Atom-Based Linear Indices, Dragon Software, Computational ADME, QSAR, '*in silico*' modeling, Virtual Screening, Human Intestinal Absorption.

Introduction

The molecular properties for absorption, distribution, metabolism, and excretion (ADME) are crucial for drug design. Following oral administration, a drug must pass through intestinal cell membranes via passive diffusion, carrier-mediated uptake, or active transport processes before reaching the systemic circulation. The development of many potential drugs has been discontinued because of their poor absorption. Several screening paradigms, which include absorption, have been employed to enhance the probability of success through the drug development stage. Methods to assess absorption rely on *in situ*, *in vivo*, *in silico*, or *in vitro* models used alone or in combination.[1,2]

Since oral is the most favorite way in various routines for drug delivery, estimating human oral bioavailability of candidates in the early stage of the drug development process is important and necessary for lead selection and optimization. Screening for absorption ability is an important part of assessing oral bioavailability, as well as attracts efforts from industry and academia.[3] In several *in vitro* cell culture models for drug absorption, the most widely used cell line is Caco-2 cells.[4] These are well-differentiated intestinal cells derived from human colorectal carcinoma. These cells retain many morphological and functional properties of the *in vivo* intestinal epithelial cell barrier, which makes the Caco-2 cell monolayer an important model for *in vitro* absorption screening. The permeability coefficient across Caco-2 cell monolayer (P) is increasingly used to estimate the oral absorption of new chemical entities (NCEs).[5-7] Nevertheless, Caco-2 cell models have several disadvantages,[8-10] being the long culture periods (21-24 days) the principal practical shortcoming of this approach, with consequently extensive cost.

A good correlation between the extent of oral drug absorption in humans and rates of transport across Caco-2 cell monolayers was obtained by Artursson and Karlsson.[7] However, there are several examples of application of Caco-2 cell models for prediction or correlation with human intestinal absorption, where the obtained results for Caco-2 cell permeability coefficients are influenced by the inter-laboratory differences.[8,10] Moreover, in the literature, there are several criteria about the use of Caco-2 cell permeability in the prediction of human absorption. Yazdanian et al.[6] reported that compounds with $\mathbf{P_{Caco\text{-}2}}$ values lesser than $0.4x10^{-6}$ cm/s exhibited very poor oral absorption, whereas compounds with $\mathbf{P_{Caco\text{-}2}}$ values greater than $7x10^{-6}$ cm/s had excellent oral absorption. In others reports, Artursson et al.[7] and Rubas et al.[11] described that compounds with $\mathbf{P_{Caco\text{-}2}}$ values over $1x10^{-6}$ and $70x10^{-6}$ cm/s, respectively, were completely absorbed in human. In addition, Chong et al.[12] concluded that compounds with $\mathbf{P_{Caco\text{-}2}}$ greater than $1x10^{-6}$ cm/s would have acceptable absorption in humans (>20%) and

Yee[13] determined that only compounds with $\mathbf{P_{Caco\text{-}2}}$ greater than $10x10^{-6}$ cm/s were well-absorbed in humans (70-100%). Based on these criteria, Chaturveldi et al.[14] have suggested that compounds with the following apparent permeability coefficients: $<1x10^{-6}$cm/s, 1-$10x10^{-6}$cm/s, $>10x10^{-6}$cm/s can be classified as poorly (0-20%), moderately (20-70%) and well (70-100%) absorbed drugs. Thus, Caco-2 permeability is a valuable index for assessing oral absorption of compounds, which, in turn, calls for methods for predicting chemical Caco-2 permeability.[3]

Table 1. Summary of Several Previously Published Caco-2 Permeability Prediction Investigations.

Year	Authors	Method	Software	Descriptors
1996	Palm *et al.*[21]	Linear Regression	PCMODEL, MacroModel	Dynamic polar surface area (PSAd)
1997	Norinder *et al.*[27]	PLS	MolSurf	Surface, logP, Polarity, HBAo[a], HBD[b], HBA, HBD, etc.
1998	Camenisch *et al.*[15]	Non-linear Regression	Statistica (statistical software)	MW, logD(oct)[c]
2000	Pickett *et al.*[28]	Not Mentioned	Chem-X, SYBYL	ClogP[d], MW, PSA[e]
2000	Cruciani *et al.*[37]	PLS	VolSurf	VolSurf Descriptors
2002	Kulkarni *et al.*[29]	membrane-interaction QSAR (MI-QSAR)	Chemlab-II, Mopac 6.0	Solute aqueous dissolution and salvation descriptors, Solute-membrane interaction and salvation descriptors, General intramolecular solute descriptors. (Many descriptors)
2002	Fujiwara *et al.*[30]	Molecular orbital (MO) calculation, 5-4-1 BP neural network	MOPAC97	Dipole moment, Polarizability, Sum(N)[f], Sum(O)[g], Sum(H)[h]
2002	Yamashita *et al.*[35]	Genetic Algorithm Based Partial Least Squares	Molconn-Z 3.50	Molconn-Z descriptors
2003	Marrero-Ponce *et al.*[31]	MRL	*TOMOCOMD-CARDD*	Atom-Based Quadratic Indices
2004	Hou *et al.*[32]	MRL	SYBYL, SASA, MSMS, etc.	HCPSA[i], logD, rgyr[j], RB
2004	Marrero-Ponce *et al.*[36]	Linear discriminant analysis (LDA)	*TOMOCOMD-CARDD*	Atom-Based Quadratic Indices
2005	Refsgaard *et al.*[34]	Nearest-Neighbor classification	SYBYL, Matlab	Number of flex bonds, number of hydrogen bond acceptors and donors, molecular and polar surface area
2005	Marrero-Ponce *et al.*[33]	LDA and MLR	*TOMOCOMD-CARDD*	Atom-Based Quadratic Indices
2006	Guangli and Yiyu.[3]	MLR and SVM	CDK	HBD[k], CPSA10[l], CPSA20[m], TPSA[n]

[a]Hydrogen bond acceptor strength for oxygen atoms. [b]Hydrogen bond donor strength. [c]distribution coefficient in 1-octanol/water. [d]Calculated logP. [e]Polar surface area. [f]Sum of charges of nitrogen atoms. [g]Sum of charges of oxygen atoms. [h]Hydrogen atoms bonding to nitrogen or oxygen atoms. [i]High charged polar surface area. [j]Radius of gyration. [k]Number of H-bond donors. [l]Partial positive surface area*total positive charge. [m]Charge weighted partial negative surface area*total molecular surface area. [n]Topological surface area based on fragments contribution.

At present, it is known that oral absorption is influenced by a different kind of interactions. In some studies, it has been demonstrated that permeability coefficients, measured for transport through Caco-2 monolayer cell cultures, are correlated with lipophilicity,[6,15-17] while in others it has been discussed the role of hydrogen bonding or charge.[7,10,15,16] Lipinski *et al*.[18] proposed "the Rule of Five" for a preliminary estimation of a compound's absorption on the basis of molecular weight, lipophilicity and the number of hydrogen bond donor and acceptor atoms in the molecule.

The significant failure rate of drug candidates in late stage development is driving the need for predictive tools that could eliminate inappropriate compounds, before substantial time and money is invested in testing.[19,20] Theoretical approaches appear to be a good alternative to '*in silico*' prediction of human absorption, for new drug candidates obtained by combinatorial chemistry method.[21-26] Therefore, it is expected an increasing use in the estimation of absorption parameters of potentially active NCEs by using quantitative structure property relationship (QSPR) methods during the drug discovery and development process. Research on predicting Caco-2 permeability from structures of compounds using quantitative structure property relationship (QSPR) modeling is on the way. In the last years, some researchers have explored QSPerR studies involving Caco-2 cell permeability. Table 1 summarizes most of these studies.[3,15,21,27-37] They generally used regression methods, small sets of molecules, and results were not fully validated by external test sets. Certainly, Hou *et al*.[32] published the largest set employed for regression analysis, about 100 compounds collected from several published datasets. On the other hand, some studies performed linear discriminant analysis (LDA) to classify between compounds with higher or moderate-poorer absorption; among them the largest training set, made up for 134 compounds, was published in 2005 by Marrero-Ponce *et al*;[33] Refsgaard et al.[34] have also built a classification model based on in-house data, but they cannot publish the dataset because at this time it is in discovery stage.

Bearing in mind that mentioned above the aims of the present work were: 1) to obtain a classification model, for the largest (as far as we know) and more heterogeneous compiled data set (157 compounds), which permits the classification of the molecules with poor-moderate and high absorptions from their molecular structures, by using non-stochastic and stochastic atom-based linear indices as molecular fingerprints, 2) to compare our result with other obtained by using topological descriptors implemented in the Dragon software,[38] with the same data set, 3) to simulate a virtual screening experiment with the obtained models, in order to find a relationship between the predicted permeability coefficients of Caco-2 cell and the human intestinal absorption. Finally, we will also build Caco-2 permeability prediction models based on multiple linear regressions (MLR) and by using the same data set, compare our results with other previously obtained by other researches.

Computational and Experimental Procedures

Computational strategies. Molecular fingerprints were generated by using the 'in house' ***TOMOCOMD*** software.[39] It is an interactive program for molecular design and bioinformatics research, which consists of four subprograms: ***CARDD*** (***C***omputed-***A***ided ***R***ational ***D***rug ***D***esign), ***CAMPS*** (*C*omputed-*A*ided *M*odeling in *P*rotein *S*cience), ***CANAR*** (*C*omputed-*A*ided *N*ucleic *A*cid *R*esearch), and ***CABPD*** (*C*omputed-*A*ided *B*io-*P*olymers *D*ocking). Each one of them allows both drawing the structures (drawing mode) and calculating molecular 2D/3D descriptors (calculation mode). In the present report, we outline salient features concerned with only one of these subprograms, ***CARDD,*** and with the calculation of non-stochastic and stochastic atom-based linear indices, considering and not considering H-atoms in the molecular pseudograph (G). The total and local linear indices for small-to-medium-sized organic compounds have been explained in some detail in the literature.[40-46]

The main steps for the application of the present method in QSAR/QSPR and drug design can be summarized briefly in the following algorithm:

1) Draw the molecular structure for each molecule in the data set, by using the software drawing mode. This procedure is performed by a selection of the active atomic symbol in the corresponding group of the periodic table of the elements;
2) Use appropriate weights in order to differentiate the atoms in the molecule. The weights used in this study are those previously proposed for the calculation of the DRAGON descriptors,[47-49] i.e., atomic mass (M), atomic polarizability (P), van der Waals atomic volume (V), Mulliken electronegativity (K), plus the atomic electronegativity in Pauling scale (G). The values of these atomic labels are shown in Table 2.[47-50]
3) Compute the total and local (atomic, group and atom-type) non-stochastic and stochastic linear indices. It can be carried out in the software calculation mode, where one can select the atomic properties and the descriptor family before calculating the molecular indices. This software generates a table in which the rows correspond to the compounds, as well as columns correspond to the atom-based (both total and local) linear maps or other MD family implemented into this program;
4) Find a QSPR/QSAR equation by using several multivariate analytical techniques, such as multilinear regression analysis (MRA), neural networks, linear discrimination analysis, and so on. Therefore, one can find a quantitative relation between a property **P** and the linear fingerprints having, for instance, the following appearance,

$$\boldsymbol{P} = a_0\boldsymbol{f}_0(x) + a_1\boldsymbol{f}_1(x) + a_2\boldsymbol{f}_2(x) + \ldots + a_k\boldsymbol{f}_k(x) + c \quad \textbf{(1)}$$

where $\boldsymbol{P}$ is the measure of the property, $\boldsymbol{f}_k(x)$ is the k^{th} total (atom and atom-type) linear indices, as well as the $\boldsymbol{a}_k$*'s* and c are the coefficients obtained by MRA.

5) Test the robustness and predictive power of the QSPR/QSAR equation, by using internal (cross-validation) and external validation techniques.

The descriptors calculated in this work were the following:

i) $\boldsymbol{f_k}(x)$ and $\boldsymbol{f_k}^{H}(x)$ are the k^{th} atom-based total linear indices considering and not considering H-atoms, respectively, in the molecule.

ii) $\boldsymbol{f_k}_L(x_E)$ and $\boldsymbol{f_k}_L^{H}(x_E)$ are the k^{th} atom-based local (atom-type = heteroatoms: S, N, O) linear indices considering and not considering H-atoms, respectively, in the molecule.

iii) $\boldsymbol{f_k}_L^{H}(x_{E\text{-}H})$ are the k^{th} local (atom-type = H-atoms bonding to heteroatoms: S, N, O) linear indices considering H-atoms in the molecular pseudograph (G).

The k^{th} stochastic total [$^{s}\boldsymbol{f_k}(x)$ and $^{s}\boldsymbol{f_k}^{H}(x)$], as well as local [$^{s}\boldsymbol{f_k}(x_E)$, $^{s}\boldsymbol{f_k}^{H}(x_E)$ and $^{s}\boldsymbol{f_k}_L^{H}(x_{E\text{-}H})$] atom-based linear indices were also computed.

The Dragon[38] computer software was also employed to calculate the molecular descriptors. The Dragon descriptors included in this work were the following groups: Topological indices, Molecular walk counts,[51,52] BCUT descriptors,[53] Gálvez´s topological charge indices,[54,55] and 2D autocorrelations. The complete list of 2D molecular descriptors used in this study as well as the symbols and their description is given as Supplemental Material.

Table 2. Values of the Atom Weights Used for Linear Indices Calculation.[47-50]

ID	Atomic Mass	VdW Volume	Mulliken Electronegativity	Polarizability	Pauling Electronegativity
H	1.01	6.709	2.592	0.667	2.2
B	10.81	17.875	2.275	3.030	2.04
C	12.01	22.449	2.746	1.760	2.55
N	14.01	15.599	3.194	1.100	3.04
O	16.00	11.494	3.654	0.802	3.44
F	19.00	9.203	4.000	0.557	3.98
P	30.97	26.522	2.515	3.630	2.19
S	32.07	24.429	2.957	2.900	2.58
Cl	35.45	23.228	3.475	2.180	3.16
Br	79.90	31.059	3.219	3.050	2.96
I	126.90	38.792	2.778	5.350	2.66

Permeability Data. In this study, we used the largest, from our knowledge, data set of measured $\mathbf{P}_{\mathbf{Caco\text{-}2}}$ consisting of 157 structurally diverse compounds. It was compiled from several published works[6,7,11-13,15,32,56-72] and divided into two subsets, used as training and test sets. The compounds were classified into high and moderate-poor absorbed compounds according to a boundary quantitative value of $\mathbf{P}_{\mathbf{Caco\text{-}2}}(8\times10^{-6}$ cm/s). This value of $\mathbf{P}_{\mathbf{Caco\text{-}2}}$ was fixed taking into consideration the experimental results reported in the literature and the wide inter-laboratory variability.[8,10] Experimental values of $\mathbf{P_{Caco\text{-}2}}$ (AP→BL), for the training data set, are illustrated in Table 3 with their references. The data set used for '*in silico*' permeability studies included compounds with a diverse molecular weight and net charge. In addition compounds with different absorption mechanisms were included in the model. Even, an external prediction set (test set) was used to validate our model. It was composed by 19 (five Sulfasalazine's metabolites and analogs, five 6–fluoroquinolones derivatives and other seven compounds) drugs–like compounds obtained from several previously published works. The experimental values of $\mathbf{P_{Caco\text{-}2}}$ (AP→BL) (APical→BasoLateral), for the test data set, are also shown in Table 3.

The decimal logarithms of the apparent Caco-2 permeability coefficients (log P_{app}), of 100 structurally diverse compounds, were taken from the literature.[32] To compare with previous works,[3,32] the data set was separated into a training set of 77 compounds and into a test set of 23 compounds, as in Hou *et al*.[32] The training set was used to build the model, and the test set was used to evaluate its predictability.

Chemometric analysis. Statistical analysis was carried out with the STATISTICA software.[73] The considered tolerance parameter (proportion of variance that is unique to the respective variable) was the default value for minimum acceptable tolerance, which is 0.01. Forward stepwise procedure was fixed as the strategy for variable selection. 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, but having as few parameters (a_k) as possible.

A linear discriminant analysis (LDA) was performed to classify between compounds with poorer-moderate and higher absorption. The quality of the models was determined by examining Wilks' λ parameter (*U*-statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and its corresponding *p*-level (p(F)), as well as the percentage of good classification in the training and test sets. The statistical robustness and predictive power of the obtained model were assessed by using an external prediction (test) set. In developing classification models, the values of 1 and -1 were assigned to active (high absorption) and inactive (moderate-poor absorption) compounds, respectively. By using the models, one compound can then be classified as either active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{Active}) - P(\text{Inactive})]x100$, or as inactive, otherwise. P(Active) and P(Inactive) are the probabilities with which the equations classify a compound as active and inactive, correspondingly.

Table 3. Caco-2 Cell Permeability for 157 Structurally Diverse Compounds.

Compounds	Ref.	Px10^{-6}	Compounds	Ref.	Px10^{-6}	Compounds	Ref.	Px10^{-6}
Acebutolol	32	1.48	Cimetidine	32	1.29	Imipramine	32	14.13
	58	4.47		60	0.50		13	14.13
	6	0.51		13	3.09	Indomethacin	6	20.42
Acebutolol ester	32	24.55		6	1.38		6	25.12
	58	77.62		68	0.35	Labetalol	6	9.33
Acetylsalicylic acid	32	8.71	Clonidine	32	25.70		6	10.00
	56	2.40		13	30.20	Mannitol	32	0.62
	13	30.90		6	21.88		6	
	6	9.12	Corticosterone	32	33.88		56	0.18
Aciclovir	32	0.70		56	54.95		11	3.24
	13	2.00		6	21.38		13	0.65
	6	0.25	Desipramine	32	22.91		6	0.40
Alprenolol	32	23.99		13	21.38		61	2.63
	56	40.74		6	24.55		68	0.12
	7	40.74	Dexamethasone	32	17.78	Meloxicam	6	19.50
	58	75.86		56	12.59		6	19.95
	6	25.12		57	26.92	Methanol	32	131.83
	15	38.30		13	23.44		11	131.83
	69	242.00		6	12.30	Methotrexate	32	1.20
Alprenolol ester	32	33.88	Dexamethasone-β-D-glucoside	32	0.44		13	1.20
	58	107.15		57	0.44	Methylscopolamine	6	0.69
Aminopyrine	32	36.31	Dexamethasone-β-D-glucuronide	32	1.15		6	0.79
	6	36.31		57	1.15	Metoprolol	32	25.70
Artemisinin	32	30.20	Diazepam	32	47.86		6	
	59	30.40		13	70.79		56	26.92
Artesunate	32	3.98		6	33.11		7	26.92
	59	4.00		69	756.00		68	18.80
Atenolol	32	0.32	Dopamine	32	9.33		15	26.71
	56	0.20		6	9.33		69	92.00
	7	0.20	Doxorubicin	32	0.16	Nadolol	32	3.89
	60	0.47		13	0.16		6	3.89
	6	0.52	Erithromycin	32	3.72	Naproxen	32	74.13
	61	0.23		13	3.72	Nevirapine	32	30.20
Betaxolol	58	15.49	Estradiol	32	16.98		6	30.20
	58	95.50		6	19.95	Nicotine	32	19.50
	68	0.19	Felodipine	32	22.91		6	19.50
	15	0.20		56	22.91	Olsalazine	56	0.11
	69	1.00	Ganciclovir	32	1.02	Oxprenolol	32	20.89
Betaxolol ester	32	30.20		11	2.69		58	66.07
	58	95.50		6	0.38		69	120.00
Bremazocine	32	7.94	Griseofulvin	32	36.31	Oxprenolol ester	32	30.90
	6	7.94		6	36.31		58	97.72
Caffeine	32	38.91	Hidrochlorothiazide	32	0.87	Phencyclidine	32	24.55
	13	50.12		60	1.51		6	24.55
	6	30.90		6	0.51	Phenytoin	32	26.92
Chloramphenicol	32	20.42	Hydrocortisone	32	21.88		6	26.92
	13	20.42		6		Pindolol	32	16.60
Chlorotiazide	32	0.19		56	21.38		6	16.60
	6	0.19		11	35.48	Pirenzepine	32	0.44
Chlorpromacine	32	19.95		6	15.85		6	0.44
	6	19.95	Ibuprophen	32	52.48			
Cephalexin	59	0.18		13	52.48			

Table 3. *Cont.*

Compounds	Ref.	$Px10^{-6}$	Compounds	Ref.	$Px10^{-6}$	Compounds	Ref.	$Px10^{-6}$
Piroxicam	32	35.48	Timolol	32	14.13	Bosentan	15	1.05
	68	46.25		58	44.67		61	1.05
Practolol	32	0.89		6	12.88	Proscillaridin	15	0.63
	56	0.89	Timolol ester	32	25.12		61	0.63
	7	0.89		58	79.43	Ceftriaxone	15	0.13
	61	1.38	Uracil	32	4.27		61	0.13
	15	0.92		6	4.27	Remikiren	15	0.74
	69	3.50	Urea	32	4.57		61	0.74
Propranolol	32	26.30		6	4.57	Squinavir	15	0.55
	56	41.69	Warfarin	32	28.18		61	0.55
	7	41.69		56	38.02	Mibefradil	15	13.49
	58	83.18		6	20.89		61	13.49
	13	27.54		61	53.70	Azithromycin	13	1.05
	6	21.88	Ziduvudine	32	6.92	Penicillin	13	1.95
	59	11.20		6	6.92	H216/44	7	0.91
	15	43.03	PNU200603	61	0.14		70	1.14
Prazocin	32	43.65	Amoxicillin	62	0.79	Naloxone	13	28.18
	13	43.65	Antipyrine	62	28.18	Sumatriptan	13	3.02
Progesterone	32	42.66		68	47.23	Taurocholic acid	13	34.67
	11	79.43		69	2.15	Tenidap	13	51.29
	6	23.44	Cyclosporine	62	0.89	Trovafloxacin	13	30.20
Propranolol ester	32	33.11	Diltiazem	62	41.69	Valproic acid	13	47.86
	58	104.71		61	48.98	Ziprasidone	13	12.30
Quinidine	32	20.42	Enalapril	62	2.29	Gly-Pro	71	6.10
	13	20.42	Furosemide	62	0.31	D-glucose	71	17.53
Ranitidine	32	0.49		68	0.29	L-Phenylalanine	71	18.37
	6	0.49	Guanabenz	62	31.62	Ketoprofen	71	23.15
Salicylic acid	32	16.22		61	72.44	SB 209670	69	8.80
	56	12.02	Cumarin	15	77.63	SB 217242	69	70.00
	6	21.88		61	77.62	Raffinose	69	0.05
Scopolamine	32	11.75	Theophyline	15	44.67	Sildenafil	69	87.00
	6	11.75		61	44.67	Oxazepam	69	246.00
Sucrose	32	1.70	Epinephrine	15	0.96	Nordazepam	69	307.00
	6	1.70		61	0.95	Metolazone	69	6.10
Sulphasalazine	32	0.20	Guanoxan	15	19.50	Lactulose	69	0.27
	56	0.13		61	19.50	Foscarnet	69	0.05
	6	0.30	Lidocaine	15	61.66	Azithromycin	13	1.05
Telmisartan	32	15.14		61	61.66	Ciprofloxacin	69	1.90
	6	15.14	Tiacrilast	15	12.59	Amiloride	69	0.78
Terbutaline	32	0.42		61	12.59	Alfentanil	69	310.00
	56	0.38	Sulpiride	15	0.69	H/334/75	71	0.13
	6	0.47		69	0.39	Lef 553	69	0.02
	68	0.97	Nitrendipine	15	16.98	Glycine	13	80.00
Testosterone	32	45.71		61	16.98	D-Phe-L-Pro	13	44.30
	56	51.29	Fleroxacin	15	15.49	BvaraU	12	4.00
	13	72.44		61	15.49	Lisinopril	72	0.05
	6	25.12	Verapamil	15	26.30	SQ-29852	12	0.02
	61	58.88		61	26.30	L-Glutamine	72	0.85

Table 3. *Cont.*

Compounds	Ref.	$Px10^{-6}$	Compounds	Ref.	$Px10^{-6}$	Compounds	Ref.	$Px10^{-6}$
Atorvastatin	65	4.90	Amprenavir	63	21.60	CNV97103	64	15.20
DMP450	63	36.80	Indinavir	63	6.00	CNV97104	64	13.90
DMP850	63	12.40	Nelfinavir	63	3.40	Nicardipine	67	19.80
DMP851	63	5.20	Ritonavir	63	3.90	Sulfapyridine	66	21.50
Losartan	63	0.88	Vinblastine	67	3.83	5-Aminosalicylic acid	66	0.32
Lucifer Yellow	63	0.05	CNV97100	64	3.60	Descarboxysulfasalazine	66	32.80
Phenol red	63	0.41	CNV97101	64	21.80	Homosulfasalazine	66	0.09
Rhodamine	63	0.55	CNV97102	64	16.80	Sulfasalamide	66	1.96

Calculation of percentages of global good classification (accuracy) and Matthews' correlation coefficients (MCC), in the training and test sets, permitted the assessment of the model.[74] The MCC is always between -1 and +1. A value of -1 indicates total disagreement (all-false predictions), and +1, total agreement (perfect predictions). The MCC is 0 for completely random predictions and, therefore, it yields an easy comparison with regard to random baseline.

Finally, we have used the receiver operating characteristic curve (ROC) to evaluate the accuracy of the discriminant functions selected (DFs) through the sensitivity (true positive fraction) and specificity (true negative fraction) for different thresholds of DFs. ROC curve is the representation of sensitivity vs 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.[75]

On the other hand, *multiple linear regression* (MLR) was carried out to predict Caco-2 permeability. The quality of the models was determined by examining the regression's statistic parameters and those of the cross validation procedures.[76,77] Therefore, the quality of models was determined by examining the correlation coefficients, squared correlation coefficients, Fisher-ratio's p-level [p(F)], standard deviation of the regression (s) and the leave *one*-out cross-validated statistic (q^2). The statistical robustness and predictive power of the obtained model were assessed by using an external prediction (test) set.

Finally, an *Analysis of Principal Components* (ACP) was performed to compare the information content of the atom-based non-stochastic and stochastic total and local linear indices with that of other descriptors used in this study. The existence of linear independence has been claimed by Randic[78] as one of the desirable attributes for novel topological indices.

To conduct this analysis, we will carry out a factor analysis by using the principal components method. The theoretical aspects of this statistical technique have been extensively exposed in the literature including many chemical applications.[73,79-85] The main applications of factor analytic techniques are: to *reduce* the number of variables, and 2) to *detect structure* in the relationships between variables, that is to *classify* variables.[73,82] In this approach, factor loadings (or "new" variables) are obtained from original (molecular descriptors) variables. Thus, these factors capture most of the "essence" of these molecular descriptors because they are a linear combination of the original items. Because each consecutive factor is defined to maximize the variability that is not captured by the preceding factor, consecutive factors are independent of each other. Put another way, consecutive factors are uncorrelated or orthogonal to each other. In this sense, the first factor obtained is generally more highly correlated with the variables than the others factors. This is to be expected because, as previously described, these factors are extracted successively and will account for less and less variance overall. Finally, some of the most important conclusions that can be drawn from a factor analysis that will be of large usefulness in the present paper are the following: [73,79-85] 1) variables with a high loading in the same factor are interrelated and will be the more so the higher the loadings, and 2) no correlation exists between variables having nonzero loadings only in different factors. These are the principal ideas that permit the interpreting the *factor structure* obtained using the factor analysis as a classification method.

The factor analysis was performed with the STATISTICA software,[73] and “varimax normalized” was used as rotational strategy to obtain the factor loadings from the principal component analysis. The goal of this rotational procedure is to obtain a clear pattern of loadings, that is, factors that are somehow clearly marked by high loadings for some variables and low loadings for others. The “varimax normalized” is the method most commonly used as “varimax” rotation.[73] This rotation strategy is aimed at maximizing the variances of the squared *normalized factor loadings* (row factor loadings divided by squared roots of the respective communalities) across variable for each factor. This strategy makes the structure of factors pattern as simple as possible, permitting a clearer interpretation of the factors without loss of orthogonality between them.[73,85-87]

Results and Discussion

Linear Discriminant Analysis Models.

QSPerR Models Obtained by Using TOMOCOMD-CARDD Descriptors. In order to develop the LDA models, the data was conformed by 82 compounds with higher absorption (**P** $\geq 8x10^{-6}$cm/s) and 56 compounds with moderate-poorer absorption (**P** $< 8x10^{-6}$cm/s). The best obtained discrimination models are given below, together with the LDA statistical parameters:

$$\boldsymbol{Class} = 4.533 - 0.364\, {}^{G}\boldsymbol{f}_{3L}{}^{H}(x_{E\text{-}H}) + 0.130\, {}^{G}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H}) + 0.159\, {}^{K}\boldsymbol{f}_{4L}{}^{H}(x_{E\text{-}H}) - 0.032\, {}^{V}\boldsymbol{f}_{1L}{}^{H}(x_E) + 1.4\text{x}10^{-8}\, {}^{K}\boldsymbol{f}_{12}{}^{H}(x) + 0.065\, {}^{V}\boldsymbol{f}_{0L}(x_E) - 0.004\, {}^{V}\boldsymbol{f}_{5L}{}^{H}(x_{E\text{-}H}) + 1.1\text{x}10^{-4}\, {}^{M}\boldsymbol{f}_{5}(x_E) \qquad (\mathbf{2})$$

$N = 138$ $\lambda = 0.435$ $Q_{Total} = 90.58$ $MCC = 0.81$

$D^2 = 5.31$ $F(8,129) = 20.94$ $p < 0.00001$

$$\boldsymbol{Class} = 4.652 - 0.318\, {}^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H}) - 0.483\, {}^{Vs}\boldsymbol{f}_{3}(x_E) + 0.095\, {}^{Vs}\boldsymbol{f}_{14}(x_E) + 3.058\, {}^{Ps}\boldsymbol{f}_{5}(x_E) + 2.065\, {}^{Ks}\boldsymbol{f}_{2}{}^{H}(x_{E\text{-}H}) - 3.273\, {}^{Ps}\boldsymbol{f}_{4L}{}^{H}(x_{E\text{-}H}) + 0.149\, {}^{Vs}\boldsymbol{f}_{5L}{}^{H}(x_E) + 0.026\, {}^{Ks}\boldsymbol{f}_{0}(x) \qquad (\mathbf{3})$$

$N = 138$ $\lambda = 0.436$ $Q_{Total} = 89.13$ $MCC = 0.77$

$D^2 = 5.29$ $F(8,129) = 20.85$ $p < 0.00001$

where N is the number of compounds, λ is the Wilks' statistic, Q_{Total} is the accuracy of the model for the training set, MCC is the Matthews' correlation coefficient, D^2 is the squared Mahalanobis distance, F is the Fisher ratio and *p*-value is its significance level.

The non-stochastic model (Eq. **2**), which includes non-stochastic indices, has an accuracy of 90.56% for the training set. This model showed a high MCC of 0.81; MCC quantifies the strength of the linear relation between the molecular descriptors and the classifications, and it may often provide a much more balanced evaluation of the prediction than, for instance, the percentages (accuracy).[88] Nevertheless, the most important criterion, for the acceptance or not of a discriminant model, is based on the statistic for external prediction set. The non-stochastic model showed an accuracy of 84.21% (MCC = 0.72) for the compounds in the test set. In Table 4, the results of classification and a posteriori probabilities for the compounds of the training and test sets are shown.

A rather similar behavior was obtained with the stochastic linear indices (Eq. **3**). In this case, the model achieved an accuracy of 89.13% with a MCC of 0.77; for the test set the results of this model were an accuracy of 84.21% and MCC of 0.72; these values are similar to those obtained with non-stochastic linear indices. The results of classification and a posteriori probabilities, with stochastic linear indices, are also shown in Table 4.

Table 4. Results for the Classification of Compounds in Training and Test Sets thought the Discriminant Functions Obtained Using Non-stochastic and Stochastic Linear Indices.

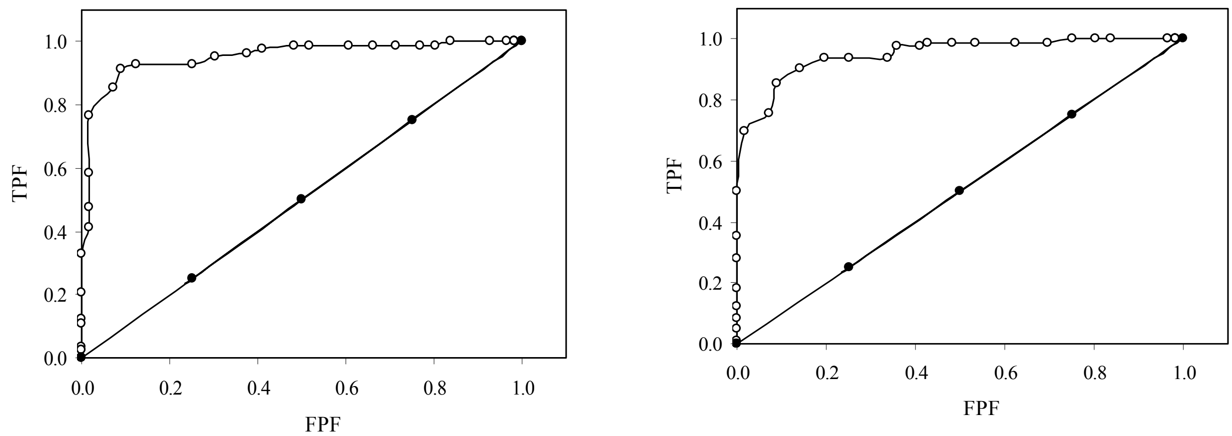
Compounds	ΔP%[a]	ΔP%[b]	Compounds	ΔP%[a]	ΔP%[b]	Compounds	ΔP%[a]	ΔP%[b]
				***High absorption group* (H)**				
				Training set				
Acebutolol ester[c,d]	-58.30	-89.37	Metoprolol	76.63	82.69	Guanoxan[c,d]	-78.60	-71.22
Acetylsalicylic acid	58.10	37.92	Naproxen	90.60	90.60	Lidocaine	38.96	66.66
Alprenolol	88.48	89.00	Nevirapine	59.43	71.41	Tiacrilast	70.76	55.65
Alprenolol ester	79.64	76.34	Nicotine	95.81	98.16	Nitrendipine	87.83	94.18
Aminopyrine	95.07	97.28	Oxprenolol	73.08	82.52	Fleroxacin	33.34	57.66
Artemisin	92.40	91.76	Oxprenolol ester	75.60	68.55	Verapamil	95.27	98.93
Betaxolol	70.90	82.52	Phencyclidine	98.35	98.98	Mibefradil	50.10	51.53
Betaxolol ester	73.59	66.80	Phenitoin[c,d]	-74.77	-21.01	Naloxone	80.72	74.76
Bremazocine	92.39	89.84	Pindolol[c]	-0.22	0.14	Taurocholic acid[c,d]	-99.67	-98.99
Caffeine	66.46	79.75	Piroxicam	86.35	93.63	Tenidap	81.93	75.77
Chloranphenicol	63.57	43.89	Prazocin	44.15	77.59	Trovafloxacin	58.67	25.80
Chlorpromazine	99.68	99.20	Progesterone	98.29	97.71	Acid valproic	68.42	70.46
Clonidine[d]	48.27	-6.76	Propranolol	92.04	89.46	Ziprasidone	85.00	84.55
Corticosterone	62.43	81.91	Propranolol ester	92.86	79.55	D-Glucose[c,d]	-58.60	-69.69
Desipramine	92.21	91.92	Quinidine	90.51	94.42	L-Phenylalanine	86.11	91.76
Dexamethasone	55.54	62.89	Salicylic acid	27.53	20.02	Ketoprofen	93.22	90.71
Griseofulvin	98.74	94.18	Scopolamine	59.92	70.51	SB 209670	10.72	25.04
Hydrocortisone	32.07	46.07	Telmisartan	97.80	92.56	SB 217242	57.77	79.71
Ibuprophen	86.28	91.16	Testosterone	94.73	94.28	Sidenafil	13.10	74.43
Imipramine	98.61	99.15	Timolol	24.14	50.00	Oxazepam	79.94	72.60
Indomethacin	96.19	76.12	Timolol ester	29.36	30.86	Nordazepam	94.50	80.32
Labetalol[c,d]	-88.97	-72.53	Warfarin	95.84	79.31	Alfentanil	51.01	43.08
Meloxicam	60.78	86.68	Theophylline[c]	-12.33	7.42	Glycine	59.00	64.16
Methanol	89.49	85.68						
				Test set				
CNV97101	39.78	59.63	CNV97103	42.11	62.82	Nicardipine	85.64	96.34
CNV97102	41.08	57.51	CNV97104	42.80	64.57	Sulfapyridine	31.89	59.85
Descarboxysulfasalazine	49.39	69.93						
				***Moderate-poor absorption group* (M-P)**				
				Training set				
Acebutolol	-61.90	-81.25	Practolol	-38.01	-46.41	Ziduvudine	-97.99	-98.26
Acyclovir	-99.28	-96.52	Ranitidine	-24.88	-39.19	Urea	-76.04	-33.50
Artesunate[c,d]	89.10	62.25	Sucrose	-99.75	-99.78	Azithromycin	-98.95	-99.71
Atenolol[d]	-37.84	24.15	Sulphasalazine	-78.18	-62.09	Penicillin	-53.55	-79.54
Chlorothiazide	-97.80	-97.84	Terbutaline	-44.05	-62.33	H21644	-98.99	-98.53
Cimetidine	-98.86	-99.40	Uracil	-95.51	-92.15	Sumatriptan	-89.70	35.39
Dexamethasone-β-D glucoside	-87.30	-93.30	Dexamethasone-β-D glucuronide	-95.42	-98.72	Cephalexin	-72.16	-58.30
Doxorrubicin	-86.89	-97.37	Amoxicillin	-93.22	-95.27	Gly-Pro[d]	-20.82	14.46
Erythromycin	-98.24	-99.80	Enalapril	-43.26	-76.43	Raffinose	-99.99	-100.00
Ganciclovir	-99.86	-99.36	Furosemide	-99.55	-99.81	Metolazone	-85.85	-96.61
Hydrochlorothiazide	-99.87	-99.73	Epinephrine	-66.56	-51.88	Lactulose	-99.79	-99.75
Mannitol	-84.31	-86.37	Sulpiride	-99.69	-99.02	Foscarnet	-96.85	-99.93
Methotrexate	-99.82	-99.64	Bosentan[c,d]	19.51	60.42	Ciprofloxacin	-28.70	-10.01
Methylscopolamine[c,d]	39.02	70.31	Proscillaridin[c]	25.99	-2.43	Amiloride	-98.43	-98.56
Nadolol[c,d]	43.21	47.49	Ceftriaxone	-99.93	-99.64	BVARAU	-95.40	-96.90
Olsalazine	-92.44	-95.29	Remikiren	-99.21	-94.30	Lisinopril	-96.65	-96.94
Pirenzepine[d]	-19.51	12.48	Saquinavir	-99.81	-99.81	SQ-29852	-95.19	-98.72

Table 4. *Cont.*

Compounds	ΔP%[a]	ΔP%[b]	Compounds	ΔP%[a]	ΔP%[b]	Compounds	ΔP%[a]	ΔP%[b]
Nelfinavir	-75.88	-44.80	DMP851	-44.82	-10.27	L-Glutamine	-52.97	-64.12
Ritonavir	-99.71	-99.35	Indinavir	-96.45	-94.01			
				Test set				
PNU200603	-83.12	-65.56	Artorvastatin	-50.53	-96.59	Rhodamine[c,d]	93.80	88.65
Cyclosporine	-100.00	-100.00	Losartan[c,d]	97.93	95.53	Vinblastine	-53.42	-11.39
Homosulfasalazine	-74.10	-54.88	Lucifer Yellow	-98.28	-98.86	CNV97100	-3.44	-0.79
5-Aminosalicylic Acid	-29.33	-44.93	Phenol Red[c,d]	99.86	99.63	Sulfasalamide	-90.17	-49.57

[a,b]ΔP% = [P(Active) - P(Inactive)]x100.
[a]Classification of each compounds using the obtained model with non-stochastic linear indices.
[b]Classification of each compounds using the obtained model with stochastic linear indices.
[c]Compounds classified incorrectly by Eq. 2
[d]Compounds classified incorrectly by Eq. 3

Receiver operating characteristic curve, ROC, for the training sets are shown in the Figure 1a,b. The areas under the curve (AUC) are 0.9575 and 0.9386 for Eq. **2** (Fig. 1a) and Eq. **3** (Fig. 1b), respectively. These high values of AUC are a measure of the excellent test accuracy of the discriminant functions selected.



(a) Eq. **2**, obtained by using atom-based non-stochastic linear indices (Matthews corr coeff, MCC = 0.81; AUC = 0.9575)

(b) Eq. **3**, obtained by using atom-based stochastic linear indices (Matthews corr coeff, MCC = 0.77; AUC = 0.9386)

Figure 1. Receiver operating characteristic curves (ROC) (a) for the Eq. **2** (atom-based non-stochastic linear indices) and (b) for the Eq. **3** (atom-based stochastic linear indices) in training sets (white points) and random classifier (black points). TPF = sensitivity and FPF = 1-specificity for different thresholds of class function (between -10 and +10).

Comparison with other approaches. The use of atom-based non stochastic and stochastic linear indices, for the classification of molecules with poorer-moderate and higher absorption, was compared with other method implemented in the Dragon software.[38] Five kinds of indices were specifically used: Topological, BCUT, Gálvez´s topological charge indices, 2D Autocorrelations and Molecular walk counts; all these families are made up for bidimentional molecular descriptors like linear indices. The development of these five models involved the use of the same data set that it was used in developing the model of linear indices and the same number of variables in the equation, in most of the approaches. The Dragon software was not able to calculate compounds Cyclosporine, Squinavir and Ritonavir, because they have more atoms than the limit that software can calculate, so they molecules were not considered in further analysis. The obtained models with their statistical parameters are shown in Table 5.

Table 5. **Statistical Parameters of the QSAR Models Obtained Using Different Molecular Descriptors to Predict the Caco-2 Permeability.**

Index		N	λ	D^2	F	p	Eq N°
2D autocorrelations	ATS1e, ATS8e, ATS4v, ATS7v, GATS3v, MATS6e, GATS7m, GATS2e	136	0.586	2.91	11.24	0.0001	4
BCUT	BEHm1, BEHm2, BELm1, BELm4, BELm7, BEHv1, BELv2	136	0.775	1.20	5.32	0.0001	5
Gálvez topological charge indices	GGI3, GGI4, GGI9, JGI1, JGI3, JGI6, JGI7, JGT	136	0.719	1.61	6.20	0.0001	6
Topological descriptors	SEigv, PW5, AAC, T(F..F), T(O..Cl), X4Av, D/Dr07, X0Av	136	0.547	3.41	13.16	0.0001	7
Molecular Walk Count	MWC09, MWC10, TWC, SRW03, SRW05, SRW07	136	0.938	0.27	1.42	0.21	8

The first comparison was based on the quality of the statistical parameters of the discriminant function, as well as predictive capacity of the models generated. As can be seen, the present approach (non stochastic and stochastic atom-based linear indices) achieved the best values, for the statistical parameters of the developed QSAR models. Moreover, our models showed the lowest values of the Wilks' lambda and the highest values of square Mahalanobis distance and Fisher ratio. All the models were significant, from statistical point of view, with the exception of the model obtained with molecular walk count descriptors. For the training set, the most accurate models, were those obtained with non-stochastic linear indices (Q = 90.56%), with atom-based quadratic indices (Q = 90.30%) and with stochastic atom-based linear indices (Q = 89.13), respectively. The best value of the sensitivity was obtained with quadratic indices (96.29%), but notice that our two models Eqs. **2** and **3** had a sensitivity of 90.24% and 91.46%, correspondingly. In addition, it should be noted that the model obtained with non-stochastic quadratic indices was built with a slightly shorter training set; besides, their test set (11 compounds) is shorter than our test set (19 compounds). The specificity of our models is greater than 90%, while the specificity of the rest of the approaches is between 70% and 88%. It is remarkable that both models showed a false alarm rate lesser than 14.5%; specifically, that the model obtained with non-stochastic linear indices has only 8.93% of false positive rate; and the other approaches show values between 18% and 38%.

Table 6. Comparison Between Atom-Based Linear Indices and Others Approaches Predicting Permeability across Caco-2 Cells.

Models' features to be compared	Non-stochastic Linear Indices (Eq. 2)	Stochastic Linear Indices (Eq. 3)	2D autocorr (Eq. 4)	BCUT (Eq. 5)	Galvez charge Indices (Eq. 6)	Topological (Eq. 7)	Walk count (Eq. 8)	Quadratic Indices (Eq. 12 in Ref. 33)
				Training set				
N	138	138	136	136	136	136	136	134
Wilks'λ (U-statistic)	0.435	0.436	0.586	0.775	0.719	0.547	0.938	0.480
F	20.94	20.86	11.24	5.32	6.20	13.16	1.42	16.88
D^2	5.31	5.29	2.91	1.20	1.61	3.41	0.27	4.52
p-level	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.21	0.0001
Accuracy (%)	90.58	89.13	80.88	78.68	78.68	78.68	60.29	90.30
MCC	0.81	0.77	0.61	0.57	0.56	0.57	0.20	0.80
Sensitivity (%)	90.24	91.46	80.49	78.05	79.27	78.05	59.76	96.29
Specificity (%)	93.67	90.36	86.84	85.33	84.42	85.33	70.00	88.64
False Positive Rate (%)	8.93	14.29	18.52	20.37	22.22	20.37	38.88	18.86
				Test set				
N	19	19	18	18	18	18	18	11
Predictability (%)	84.21	84.21	72.22	59.09	66.66	50.00	66.66	83.33
MCC	0.73	0.73	0.43	0.35	0.30	0.08	0.30	0.71

An important point of view to either accept or reject a QSAR model is the statistics for the external prediction set.[77] Both obtained models achieved the highest value of accuracy of 84.21% and MCC of 0.73. Among the other approaches only Marrero-Ponce's model obtained comparable results to our two models, the other ones showed poor results predicting permeability across Caco-2 cells. All these results are summarized in Table 6, where a comparison among different computational schemes can be more easily performed.

Virtual Screening. The relevance of QSAR studies in the prediction of human intestinal absorption has been demonstrated in recent publications,[26,89-92] and the so-called "*Rule-of-5*" has proved very popular as a rapid screen for compounds that are likely to be poorly absorbed.[18] In the present study, it was simulated a virtual search to predict the absorption profile of 241 compounds,[92] using discriminant obtained models with non-stochastic and stochastic linear indices. The aim of the present report is to evaluate the capacity of human absorption prediction from the classification models (Eqs. 2 and 3), into high and moderate-poor, for drug absorption in Caco-2 cells. Moreover, some compounds included in the obtained models (training or test set) were also used in this screening. As the compounds selected for the virtual screening were obtained from different sources, only the first 145 compounds (data of best quality, classified as OK and Good by Zhao *et al.*[92]) should be used to bring a better comparative criterion; nevertheless, the rest of the compounds can be evaluated, but their human absorption values (Abs %) were not comparatively reliable.[92] These experimental values and the evaluation results of these compounds are depicted in Table 7. In this Table we give the values of $\Delta P\% = [P(H) - P(M\text{-}P)]x100$, where P(H) is the probability that the equation classify a compound with $P_{Caco\text{-}2} \geq 8\ x10^{-6}$cm/s. Conversely, P(M-P) is the probability that the model classify a compound with $P_{Caco\text{-}2} < 8\ x10^{-6}$cm/s. This ΔP% takes positive values when P(H) > P(M-P), and negative, otherwise. Therefore, when ΔP% is positive (negative), the compounds are classified with Higher (Moderate-Poorer) absorption profile.

Table 7. Results for the Virtual Screening of 241 drugs. Permeability Coefficient from Models (Eq. 2 and Eq.3) and Observed Human Absorption and Bioavailability from Literature.

Compounds	ΔP%[a]	ΔP%[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
1-Cisapride	1.71	-48.69	100		100	100
2-Valproic acid	68.42	70.46	100	~100	90(68-100)	100
3-Salicylic acid	27.53	20.02	100	100		100
4-Diazepam	97.88	95.31	97-100	100		100
5-Sudoxicam	57.96	85.62		100		100
6-Glyburide	-99.00	-90.91				100
7-Gallopamil	93.44	98.87		~100	15	100
8-Mexiletine	98.77	99.16		100	88	100
9-Nefazodone	89.58	76.81		100	15-23	100
10-Naproxen	90.60	90.60	94-99	100	99	99
11-Lamotrigine	98.42	96.98	70		98	98
12-Tolmesoxide	97.49	99.59	100		85	98
13-Disulfiran	97.05	99.76		91		97
14-Torasemide	-98.14	-87.89			96	96
15-Metoprolol	80.29	75.39	95-100	>90	50	95
16-Naloxone	69.49	75.63				91
17-Terazocin	16.19	59.31	91	~100	90	90
18-Sulindac	85.72	44.96		90		90
19-Sultopride	71.91	95.43	100	~100		89
20-Topiramate	-99.25	-97.51			81-95	86
21-Tolbutamide	-91.88	1.98				85
22-Propiverine	97.14	98.31		84		84
23-Digoxin	-99.71	-99.98			67	81
24-Mercapto ethane sulfonic acid	-4.45	67.49				77
25-Cimetidine	-98.86	-99.40	62-98		60	64
26-Furosemide	-99.55	-99.81	61	61	61	61
27-Metformin	-99.61	-99.69			50-60	53
28-Rimiterol	-40.32	-50.43				48
29-Cymarin	-28.30	-89.58		47		47
30-Ascorbic Acid	-36.60	-67.71				35
31-Fosfomycin	-94.48	-99.10				31
32-Fosmidomycin	-97.47	-98.52		30		30
33-k-Strophanthoside	93.44	98.87		16		16
34-Adefovir	-99.99	-99.99	12		12	16

Table 7. *Cont.*

Compounds	ΔP%[a]	ΔP%[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
35-Acarbose	-100.00	-100.00		1-2.		2
36-Ouabain	97.40	91.81		1.4		1.4
37-Kanamycin	-99.87	-99.93				1
38-Lactulose	-99.79	-99.75	0.6	0.6		0.6
39-Camazepan	97.77	95.92	99		100	100
40-Indomethacin	96.19	76.12	100		100	100
41-Levomorgestrel	96.86	94.45			100	100
42-Tenoxicam	99.95	96.36			100	100
43-Theophyline	-12.33	7.42	96		100	100
44-Oxatomide	27.44	75.11	100			100
45-Desipramine	92.21	91.92	95-100	>95	40	100
46-Fenclofenac	98.52	85.26	100			100
47-Imipramine	98.61	99.15	95-100	>95	22-67	100
48-Lormetazepan	99.61	97.11	100	100	80	100
49-Diclofenac	89.53	17.53	100		90	100
50-Granisetron	78.87	87.07	100	100		100
51-Testosterone	94.73	94.28	100	100		100
52-Caffeine	66.46	79.75	100	100		100
53-Corticosterone	62.43	81.91	100	100		100
54-Ethinyl stradiol	95.41	88.93	100	~100	59	100
55-Isoxicam	80.25	88.32		100		100
56-Lornoxicam	99.98	95.48		100		100
57-Nicotine	95.81	98.16	100	100		100
58-Ondansetron	97.37	95.57	100	100	60	100
59-Piroxicam	86.35	93.63	100	100		100
60-Verapamil	95.27	98.93	100	>90	10-52.	100
61-Progesterone	98.29	97.71	91-100	91		100
62-Stavudine	-91.55	-67.62			100	100
63-Toremifene	99.88	99.07			100	100
64-Cyproterone acet.	99.53	91.61			100	100
65-Praziquantel	92.86	95.28		100		100
66-Cicaprost	-32.46	-39.24		100		100
67-Aminopyrine	95.07	97.28	100			100
68-Nordazepam	86.99	62.02	99		99	99
69-Carfecillin	-96.98	-98.32	100			99
70-Prednisolone	45.41	54.45	99		70-100	99
71-Propranolol	92.04	89.46	90-100	>90	30	99
72-Viloxazine	70.01	80.48	100	~100	61-98	98
73-Warfarin	95.84	79.31	98	~100	93-98	98
74-Atropine	70.08	73.67		90		98
75-Minoxidil	90.67	90.96		95		98
76-Clofibrate	98.54	95.19	96		95-99	97
77-Trimethoprim	43.54	91.44	97		92-102	97
78-Venlafaxine	78.46	94.00	92			97
79-Antipyrine	97.19	97.19	100	~100	97	97
80-Bumetanide	-99.47	-99.56	100	100	~100	96
81-Trapidil	83.17	82.70			96	96
82-Fluconazole	42.81	49.35	95-100		>90	95
83-Sotalol	-81.09	6.26	95	~100	90-100	95
84-Codeine	96.03	97.07	95		91	95
85-Flumazenil	92.66	83.10	95	>95	16	95
86-Ibuprofen	86.28	91.16	100			95
87-Labetalol	-88.97	-72.53	90-95	>90	33	95
88-Oxprenolol	73.08	82.52	97	90	50	95
89-Practolol	-38.01	-46.41	95	~100		95
90-Timolol	24.14	50.00	72	>90	75	95
91-Alprenolol	88.48	89.00	93-96	>93		93
92-Amrinone	35.47	71.14		93		93

Table 7. *Cont.*

Compounds	ΔP%[a]	ΔP%[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
93-Ketoprofen	93.22	90.71	100	~100	>92	92
94-Hydrocortisone	32.07	46.07	89-95	84-95		91
95-Betaxolol	70.90	82.52	90	90	80-89	90
96-Ketorolac	75.46	81.18	100	Well	80-100	90
97-Meloxicam	60.78	86.68	90		90	90
98-Phenytoin	-74.77	-21.01	90	90	90	90
99-Amphetamine	98.25	98.47				90
100-Chloramphenicol	63.57	43.89	90		80	90
101-Felbamate	-98.17	-68.87		90-95	102	90
102-Nizatidine	-19.04	-9.15	99		>90	90
103-Alprazolam	99.67	97.44			80-100	90
104-Tramadol	87.90	94.04			65-75	90
105-Nisoldipine	85.46	78.89				89
106-Oxazepam	56.56	44.22	97	~100	92.8	89
107-Tenidap	81.93	75.77	90		89	89
108-Dihydrocodeine	96.16	97.28			20	88
109-Felodipine	97.91	67.86	100	100	16	88
110-Nitrendipine	87.83	94.18			23	88
111-Saccharin	22.94	69.27	97	88		88
112-Mononidine	-36.01	-72.04			88	87
113-Bupropion	94.12	82.27	87		87	87
114-Pindolol	-0.22	0.14	92-95	>90	87	87
115-Lamivudine	-96.24	-99.19			86-88	85
116-Morphine	90.10	90.27	100	~100	20-30	85
117-Lansoprazole	98.55	91.70			85	85
118-Oxyfedrine	71.57	60.82			85	84
119-Captopril	-7.97	-64.63	77	71	62	84
120-Bromazepam	94.57	76.80	84		84	84
121-Acetylsalicylic acid	58.10	37.92				82
122-Sorivudine	-95.40	-96.90	82	82	61	82
123-Methylprednisolone	40.73	60.55	82		82	82
124-Mifobate	95.89	100.00				81
125-Flecainide	-65.46	-97.33			81	81
126-Quinidine	93.74	93.01	80	81	81	81
127-Piroximone	-66.76	-52.08			81	80
128-Acebutolol	-58.58	-66.29	90	90	50	80
129-Ethambutol	-71.62	-82.84		75-80		80
130-Acetaminophen	24.49	3.24	80-100	80	68.95	80
131-Dexamethasone	55.54	62.89	92-100		80	80
132-Guanabenz	72.25	17.48	75			80
133-Isoniazid	89.83	98.86				80
134-Omeprazole	76.84	93.80				80
135-Methadone	98.97	98.94			80	80
136-Fanciclovir	-5.58	-63.92			77	77
137-Metolazone	-85.85	-96.61	64	62-64		64
138-Fenoterol	-69.60	-84.27		60		60
139-Nadolol	43.21	47.49	20-35	34	34	57
140-Atenolol	-37.84	24.15	50-54	50	50	50
141-Sulpiride	-99.69	-99.02	36		30	44
142-Metaproterenol	-44.05	-62.33		44	10	44
143-Famotidine	-99.97	-99.89			37-45	28
144-Foscarnet	-96.85	-99.93	17	17(12-22)		17
145-Cidofovir	-99.96	-99.98			<5	3

Table 7. *Cont.*

Compounds	ΔP%[a]	ΔP%[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
146-Isradipine	80.54	77.44	92	90-95	17	92
147-Terbutaline	-44.05	-62.33	60-73	50-73	16	62
148-Reproterol	-97.48	-96.90		60		60
149-Lincomycin	-95.16	-96.49		20-35		28
150-Streptomycin	-100.00	-100.00		poor		1
151-Fluvastatin	59.75	-66.03	100	>90	19-29	100
152-Urapidil	-57.35	-49.78			68	78
153-Propylthiouracil	-28.43	-81.21	75		76(53-88)	76
154-Recainam	-95.52	-89.52				71
155-Cycloserine	94.66	85.89				73
156Hydrochlorothiazide	-99.87	-99.73	67-90	65-72		69(65-72)
157-Pirbuterol	-69.50	-68.40				60
158-Sumatriptan	-89.70	35.39	55-75	>57	14	57
159-Amiloride	-98.43	-98.56				50
160-Mannitol	-84.31	-86.37	16-26			16
161-Ganciclovir	-99.86	-99.36	3-3.8	3	3	3
162-Neomycin	-99.55	-99.96				1
163-Raffinose	-99.99	-100.00	0.3			0.3
164-Phenglutarimide	-5.47	37.10	100			100
165-Bornaprine	95.59	96.85	100			100
166-D-Phe-L-Pro	44.38	61.59	100			100
167-Scopolamine	59.92	70.51	90-100			95
168-Naloxone	78.77	74.17	91			91
169-Ziprasidone	85.00	84.55	60			60
170-Guanoxan	-78.60	-71.22		50		50
171-Netivudine	-97.87	-85.79		28		28
172-Gentamicin-C1	-99.85	-99.88	0	poor		poor
		Zwitterionic drugs				
173-Cefadroxil	-91.02	-87.96			100	100
174-Ofloxacin	26.38	58.82			100	100
175-Pefloxacin	94.87	93.89			100	100
176-Cephalexin	-72.16	-58.30	98	100		100
177-Loracarbef	-37.30	-43.66	100	100		100
178-Glycine	76.98	63.03	100			100
179-Amoxicillin	-93.22	-95.27	94		93	93
180-Tiagabine	93.84	79.38			90	90
181-Telmisartan	97.80	92.56	90	rapid	43	90
182-Trovafloxacin	58.67	25.80	88		88	88
183-Acrivastine	89.84	88.38	88			88
184-Nicotinic acid	63.46	72.51				88
185-Levodopa	5.57	20.91	100	80-90	86	86
186-Cefatrizine	-96.95	-97.04			75	75
187-Ampicillin	-76.40	-79.33				62
188-Vigabatrin	44.94	5.99				58
189-Tranexamic acid	42.19	42.29	55			55
190-Eflurnithine	85.36	-1.65				55
191-Metyldopa	50.46	35.55		41		41
192-Ceftriaxone	-99.93	-99.64	1	1		1
193-Distigminebromide	26.90	84.20			47	8
194-Ziduvudine	-97.99	-98.26	100	100	63	100
195-Ximoprofen	67.67	87.40	100		98	98
196-Clonidine	48.27	-6.76	85-100	100	75-95	95

Table 7. *Cont.*

Compounds	ΔP%[a]	ΔP%[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
197-Viomycin	-100.00	-100.00				85
198-Ceftizoxime	-99.94	-99.82				72
199-Capreomycin	-100.00	-100.00				50
200-AAFC	-78.97	-91.81		32		32
201-Bretylium tosylate	99.02	98.80	23		23	23
Dose-limited, dose-dependent, and formulation-dependent drugs						
202-Spironolactone	96.96	79.13		>73		73
203-Etoposide	-19.08	-63.18	50		50(25-75)	50(25-75)
204-Cefetamet pivoxil	-99.26	-93.34			47	47
205-Cefuroximeaxetil	-98.81	-98.12	36		36-58	44(36-52)
206-Azithromycin	-96.98	-99.68	35-37		37	37
207-Fosinopril	-22.65	96.68		36	25-29	36
208-Pravastatin	-88.12	-98.16	34	34	18	34
209-Cyclosporin	-100.00	-100.00	35		10-60	28(10-65)
210-Bromocriptine	-97.33	-98.66	28	28	6	28
211-Doxorubicin	-86.89	-97.37	5	trace	5	12(0.7-23)
212-Cefuroxime	-99.81	-98.97				1
213-Iothalamate sodium	49.07	36.30	1.9	1.9		1.9
214-Sulphasalazine	-78.18	-62.09	12-13.			59(56-61)
215-Benazepril	27.02	-34.75	37	>37		≥37
216-Lisinopril	-96.65	-96.94	25	25	25-50	28(25-50)
217-Esalaprilat	-82.78	-94.33	9-10.	10-40.		25(10-40)
218-Anfotericina	-100.00	-100.00	5	poor		3(2-5)
219-Aztreonam	-99.98	-99.67		<1	<1	1
220-Mibefradil	50.10	51.53			37-109	69(37-100)
221-Ranitidine	-24.88	-39.19	50-61		50(39-88)	64(39-88)
222-Chlorotiazide	-97.80	-97.84	13-56			49(36-61)
223-Acyclovir	-99.28	-96.52	20-30		15-30	23(15-30)
224-Norfloxacin	-35.96	-13.19	35	30-40	~70	71
225-Metthotrexate	-99.82	-99.64	20-100	100		70(53-83)
226-Gabapentin	56.71	26.28	50	well	60A(36-64)	59(43-64)
227-Prazocin	44.15	77.59	100		44-69	86(77-95)
228-Olsalazine	-92.44	-95.29	2.3		2.3	24(17-31)
Drugs expected to have higher absorption						
229-Ciprofloxacin	-28.70	-10.01	69-100		69	≥69
230-Ribavirin	-96.57	-88.73			33	≥33
231-Pafenolol	-93.22	-95.17			28	≥29
232-Azosemide	-98.86	-98.16			10	≤10
233-Xamoterol	-98.15	-98.96			5	≥5
234-Enalapril	-27.13	-79.04	66	60-70	29-50	66(61-71)
235-Phenoxymethyl penicillin	-81.94	-89.49	45	45(31-60)		59(49-68)
236-Gliclazide	-76.10	45.86				≥65
237-Benzylpenicillin	-53.55	-79.54	30	15-30		≥30
238-Thiacetazone	-94.33	-56.28				≥20
239-Lovastatin	80.56	24.63	30	31		≥10
240-Cromolyn sodium	-52.42	-70.99				≥0.4
241-Erythromycin	-99.15	-99.82	35		35	≥35

ΔP% = [P(High absorption group) - P(moderate-poor group)]x100. [a]Results of the classification of compounds obtained from Eq.2. [b]Results of the classification of compounds obtained from Eq.3. [c]The data used for QSAR studies was taken from Clark,[78] Wessel,[26] Palm,[79] Yazdanian,[6] Yee,[13] and Chiou.[80] [d]Absorption data obtained from the original and reviewed literature. [e]Bioavaililability or absolute bioavailability of oral administration. [f]Absorption data (or average values) chosen in Reference [81] based on the analysis of literature.

Taking into account the first 145 compounds (data of best quality), the global percentage of good extrapolation (from "*in vitro*" to "*in vivo*") achieved by our models was 79.31% and 81.38%, correspondingly, when non-stochastic linear indices and stochastic linear indices were used. These values are similar to that value that obtained by Marrero-Ponce *et al.*[33] for data of best quality. If the rest of the experimental data (compounds 146-241) are considered as an only great group, where less realistic data of Abs% are reported, the percentages of correct correspondence between "*in vitro*" permeability data (Caco-2 cells, predicted by linear indices) and the human absorption were 77.89% (74/95) with non-stochastic and 75.79%(72/95) with stochastic linear indices. This group has a lesser percentage of correspondence than the first 145 previously analyzed compounds.

Now we are going to analyze the last group (146-241) but divided into several groups according to the Zhao *et al.* classification.[92] Compounds from 146 to 172 were considered as uncertain and unchecked data. For these compounds, the global good classifications were 80.77% for non-stochastic and 76.92% for stochastic linear indices. In addition, the group of twenty zwitterionic drugs, reported by Zhao *et al.*,[92] was analyzed. For this kind of drugs, our model showed only a 50% of correct correspondence between "*in vitro*" classification and "*in vivo*" results for both models. The same result was obtained with non-stochastic quadratic indices by Marrero-Ponce *et al.*[33] The prediction of compounds 193-201 (group of missing fragments) was lesser than 50%. It means that more than a half of the compounds were badly predicted; this is a logical result if we bear in mind those the criteria followed to classify these compounds by Zhao *et al.*[92]. For the group with dose limited, dose-dependent and formulation-dependent drugs the correspondence between "*in vitro*" permeabilities and the human absorption values was high. Only one compound was misclassified by Eq. **2** for an accuracy of 96.15% (26/27), and two compounds were badly classified by Eq. **3**, for 92.31% (25/27) of global good classification. Finally, for the analysis of drugs with expected higher absorption, according to Zhao *et al*,[92] it is not reported a value or average for the human absorption. However, if the data from the fifth, sixth and seventh columns in Table 7 are considered, our two models explain 91.67% of the experimental variance.

Considering the full set (241 compounds) the model obtained with the non-stochastic linear indices (Eq. **2**) showed a 78.84% (190/241) of the explanation of the human absorption values, which is a logic result by considering the structural variability and the biological property. Notice that these values are the same as those obtained by Marrero-Ponce *et al.*[33] On the other hand, the results obtained for the full set with stochastic linear indices (Eq. **3**) were slightly better; this model has an accuracy of 79.25% (191/241).

Nevertheless, it has been widely reported in the literature the influence of transport mechanism on the prediction of this biopharmaceutical property, for example: Methotrexate is absorbed by a carrier-mediated process, Zidovudine is absorbed by active transport, Amoxicillin and Cefatrizine are absorbed via dipeptide carrier system, as well as in the Etoposide case, it is suggested that its distribution into the brain is partially controlled by an active transport process.[89] In addition, Cefadroxil, Digoxin and Cepahalexin were compounds with known active transport.[93] Other compounds with the same skeleton pattern; i.e., cephalosporins (Cefatrizine and Ceftizoxime), cardiotonic glycosides (Ouabain) as well as antiviral nucleoside analogues (Stavudine, Lamivudine, Sorivudine) appear badly classified (uncorrelated between the permeability predicted in Caco-2 cells and the human absorption values), suggesting an active transport system for these drugs. In addition, in the case of the Viomycin, with an appropriate intestinal absorption (Abs % = 85), it has a molecular weight value of 685 $gMol^{-1}$ (>500), similar to those drugs with poor intestinal absorption, for what it could be suggested that this compound can be actively transported, as it was pointed out in the case of Rifampicin by Egan *et al*.[94]

According to these results, we can say that the quality of the predictions assessed the predictive power of the obtained QSAR models and justified their use in the prediction of this important biopharmaceutical property. Besides, this is not a fortuitous result, due to the data set used in this study including any sort of absorption model compounds.

Multiple Linear Regression Models.

QSAR Models Obtained Using *TOMOCOMD-CARDD* Descriptors. The purpose of this study was to develop a quantitative model that permit the prediction of Caco-2 cell permeability from the molecular structure by using a combinatorial approach of linear indices and multiple linear regression method. As we previously pointed out, the data set was divided into training and test set as in Hou *et al*.[32] The experimental apparent permeability, and the results predicted by the multiple linear regression (MLR) models for the training set were listed in Table 8. The models obtained by using non-stochastic linear indices with their statistical parameters, are given bellow:

$$\begin{aligned}\mathbf{log}\ \boldsymbol{P_{eff}} = & -3.757(\pm0.14) -0.031(\pm0.007)\,{}^{M}\boldsymbol{f}_{1L}{}^{H}(x_E) +0.004(\pm0.001)\,{}^{M}\boldsymbol{f}_{3L}{}^{H}(x_E) \\ & +1.44\text{x}10^{-6}(\pm2.43\text{x}10^{-7})\,{}^{P}\boldsymbol{f}_{11L}{}^{H}(x_E) +1.97\text{x}10^{-9}(\pm4.29\text{x}10^{-10})\,{}^{V}\boldsymbol{f}_{14L}(x_E) \\ & -0.040(\pm0.005)\,{}^{V}\boldsymbol{f}_{0L}{}^{H}(x_{E\text{-}H}) \qquad (\mathbf{9})\end{aligned}$$

N = 77 R = 0.82 R^2 = 0.672 q^2 = 0.60 F(5, 71) = 28.42 s = 0.47 p<0.0001

$$\log P_{eff} = -3.603(\pm0.136) - 0.039(\pm0.007)\,{}^{M}\boldsymbol{f}_{1L}{}^{H}(x_E) + 0.005(\pm0.001)\,{}^{M}\boldsymbol{f}_{3L}{}^{H}(x_E) + 1.59\mathrm{x}10^{-6}(\pm2.27\mathrm{x}10^{-7})\,{}^{P}\boldsymbol{f}_{11L}{}^{H}(x_E) + 2.15\mathrm{x}10^{-9}(\pm3.98\mathrm{x}10^{-10})\,{}^{V}\boldsymbol{f}_{14L}(x_E) - 0.040(\pm0.004)\,{}^{V}\boldsymbol{f}_{0L}{}^{H}(x_{E\text{-}H}) \quad (\mathbf{10})$$

$N = 75 \quad R = 0.85 \quad R^2 = 0.723 \quad q^2 = 0.68 \quad F(5, 69) = 36.03 \quad s = 0.43 \quad p<0.0001$

where N is the size of the data set, R is the regression coefficient, R^2 is the determination coefficient, s is the standard deviation of the regression, F is the Fischer ratio and q^2 is the squared correlation coefficient of the cross-validation performed by the LOO procedure. These statistics indicate that these models are appropriate for the description of the chemicals studied here. In the development of the first quantitative model for the description of activities (Eq. **9**, N = 77), two compounds were detected as statistical outliers. Once rejected the statistical outliers, Eq. **10** (N = 75) was obtained with better statistical parameters.

Table 8. Predicted Values of Log P_{eff} for Compounds in the Training Set

Name	Log P_{eff}.[a]	Pred. Eq 9[b]	Pred. Eq 10[b]	Pred. Eq 11[b]	Pred. Eq 12[b]	Pred. Eq. 13[c]	Pred. Eq. 14[c]	Pred. Eq. 15[c]	Pred. Eq. 16[c]	Pred. Eq. 17[c]
Acebutolol	-5.83	-5.78	-5.94	-5.76	-5.72	-5.15	-5.04	-5.30	-5.51	-5.30
Acebutolol ester	-4.61	-5.68	outlier	-5.45	-5.38	-5.14	-4.92	-5.47	-4.55	-5.07
Acetylsalicylic	-5.06	-4.73	-4.75	-4.83	-4.78	-4.97	-4.98	-5.25	-5.35	-5.01
Acyclovir	-6.15	-5.68	-5.77	-5.65	-5.62	-6.00	-5.66	-4.86	-5.29	-5.15
Alprenolol	-4.62	-4.72	-4.70	-4.91	-4.89	-4.74	-5.04	-5.00	-4.94	-5.12
Alprenolol ester	-4.47	-4.59	-4.62	-4.60	-4.55	-4.81	-4.92	-5.16	-4.20	-4.89
Aminopyrine	-4.44	-4.35	-4.34	-4.49	-4.40	-4.26	-4.83	-5.19	-4.43	-5.22
Artemisin	-4.52	-4.60	-4.72	-5.06	-4.95	-5.05	-5.06	-4.99	-4.94	-5.00
Artesunate	-5.40	-4.72	-4.91	-5.14	-5.01	-5.54	-4.88	-5.34	-5.29	-5.18
Atenolol	-6.50	-5.68	-5.78	-5.61	-5.61	-5.25	-5.08	-5.45	-5.44	-5.15
etaxolol	-4.81	-5.00	-5.06	-5.03	-5.00	-5.19	-4.94	-5.01	-5.34	-4.84
Betaxolol ester	-4.52	-4.89	-5.00	-4.72	-4.66	-5.14	-4.94	-4.96	-5.03	-4.60
Bremazocine	-5.10	-4.94	-4.92	-4.81	-4.78	-4.92	-4.60	-5.01	-4.31	-4.67
Caffeine	-4.41	-4.31	-4.31	-4.31	-4.22	-4.39	-5.37	-5.36	-4.65	-5.15
Chloranphenicol	-4.69	-5.23	-5.35	-5.29	-5.24	-5.24	-5.03	-4.93	-5.45	-5.22
Chlorothiazide	-6.72	-6.33	-6.35	-6.16	-6.21	-6.57	-6.30	-5.73	-6.23	-5.25
Chlorpromazine	-4.70	-4.40	-4.33	-4.22	-4.15	-4.67	-4.92	-4.56	-4.35	-5.34
Cimetidine	-5.89	-5.71	-5.80	-5.55	-5.54	-5.41	-5.06	-5.24	-5.19	-5.09
Clonidine	-4.59	-4.50	-4.37	-4.93	-4.92	-4.31	-5.05	-4.63	-5.03	-5.08
Corticosterone	-4.47	-4.92	-4.92	-4.80	-4.77	-4.70	-4.73	-4.92	-4.57	-5.15
Desipramine	-4.64	-4.65	-4.61	-4.58	-4.56	-4.58	-5.10	-4.47	-4.87	-5.29
Dexamethasone	-4.75	-5.03	-4.98	-4.97	-4.94	-4.63	-4.82	-5.13	-4.89	-5.33
Dexamethasone-β-D Glucoside	-6.36	-5.97	-6.04	-5.64	-5.63	-5.17	-4.85	-5.69	-6.18	-5.23
Dexamethasone- β-D Glucuronide	-5.94	-5.89	-5.96	-5.46	-5.45	-5.22	-4.83	-5.78	-6.17	-5.29
Diazepam	-4.32	-4.50	-4.46	-4.39	-4.31	-4.31	-4.91	-4.17	-4.87	-5.32
Dopamine	-5.03	-5.14	-5.04	-5.31	-5.32	-4.66	-4.94	-5.00	-4.71	-4.95
Doxorrubicin	-6.80	-6.40	-6.52	-6.41	-6.38	-5.78	-5.06	-5.81	-6.31	-5.61
Erythromycin	-5.43	-6.43	outlier	-5.78	-5.71	-5.66	-5.32	-6.28	-5.71	-5.87
Estradiol	-4.77	-4.56	-4.46	-4.70	-4.69	-4.83	-4.71	-4.67	-4.78	-4.95
Felodipine	-4.64	-4.81	-4.81	-4.61	-4.55	-4.63	-4.82	-4.17	-4.52	-4.98

Table 8. *Cont.*

Name	Log P_{eff}.[a]	Pred. Eq 9[b]	Pred. Eq 10[b]	Pred. Eq 11[b]	Pred. Eq 12[b]	Pred. Eq. 13[c]	Pred. Eq. 14[c]	Pred. Eq. 15[c]	Pred. Eq. 16[c]	Pred. Eq. 17[c]
Ganciclovir	-5.99	-5.94	-6.04	-5.88	-5.86	-6.12	-5.68	-4.95	-5.48	-5.21
Griseofulvin	-4.44	-3.93	-3.83	-4.20	-4.10	-4.89	-5.02	-5.14	-4.87	-5.05
Hydrochlorothiazide	-6.06	-5.82	-5.79	-6.38	-6.47	-6.35	-6.33	-5.73	-6.15	-5.25
Hydrocortisone	-4.66	-5.03	-5.00	-5.01	-4.99	-4.74	-4.75	-4.90	-4.73	-5.21
Ibuprophen	-4.28	-4.34	-4.27	-4.46	-4.43	-4.66	-4.66	-4.81	-4.69	-5.07
Imipramine	-4.85	-4.48	-4.48	-4.18	-4.13	-4.78	-5.10	-4.65	-4.73	-5.32
Indomethacin	-4.69	-5.24	-5.31	-4.92	-4.84	-5.14	-5.04	-4.92	-4.81	-5.01
Labetalol	-5.03	-5.93	-5.97	-5.79	-5.81	-5.04	-5.38	-5.00	-5.16	-5.35
Mannitol	-6.21	-5.51	-5.50	-5.81	-5.86	-5.96	-5.84	-5.18	-5.98	-4.95
Meloxicam	-4.71	-4.66	-4.57	-4.71	-4.68	-5.64	-5.03	-5.43	-5.29	-4.98
Methanol	-3.88	-4.18	-4.08	-4.60	-4.59	-3.27	-3.47	-4.30	-4.26	-4.67
Methotrexate	-5.92	-6.41	-6.50	-6.67	-6.62	-5.79	-5.81	-5.21	-6.24	-5.67
Methylscopolamine	-6.16	-5.05	-5.16	-4.69	-5.05	-5.60	-5.35	-5.61	-4.77	-4.61
Metoprolol	-4.59	-5.04	-5.13	-5.08	outlier	-5.22	-5.01	-5.00	-5.37	-5.14
Nadolol	-5.41	-5.52	-5.58	-5.67	-5.68	-5.18	-5.08	-5.63	-5.13	-5.33
Naproxen	-4.13	-4.62	-4.61	-4.63	-4.59	-4.87	-4.51	-4.66	-4.76	-5.18
Nevirapine	-4.52	-4.99	-4.94	-5.08	-5.00	-4.36	-5.49	-4.49	-5.01	-4.49
Nicotine	-4.71	-4.59	-4.60	-4.52	-4.46	-4.57	-4.93	-4.62	-4.47	-5.02
Olsalazine	-6.96	-6.25	-6.37	-6.28	-6.22	-6.47	-5.87	-5.75	-6.54	-5.32
Oxprenolol	-4.68	-4.80	-4.79	-4.96	-4.93	-4.85	-5.23	-4.94	-5.13	-5.14
Oxprenolol ester	-4.51	-4.70	-4.74	-4.65	-4.60	-4.90	-5.11	-5.12	-4.51	-4.91
Phencyclidine	-4.61	-4.10	-4.03	-4.12	-4.07	-4.52	-4.98	-4.38	-4.60	-5.25
Phenitoin	-4.57	-4.96	-4.95	-5.04	-5.03	-4.22	-4.88	-4.97	-4.85	-5.31
Pindolol	-4.78	-5.35	-5.37	-5.45	-5.44	-5.01	-4.86	-5.08	-4.76	-5.18
Pirenzepine	-6.36	-5.40	-5.55	-5.07	outlier	-5.13	-5.17	-4.98	-5.06	-5.53
Piroxicam	-4.45	-4.78	-4.69	-5.14	-5.09	-5.56	-5.71	-5.03	-5.50	-4.97
Practolol	-6.05	-5.48	-5.58	-5.58	-5.57	-5.26	-5.22	-5.45	-5.47	-5.15
Prazocin	-4.36	-5.21	-5.26	-5.09	-4.96	-5.14	-5.42	-4.78	-5.89	-5.12
Progesterone	-4.37	-4.28	-4.23	-4.17	-4.12	-4.71	-4.69	-4.85	-4.37	-5.08
Propranolol	-4.58	-4.73	-4.71	-4.93	-4.91	-4.68	-4.82	-5.05	-4.81	-5.21
Propranolol ester	-4.48	-4.62	-4.65	-4.62	-4.58	-4.77	-4.81	-5.32	-4.66	-4.97
Quinidine	-4.69	-5.24	-5.32	-4.83	-4.76	-5.10	-5.48	-4.79	-4.64	-5.02
Ranitidine	-6.31	-5.69	-6.01	-6.02	-5.92	-5.56	-5.40	-6.03	-5.79	-5.20
Salicylic acid	-4.79	-4.72	-4.65	-4.96	-4.94	-4.52	-4.55	-5.31	-5.35	-4.93
Scopolamine	-4.93	-4.84	-4.91	-4.70	-4.63	-5.48	-5.04	-5.22	-4.72	-4.55
Sucrose	-5.77	-6.07	-6.15	-6.05	-6.10	-6.30	-5.64	-5.07	-5.54	-4.95
Sulphasalazine	-6.71	-6.74	-6.87	-6.54	-6.48	-6.11	-5.83	-5.29	-6.44	-5.53
Telmisartan	-4.82	-5.44	-5.47	-4.70	-4.61	-4.73	-4.96	-5.37	-5.45	-5.55
Terbutaline	-6.38	-5.46	-5.44	-5.68	-5.70	-5.69	-5.18	-5.60	-5.12	-5.11
Testosterone	-4.34	-4.42	-4.35	-4.46	-4.43	-4.81	-4.72	-4.77	-4.28	-5.01
Timolol	-4.85	-4.69	-4.86	-4.72	-4.67	-5.21	-5.66	-5.51	-5.15	-5.27
Timolol ester	-4.60	-4.59	-4.80	-4.48	-4.40	-5.23	-5.32	-5.39	-4.69	-5.04
Uracil	-5.37	-4.98	-5.01	-5.16	-5.16	-5.39	-5.29	-5.62	-5.71	-4.86
Urea	-5.34	-5.17	-5.17	-5.16	-5.18	-5.89	-5.57	-6.56	-5.50	-4.72
Warfarin	-4.55	-4.80	-4.78	-4.66	-4.61	-4.73	-5.25	-4.66	-4.71	-5.41
Ziduvudine	-5.16	-5.35	-5.29	-5.98	-5.93	-5.85	-5.65	-4.88	-5.13	-5.25
PNU200603	-6.85	-6.37	-6.42	-5.99	-5.98	-5.73	-5.18	-5.13	-5.27	-5.27

[a]Experimental log P_{eff} obtained from Hou *et al.*[32]
[b]Predicted value of Log P_{eff} using non-stochastic and stochastic linear indices described in this report.
[c]Predicted value of Log P_{eff} using Dragon descriptors described in this report.

The achieved values of q^2 (0.60 and 0.68) showed that Eqs. **9** and **10** are reliable. Undoubtedly, the high value of LOO q^2 appears to be necessary, but not sufficient condition for the models to have a good predictive power. Golbraikh and Tropsha ever emphasized that the real predictive ability of a QSAR model could only be estimated by using an external test set of compounds that were not used for building the model.[77] Here, the actual prediction power of Equations **9** and **10** were validated by using an external test set of 23 compounds. The predicted values for the compounds of the test set, using non-stochastic linear indices are shown in Table 9. In this sense, bosentan was detected as a statistical outlier for both models; ceftriaxone and nitrendipine were also detected as outliers for Eqs. **9** and **10**, correspondingly; after that, the squared correlation coefficient for the test set for both models were 0.60 and 0.50, correspondingly. In the test set the predicted error for the other tested molecules is smaller than 1.0 log unit. The good prediction for the tested compounds confirms the significance of the selected molecular descriptors and the model based on them.

When the stochastic linear indices were used, the obtained models for the prediction of Caco-2 cell permeability are the following:

$$\mathbf{log}\ \boldsymbol{P_{eff}} = -4.298(\pm 0.175) + 9.18\ \mathrm{x}10^{-4}(\pm 4.18\mathrm{x}10^{-4})^{Ms}\boldsymbol{f}_1{}^{H}(x) + 0.063(\pm 0.015)^{Ms}\boldsymbol{f}_{8L}(x_E)$$
$$-0.043(\pm 0.006)^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H}) - 1.70(\pm 0.22)^{Ks}\boldsymbol{f}_{1L}{}^{H}(x_E) + 0.786(\pm 0.188)^{Gs}\boldsymbol{f}_{3L}{}^{H}(x_E) \quad (\mathbf{11})$$

N = 77 R = 0.79 $R^2 = 0.624$ $q^2 = 0.56$ $F(5, 71) = 23.91$ $s = 0.50$ $p<0.0001$

$$\mathbf{log}\ \boldsymbol{P_{eff}} = -4.271(\pm 0.155) + 9.088\ \mathrm{x}10^{-4}(\pm 4.18\mathrm{x}10^{-4})^{Ms}\boldsymbol{f}_1{}^{H}(x) + 0.059(\pm 0.014)^{Ms}\boldsymbol{f}_{8L}(x_E)$$
$$-0.047(\pm 0.006)^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H}) - 1.03(\pm 0.196)^{Ks}\boldsymbol{f}_{1L}{}^{H}(x_E) + 0.774(\pm 0.167)^{Gs}\boldsymbol{f}_{3L}{}^{H}(x_E) \quad (\mathbf{12})$$

N = 75 R = 0.84 $R^2 = 0.706$ $q^2 = 0.63$ $F(5, 69) = 31.70$ $s = 0.44$ $p< 0.0001$

where N is the size of the data set, R is the regression coefficient, R^2 is the determination coefficient, *s* is the standard deviation of the regression, F is the Fischer ratio, and q^2 is the squared correlation coefficient of the cross-validation performed by the LOO procedure. These statistics indicate that the models are appropriate for the description of those chemicals studied here. In the development of the stochastic model for the description of the activities (Eq. **11**, N = 77), two compounds were detected as statistical outliers. Once rejected the statistical outliers, the Eq. **12** (N = 75) was obtained with better statistical parameters. The results predicted by the stochastic MLR models for the training set were listed in Table 8.

As we pointed out, the predictive power of a QSAR model must be estimated using an external test set. At this point, the real predictive power of stochastic linear indices' models (Eqs. **9** and **10**) were validated by the same external test set of 23 compounds. Two compounds were detected as statistical outliers by Eqs. **11** and **12**; they were ceftriaxone and guanabenz. The final squared correlation coefficients for the two models were 0.64 and 0.62, for the external prediction set. The obtained values for the prediction of the test set, using stochastic linear indices are also shown in Table 9.

Table 9. Predicted Values of Log P_{eff} for Compounds in the Test Set

Name	Log P_{eff}.[a]	Pred. Eq 9[b]	Pred. Eq 10[b]	Pred. Eq 11[b]	Pred. Eq 12[b]	Pred. Hou´s[c]	Pred. G and Y[d]	Pred. Eq. 13[e]	Pred. Eq. 14[e]	Pred. Eq. 15[e]	Pred. Eq. 16[e]	Pred. Eq. 17[e]
Amoxicillin	-6.10	-5.78	-5.81	-5.38	-5.38	-6.16	-5.65	-5.42	-5.26	-5.977	outlier	-5.064
Antipyrine	-4.55	-4.55	-4.55	-4.64	-4.56	-4.82	-4.55	-4.53	-4.92	-5.062	-4.540	-5.106
Cyclosporine	-6.05	-6.42	-6.91	-5.48	-5.35	-5.81	-5.82	np	np	np	np	np
Diltiazen	-4.38	-4.86	-5.01	-4.26	-4.16	-4.84	-4.57	-5.14	-5.10	-4.355	-4.952	-5.093
Enalapril	-5.64	-4.68	-4.75	-4.69	-4.65	-5.66	-5.04	-4.89	-4.81	-4.917	-4.474	-5.440
Furosemide	-6.51	-5.86	-5.87	-6.09	-6.12	-5.81	-5.81	-6.23	-5.27	-5.459	-5.478	-5.316
Guanabenz	-4.50	-5.48	-5.49	outlier	outlier	-4.63	-5.24	outlier	-5.03	-4.491	-5.443	-5.034
Cumarin	-4.11	-4.25	-4.18	-4.37	-4.31	-5.31	-4.44	-3.83	-5.02	-4.924	-4.804	-4.997
Theophylline	-4.35	-4.58	-4.57	-4.66	-4.60	-5.06	-4.66	-4.57	-5.31	-5.106	-5.462	-5.105
Epinephrine	-6.02	-5.30	-5.25	-5.55	-5.57	-5.47	-5.56	-5.13	-5.15	-4.772	outlier	-5.015
Guanoxan	-4.71	-5.44	-5.47	-5.61	-5.61	-5.39	-5.27	-5.56	-5.48	-4.779	-4.960	-5.100
Lidocaine	-4.21	-4.59	-4.59	-4.67	-4.63	-4.45	-4.75	-5.03	-5.05	-4.903	-4.689	-5.121
Tiacrilast	-4.9	-5.43	-5.53	-4.83	-4.76	-5.68	-4.97	-5.34	-4.94	-4.938	-5.679	-5.208
Sulpiride	-6.16	-6.11	-6.27	-5.99	-5.99	-5.81	-5.47	-5.86	-5.21	-5.443	-5.150	-5.365
Nitrendipine	-4.77	-5.70	outlier	-5.88	-5.74	-5.08	-4.73	-5.13	-4.96	-4.598	-4.856	-5.472
Fleroxacin	-4.81	-4.94	-5.04	-4.75	-4.64	-5.39	-4.76	-5.19	-5.36	-4.563	-5.921	-5.056
Verapamil	-4.58	-5.16	-5.30	-4.27	-4.15	-4.87	-4.52	-4.75	-5.21	-4.884	-4.490	-5.640
Bosentan	-5.98	outlier	outlier	-5.59	-5.49	-5.32	-5.43	-5.41	-5.56	-4.778	-5.199	-5.434
Proscillaridin	-6.20	-5.76	-5.87	-5.17	-5.15	-5.42	-5.77	outlier	outlier	-5.042	-6.448	-5.652
Ceftriaxone	-6.88	outlier	-5.95	outlier	outlier	-6.53	-5.32	-5.82	-5.614	-5.928	-5.964	-5.390
Remikiren	-6.13	-6.73	-6.90	-5.75	-5.76	-5.36	-6.17	-5.18	-5.096	-5.295	outlier	-5.136
Saquinavir	-6.26	-6.76	-6.99	-6.29	-6.27	-5.39	-5.87	np	np	np	np	np
Mibefradil	-4.87	-5.44	-5.60	-4.67	-4.58	-5.06	-4.97	-4.84	-4.92	-5.16	-5.26	-5.35
R^2		0.61	0.50	0.64	0.62	0.61	0.58	0.55	0.19	0.31	0.32	0.11
UME		0.45	0.51	0.39	0.41	0.49	0.49	0.47	0.67	0.57	0.56	0.73

[a] Log P_{eff} experimentally obtained from Hou *et al.*[32]
[b]Predicted value of Log P_{eff} using non-stochastic and stochastic linear indices described in this report.
[c]Predicted value of Log P_{eff} using Hou's models.
[d]Predicted value of Log P_{eff} using Guangli and Yiyu's model.
[e]Predicted value of Log P_{eff} using Dragon descriptors described in this report.
UME unsigned mean error.

Comparison with other approaches. In two previous works,[3,32] the authors compare their results with the Caco-2 permeation model published by Marrero-Ponce *et al.*,[31] but the training data set used by Marrero-Ponce's group was small (only 17 compounds); thus a direct comparison is not correct. Therefore, we perform a comparison by using the same data set reported by Hou *et al*.[32], as well as used later by Guangli and Yiyu,[3] between our models and those models mentioned above. The comparison between the QSAR models reported here and other methods previously reported with the same data set is shown in Table 10.

The obtained models for Caco-2 permeability prediction, using atom-based non stochastic and stochastic linear indices were also compared with other method implemented in the Dragon software.[38] Five families of bidimensional molecular descriptors were used to develop quantitative models; they were: topological, BCUT, Gálvez´s topological charge indices, 2D Autocorrelations and Molecular walk counts. All of these models were performed with the same training data set. The results predicted by these multiple regression models for the training set were listed in Table 8

Table 10. Statistical Parameters of the QSAR Models Obtained Using Different Molecular Descriptors to Predict the Caco-2 Permeability.

index		N	R^2	s	F	q^2	Eq. Nº
Non-Stochastic Linear Indices	${}^{M}\boldsymbol{f}_{1L}{}^{H}(x_E)$, ${}^{M}\boldsymbol{f}_{3L}{}^{H}(x_E)$, ${}^{P}\boldsymbol{f}_{11L}{}^{H}(x_E)$, ${}^{V}\boldsymbol{f}_{14L}(x_E)$, ${}^{V}\boldsymbol{f}_{0L}{}^{H}(x_{E\text{-}H})$	77	0.67	0.469	28.43	0.60	9
Non-Stochastic Linear Indices	${}^{M}\boldsymbol{f}_{1L}{}^{H}(x_E)$, ${}^{M}\boldsymbol{f}_{3L}{}^{H}(x_E)$, ${}^{P}\boldsymbol{f}_{11L}{}^{H}(x_E)$, ${}^{V}\boldsymbol{f}_{14L}(x_E)$, ${}^{V}\boldsymbol{f}_{0L}{}^{H}(x_{E\text{-}H})$	75	0.72	0.432	36.03	0.68	10
Stochastic Linear Indices	${}^{Ms}\boldsymbol{f}_{1}{}^{H}(x)$, ${}^{Ms}\boldsymbol{f}_{8L}(x_E)$, ${}^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H})$, ${}^{Ks}\boldsymbol{f}_{1L}{}^{H}(x_E)$, ${}^{Gs}\boldsymbol{f}_{3L}{}^{H}(x_E)$	77	0.63	0.496	23.92	0.56	11
Stochastic Linear Indices	${}^{Ms}\boldsymbol{f}_{1}{}^{H}(x)$, ${}^{Ms}\boldsymbol{f}_{8L}(x_E)$, ${}^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H})$, ${}^{Ks}\boldsymbol{f}_{1L}{}^{H}(x_E)$, ${}^{Gs}\boldsymbol{f}_{3L}{}^{H}(x_E)$	75	0.70	0.441	31.70	0.63	12
Hou´s model	$-1.8 < \log D < 2.0$, *HCPSA*, *rgyr*, *f*rotb	77	0.68	0.439	38.10	0.62	Eq 17 in Ref. 32
Hou´s model	$-1.8 < \log D < 2.0$, *HCPSA*, *rgyr*, *f*rotb	74	0.71	0.405	43.1	0.66	Eq 18 in Ref. 32
Guangli and Yiyu´s model	*HBD*, *CPSA10*, *CPSA20*, *TPSA*	77	0.50	*	*	*	See Ref.3
2D autocorrelations	ATS7v, ATS1e, MATS1m, MATS5m, MATS3e	77	0.56	0.534	18.42	0.33	13
BCUT	BELv2, BEHv7, BELm1, BELv1, BEHv1	77	0.29	0.684	5.83	0.16	14
Gálvez topological charge indices	JGI1, JGI9, JGI6, GGI3, GGI2	77	0.32	0.668	6.79	0.17	15
Topological descriptors	SEigv, JhetZ, WA, Rww, IDDE	77	0.52	0.561	15.58	0.45	16
Molecular Walk Count	SRW03, TWC, MWC10	77	0.10	0.759	2.74	0.01	17

*Values are not reported in the literature.

The predictability of these five models were also validated using the external test set of 23 compounds as in Hou *et al*.[32] The predicted values for the tested molecules are also shown in Table 9. The Dragon software was not able to calculate compounds Cyclosporine and Squinavir, because they have more atoms than the limit that software can calculate, so these molecules were not performed in the test set.

The comparison was based on the quality of the statistical parameters of the regression. In this sense, the present approach (atom-based non stochastic linear indices) showed the highest squared correlation coefficient value of 0.72 with Eq. **10**, having as outliers only two compounds, while Hou's model achieved a value of 0.71 but had tree outliers. Nevertheless, Hou's model yields the lowest standard deviation value of 0.405, while non-stochastic model (Eq. **10**) shows a value of 0.432. The calculated values of q^2 for non-stochastic and stochastic linear indices' models are between 0.60 and 0.68; these values showed that our models are reliable. Certainly, the high value of LOO q^2 appears to be necessary but not sufficient condition for the models to have a good predictive power. Golbraikh *et al*. ever emphasized that the actual predictive ability of a QSAR model can only be estimated by using an external test set of compounds that were not used for building the model.[77] Some compound were detected as statistical outliers during the development of the validation of the QSAR models and removed for the subsequent analysis. As can be seen in Table 9, one of our models (Eq. **11**) shows the highest value (0.64) of the squared correlation coefficient for the test set and the lowest value of unsigned mean error (*UME*) of 0.39. Other models show adequate values of R^2 between 0.50 and 0.62 (see Table 9). In addition, only two models show values of *UME* lesser than 0.46 the previously mentioned for Eq. **11** (0.39), as well as the values of 0.45 and 0.41 achieved by the equations **9** and **12**, respectively. The values of *UME* for the rest of the approaches are between 0.47 and 0.73.

In a previous work one of the present authors (Y. M-P) performed a factor analysis to check the existence or not of linear independence between the total and local linear indices and others 0D-3D MDs available in the literature.[40] In these sense, here we carried out a factor analysis to ensure the linear independence between the atom-based non-stochastic and stochastic linear indices and the others 2D molecular descriptors calculated in this work. The comparison was based on 98 chemicals used to develop linear regression models because two compounds were not performed by Dragon software.

Table 11. Results of the Factor Analysis by Using the Principal Component Method for 23 2D Molecular Descriptors as Well as the 10 Non-stochastic and Stochastic Linear Indices

Factors	Eigenvalue	% Total Variance	Cumulative Eigenvalue	% Cumulative Variance
F_1	13.31	40.33	13.31	40.33
F_2	6.72	20.35	20.03	60.68
F_3	3.41	10.34	23.44	71.02
F_4	2.31	7.01	25.75	78.03
F_5	1.60	4.86	27.35	82.89
F_6	1.19	3.60	28.54	86.49

The results of the factor analysis are summarized in Table 11, and the six principal factors explain approximately 86.5% of the variance. The first factor explains 40.33% of the variance in the molecular indices studied. The addition of the second factor increases to 60.68% the variance explained, and the addition of the third factor allows 71.02% of the index variance to be accounted for. The other factors explain (% Cumulative Variance) the 7.01 (78.03), 4.86 (82.89) and 3.60% (86.49) of the variance in the molecular descriptors studied (see Table 11). Factor loadings from the principal component analysis, after a Varimax normalized rotation of the factors, are shown in Table 12.

Table 12. Factor Loadings (Varimax Normalized Rotation) for 23 2D Molecular Descriptors as Well as the 10 Total and Local (Atom-type) Linear Indices

index	**F_1**	**F_2**	**F_3**	**F_4**	**F_5**	**F_6**
${}^{M}\boldsymbol{f}_{1L}{}^{H}(x_E)$	0.14	**0.93**	0.07	0.15	0.23	-0.02
${}^{M}\boldsymbol{f}_{3L}{}^{H}(x_E)$	0.11	**0.95**	0.09	0.09	0.17	-0.05
${}^{P}\boldsymbol{f}_{11L}{}^{H}(x_E)$	0.00	**0.94**	0.15	-0.03	0.00	-0.16
${}^{V}\boldsymbol{f}_{14L}(x_E)$	-0.02	**0.92**	0.14	-0.07	-0.04	-0.17
${}^{V}\boldsymbol{f}_{0L}{}^{H}(x_{E\text{-}H})$	0.13	0.09	-0.04	-0.01	**0.90**	-0.13
${}^{Ms}\boldsymbol{f}_{1}{}^{H}(x)$	**0.66**	0.12	0.32	0.58	0.20	0.08
${}^{Ms}\boldsymbol{f}_{8L}(x_E)$	0.35	**0.73**	0.04	0.12	0.53	0.10
${}^{Ms}\boldsymbol{f}_{2L}{}^{H}(x_{E\text{-}H})$	0.05	0.26	0.00	0.02	**0.86**	-0.19
${}^{Ks}\boldsymbol{f}_{1L}{}^{H}(x_E)$	0.38	**0.70**	0.06	0.13	0.54	0.13
${}^{Gs}\boldsymbol{f}_{3L}{}^{H}(x_E)$	0.40	**0.68**	0.06	0.12	0.55	0.15
SEigv	-0.61	-0.36	-0.07	-0.12	-0.61	-0.10
JhetZ	-0.54	0.28	-0.08	-0.47	0.22	-0.26
WA	0.36	0.19	0.40	**0.67**	0.30	0.21
Rww	0.22	0.24	0.16	**0.72**	0.41	0.19
IDDE	0.52	0.16	**0.65**	0.34	0.14	0.18
ATS7v	0.18	0.40	0.60	0.31	0.06	0.19
ATS1e	-0.12	0.53	-0.44	-0.39	0.50	-0.13
MATS1m	-0.13	0.61	0.08	-0.52	0.16	0.00
MATS5m	-0.04	0.03	**0.87**	-0.08	0.14	0.15
MATS3e	0.24	-0.17	0.21	0.11	0.01	**0.75**
BELv2	0.24	0.14	**0.84**	0.30	0.07	0.09
BEHv7	0.46	0.07	**0.65**	0.48	-0.03	0.20
BELm1	0.35	-0.17	**0.85**	0.08	-0.14	-0.03
BELv1	0.30	0.19	**0.87**	-0.04	-0.04	-0.15
BEHv1	0.22	0.12	**0.94**	-0.05	-0.09	-0.05
JGI1	0.31	0.08	0.06	-0.06	0.36	**-0.78**
JGI9	0.54	0.11	0.15	0.56	0.39	0.14
JGI6	**0.82**	0.07	0.35	0.16	0.09	0.03
GGI3	**0.92**	0.12	0.21	-0.03	0.18	-0.04
GGI2	**0.89**	0.07	0.26	0.06	0.23	-0.07
SRW03	-0.09	-0.06	-0.02	**0.67**	-0.15	-0.07
TWC	**0.82**	0.19	0.36	0.32	0.13	0.06
MWC10	**0.89**	0.05	0.12	-0.03	-0.03	-0.05
Expl.Var	6.86	6.42	6.04	3.51	4.01	1.71
Prp.Totl	0.21	0.19	0.18	0.11	0.12	0.05

The k^{th} total linear indices [${}^{Ms}\boldsymbol{f}_{1L}{}^{H}(x)$] is strongly loaded in factor 1 (F_1). Most of the Dragon 2D descriptor used in this study as Galvez topological charge indices (JGI6, GGI2 and GGI3) and molecular walk counts (TWC and MWC10) are also robustly loaded (loadings > 0.65) in this factor. Thus, total linear indices and the others "F_1-indices" produce much redundancy and overlapping among them and their relations are very complex. The second factor (F_2) is almost exclusively for the atom-type linear indices of heteroatoms [$\boldsymbol{f}_{kL}(x_E)$] as well as the five factor (F_5) appears to be most significant for the atom-type (H-atoms bonding to heteroatoms) linear indices [$\boldsymbol{f}_{kL}(x_{H\text{-}E})$] as can be seen in Table 12. As previously stated, the indices with a high loading in the same factor are interrelated, while no correlation exists between indices having nonzero loadings only in different factors.[73,79-85] Consequently, it is clear that the atom-type linear indices are orthogonal to most of the 2D molecular descriptors, Thus, we can say that the atom-type (heteroatoms and H-atoms bonding to heteroatoms) linear indices contain structural information not contained in any other 2D molecular descriptors used in this study. These results are quite similar to those achieved in an earlier work by Marrero-Ponce.[40]

Interpretation of QSPerR Models. Now, although this topic is not an objective of our work, we are going to give a brief and general explanation about physical interpretation in structural terms of the obtained models.

Hitherto, it is known that the absorption is influenced by a different kind of interactions. Several studies have demonstrated that the permeability coefficient, measured by a transport through Caco-2 monolayer cell cultures, is correlated with lipophilicity,[6,15-17] while others emphasize on the role of hydrogen-bonding capacity or charge.[7,10,15,16] A paradigm of structure-permeability relationship has been expressed as:[16]

Permeability = *f* (lipophilicity, molecular size, H-bonding capacity, charge) **(18)**

As it can be observed, in the discriminant and the regression models, the included variables are very close to the factors that influence on the P values. These factors are related with the structural features of molecules. The variables $\boldsymbol{f}_{k\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E\text{-}H}})$ and ${}^{s}\boldsymbol{f}_{k\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E\text{-}H}})$ are connected with the hydrogen atoms as donors, while the variables $\boldsymbol{f}_{k\mathrm{L}}(x_{\mathrm{E}})$, $\boldsymbol{f}_{k\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}})$, ${}^{s}\boldsymbol{f}_{k\mathrm{L}}(x_{\mathrm{E}})$ and ${}^{s}\boldsymbol{f}_{k\mathrm{L}}^{\mathrm{H}}(x_{\mathrm{E}})$ contain information about the number of hydrogen acceptors and the charge of molecules. All of them are related with the total hydrogen bond capacity. The values of P negatively depend on these descriptors. The values of these molecular indices increase with the rise of the numbers of heteroatoms and the hydrogen bond to heteroatoms in the molecules. For this reason, we can say that these molecular descriptors have a negative contribution to P. These are logical results because increasing the number of heteroatoms and the hydrogen bound to heteroatoms in the molecules decrease the permeability across the biological membrane. This effect is rather close to the molecular lipophilicity decrease and the possibility of the molecule of ionization and to obtain a charge. Other descriptors as $\boldsymbol{f}_{k}^{\mathrm{H}}(x)$, ${}^{s}\boldsymbol{f}_{k}(x)$ and ${}^{s}\boldsymbol{f}_{k}^{\mathrm{H}}(x)$ contains information about the molecular weight and consequently of the size of molecules. For this reason, we can say that these molecular descriptors have a positive contribution to P. In addition, these properties (molecular weight, size), H-bonding and charge are components of lipophilicity.[16] For each property there are limited ranges as it is established in the Rule- of-5,[18] but anyone is independent.[95]

Concluding Remarks

The development in drug discovery of combinatorial chemistry methods, where large numbers of candidate compounds are synthesized and screened in parallel for "*in vitro*" pharmacological activity, has dramatically increased the demand for rapid and efficient models for estimating human absorption. Therefore, the continuous definition of novel molecular descriptors that could explain different biological/pharmacokinetical properties, by means of QSAR, is necessary. Consequently, we have develop LDA and MLR models that could permit us to predict, by fast "*in silico*" screening, the intestinal permeability of chemicals and to outline preliminary conclusions about their possible human intestinal absorption profile.

In the current study, based on a larger set of drug or drug-like molecules, we developed discriminant functions that permit us the classification of molecules, in accordance with their molecular structure, with poorer-moderate or higher absorption. This result demonstrated that non-stochastic and stochastic atom-based linear indices are a novel and powerful approach to generate an adequate model for the correct classification of the intestinal permeability for structurally diverse drugs. Acceptable efficiency and a fairly good predictability were found for an external test set. Our method favorably compares with other approaches implemented in the Dragon software. Furthermore, rather satisfactory results were obtained by evaluating the capacity of prediction of human absorption for the obtained classification models. This approach permits that these indices could be applied to larger sets of NCEs synthesized via combinatorial chemistry approach.

Moreover, linear correlation models were developed to estimate Caco-2 permeability data. The predictions in the external test set demonstrate that this model bears good performance, as well as could be used for estimation of Caco-2 permeability for drug and drug-like molecules. The comparison with other approaches assesses a good behavior of our method.

Definitely, the models obtained in the current work are not ideal, because the data set used here, although the largest, is limited. A sufficiently large set of experimental data relating to this endpoint for the validation is crucial in the development of better prediction models. So, based on increasing data, the learning/modeling will need to be an ongoing, iterative process in which the models are continuously refined. Nevertheless, the present "*in silico*" method would be a valuable tool in the drug discovery process to select the molecules with the greatest change prior to synthesis.

References

1. Lin J, Sahakian DC, de Morais SM, Xu JJ, Polzer RJ, Winter SM 2003. The role of absorption, distribution, metabolism, excretion and toxicity in drug discovery. Curr Top Med Chem 3(10):1125-1154.
2. Fujikawa M, Ano R, Nakao K, Shimizu R, Akamatsu M 2005. Relationships between structure and high-throughput screening permeability of diverse drugs with artificial membranes: application to prediction of Caco-2 cell permeability. Bioorg Med Chem 13(15):4721-4732.
3. Guangli M, Yiyu C 2006. Predicting Caco-2 permeability using support vector machine and chemistry development kit. J Pharm Pharmacol 9(2):210-221.
4. Bohets H, Annaert P, Mannens G, Van Beijsterveldt L, Anciaux K, Verboven P, Meuldermans W, Lavrijsen K 2001. Strategies for absorption screening in drug discovery and development. Curr Top Med Chem 1(5):367-383.
5. Hidalgo IJ, Raub TJ, Borchardt RT 1989. Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability. Gastroenterology 96(3):736-749.
6. Yazdanian M, Glynn SL, Wright JL, Hawi A 1998. Correlating partitioning and caco-2 cell permeability of structurally diverse small molecular weight compounds. Pharm Res 15(9):1490-1494.
7. Artursson P, Karlsson J 1991. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. Biochem Biophys Res Commun 175(3):880-885.
8. Delie F, Rubas W 1997. A human colonic cell line sharing similarities with enterocytes as a model to examine oral absorption: advantages and limitations of the Caco-2 model. Crit Rev Ther Drug Carrier Syst 14(3):221-286.
9. Anderle P, Niederer E, Rubas W, Hilgendorf C, Spahn-Langguth H, Wunderli-Allenspach H, Merkle HP, Langguth P 1998. P-Glycoprotein (P-gp) mediated efflux in Caco-2 cell monolayers: the influence of culturing conditions and drug exposure on P-gp expression levels. J Pharm Sci 87(6):757-762.
10. Artursson P, Palm K, Luthman K 1996. Caco-2 Monolayer in Experimental and Theoretical Predictions of Drug Transport. Adv Drug Deliv Rev 22:67-84.
11. Rubas W, Jezyk N, Grass GM 1993. Comparison of the permeability characteristics of a human colonic epithelial (Caco-2) cell line to colon of rabbit, monkey, and dog intestine and human drug absorption. Pharm Res 10(1):113-118.

12. Chong S, Dando SA, Morrison RA 1997. Evaluation of Biocoat intestinal epithelium differentiation environment (3-day cultured Caco-2 cells) as an absorption screening model with improved productivity. Pharm Res 14(12):1835-1837.
13. Yee S 1997. In vitro permeability across Caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man--fact or myth. Pharm Res 14(6):763-766.
14. Chaturveldi PR, Deker CJ, Odinecs A 2001. Prediction of Pharmacokinetic Properties Using Experimental Approaches During Early Drug Discovery. Curr Opin Chem Biol 5:452-463.
15. Camenisch G, Alsenz J, van de Waterbeemd H, Folkers G 1998. Estimation of permeability by passive diffusion through Caco-2 cell monolayers using the drugs' lipophilicity and molecular weight. Eur J Pharm Sci 6(4):317-324.
16. van de Waterbeemd H, Camenisch G 1996. Estimation of Caco-2 cell Permeability Using Calculated Molecular Descriptors. Quant Struct-Act Relat 15:480-490.
17. Ren S, Lien EJ 2000. Caco-2 cell permeability vs human gastrointestinal absorption: QSPR analysis. Prog Drug Res 54:1-23.
18. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development. Adv Drug Deliv Rev 23(1):3-25.
19. Sugawara M, Takekuma Y, Yamada H, Kobayashi M, Iseki K, Miyazaki K 1998. A general approach for the prediction of the intestinal absorption of drugs: regression analysis using the physicochemical properties and drug-membrane electrostatic interaction. J Pharm Sci 87(8):960-966.
20. Clark DEDE, Pickett SDSD 2000. Computational methods for the prediction of 'drug-likeness'. Drug Discovery Today 5(2):49-58.
21. Palm K, Luthman K, Ungell AL, Strandlund G, Artursson P 1996. Correlation of drug absorption with molecular surface properties. J Pharm Sci 85(1):32-39.
22. Hamilton HW, Steinbaugh BA, Stewart BH, Chan OH, Schmid HL, Schroeder R, Ryan MJ, Keiser J, Taylor MD, Blankley CJ 1995. Evaluation of physicochemical parameters important to the oral bioavailability of peptide-like compounds: implications for the synthesis of renin inhibitors. J Med Chem 38(9):1446-1455.
23. Abraham MH, Chadha HS, Mitchell RC 1994. Hydrogen bonding. 33. Factors that influence the distribution of solutes between blood and brain. J Pharm Sci 83(9):1257-1268.
24. Basak SC, Gute BD, Drewes LR 1996. Predicting blood-brain transport of drugs: a computational approach. Pharm Res 13(5):775-778.

25. Yoshida F, Topliss JG 1996. Unified model for the corneal permeability of related and diverse compounds with respect to their physicochemical properties. J Pharm Sci 85(8):819-823.
26. Wessel MD, Jurs PC, Tolan JW, Muskal SM 1998. Prediction of human intestinal absorption of drug compounds from molecular structure. J Chem Inf Comput Sci 38(4):726-735.
27. Norinder U, Osterberg T, Artursson P 1997. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. Pharm Res 14(12):1786-1791.
28. Pickett SD, McLay IM, Clark DE 2000. Enhancing the hit-to-lead properties of lead optimization libraries. J Chem Inf Comput Sci 40(2):263-272.
29. Kulkarni A, Han Y, Hopfinger AJ 2002. Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. J Chem Inf Comput Sci 42(2):331-342.
30. Fujiwara S, Yamashita F, Hashida M 2002. Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. Int J Pharm 237(1-2):95-105.
31. Ponce YM, Cabrera M, A., Romero V, Ofori E, Montero LA 2003. Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2 Permeability of Drugs. Int J Mol Sci 4:512-536.
32. Hou TJ, Zhang W, Xia K, Qiao XB, Xu XJ 2004. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. J Chem Inf Comput Sci 44(5):1585-1600.
33. Marrero-Ponce Y, Cabrera MA, Romero-Zaldivar V, Bermejo M, Siverio D, Torrens F 2005. Prediction of Intestinal Epithelial Transport of Drug in (Caco-2) Cell Culture from Molecular Structure using in silico Approaches During Early Drug Discovery. Int Electron J Mol Des 4 124-150.
34. Refsgaard HHF, Jensen BF, Brockhoff PB, Padkjaer SB, Guldbrandt M, Christensen MS 2005. In silico prediction of membrane permeability from calculated molecular parameters. J Med Chem 48(3):805-811.
35. Yamashita F, Hashida M 2004. In silico approaches for predicting ADME properties of drugs. Drug Metab Pharm 19(5):327-338.
36. Marrero Ponce Y, Cabrera Perez MA, Romero Zaldivar V, Gonzalez Diaz H, Torrens F 2004. A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. J Pharm Pharm Sci 7(2):186-199.

37. Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T, Vianello R 2005. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. J Med Chem 48(22):6970-6979.
38. Todeschini R, Consonni V, Pavan M. Dragon Software version 2.1, 2002. ed.
39. Marrero-Ponce Y, Romero V. 2002. TOMOCOMD software. TOMOCOMD (TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es. ed., Central University of Las Villas.
40. Marrero-Ponce Y 2004. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. J Chem Inf Comput Sci 44(6):2010-2026.
41. Marrero-Ponce Y, Castillo-Garit JA, Torrens F, Romero-Zaldivar V, Castro E 2004. Atom, Atom-Type, and Total Linear Indices of the ¨Molecular Pseudograph´s Atom Adjacency Matrix¨: Application to QSPR/QSAR Studies of Organic Compounds. Molecules 9:1100-1123.
42. Marrero-Ponce Y, Castillo-Garit JA, Olazabal E, Serrano HS, Morales A, Castanedo N, Ibarra-Velarde F, Huesca-Guillen A, Sanchez AM, Torrens F, Castro EA 2005. Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. Bioorg Med Chem 13(4):1005-1020.
43. Marrero-Ponce Y, Montero-Torres A, Zaldivar CR, Veitia MI, Perez MM, Sanchez RN 2005. Non-stochastic and stochastic linear indices of the 'molecular pseudograph's atom adjacency matrix': application to 'in silico' studies for the rational discovery of new antimalarial compounds. Bioorg Med Chem 13(4):1293-1304.
44. Marrero-Ponce Y, Marrero RM, Torrens F, Martinez Y, Bernal MG, Zaldivar VR, Castro EA, Abalo RG 2006. Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and "rational" selection of new lead antibacterial agents. J Mol Model 12(3):255-271.
45. Marrero-Ponce Y, Machado-Tugores Y, Pereira DM, Escario JA, Barrio AG, Nogal-Ruiz JJ, Ochoa C, Aran VJ, Martinez-Fernandez AR, Sanchez RN, Montero-Torres A, Torrens F, Meneses-Marcel A 2005. A computer-based approach to the rational discovery of new trichomonacidal drugs by atom-type linear indices. Curr Drug Discov Technol 2(4):245-265.

46. Marrero-Ponce Y, Castillo-Garit JA 2005. 3D-chiral Atom, Atom-type, and Total Non-stochastic and Stochastic Molecular Linear Indices and their Applications to Central Chirality Codification. J Comput-Aided Mol Design 19(6):369-383.
47. Kier L. B, Hall L, H. 1986. Molecular Connectivity in Structure–Activity Analysis. ed., Letchworth, U. K: Research Studies Press.
48. Consonni V, Todeschini R, Pavan M 2002. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the Novel 3D Molecular Descriptors. J Chem Inf Comput Sci 42:682-692.
49. Todeschini R, Gramatica P 1998. New 3D Molecular Descriptors: The WHIM Theory and QSAR Applications. Persp Drug Disc Des 9-11:355–380.
50. Pauling L. 1939. The Nature of Chemical Bond. ed., Ithaca, NY: Cornell University Press. p 2-60.
51. Ruecker G, Ruecker C 1993. Counts of all walks as atomic and molecular descriptors J Chem Inf Comput Sci 33:683-695.
52. Gutman I, Rücker C, Rücker G 2001. On walks in molecular graphs. J Chem Inf Comput Sci 41(3):739-745.
53. Pearlman RS, Smith KM 1999. Metric Validation and the Receptor-Relevant Subspace Concept J Chem Inf Comput Sci 39:28-35.
54. Galvez J, Garcia R, Salabert MT, Soler R 1994. Charge Indexes. New Topological Descriptors. J Chem Inf Comput Sci 34:520-525.
55. Gálvez J, Garcia-Domenech R, de Julián-Ortiz JV, Soler R 1995. Topological Approach to Drug Design J Chem Inf Comput Sci 35:272-284.
56. Artursson P 1990. Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. J Pharm Sci 79(6):476-482.
57. Haeberlin B, Rubas W, Nolen HW, 3rd, Friend DR 1993. In vitro evaluation of dexamethasone-beta-D-glucuronide for colon-specific drug delivery. Pharm Res 10(11):1553-1562.
58. Hovgaard L, Brondsted H, Buur A, Bundgaard H 1995. Drug delivery studies in Caco-2 monolayers. Synthesis, hydrolysis, and transport of O-cyclopropane carboxylic acid ester prodrugs of various beta-blocking agents. Pharm Res 12(3):387-392.
59. Augustijns P, D'Hulst A, Van Daele J, Kinget R 1996. Transport of artemisinin and sodium artesunate in Caco-2 intestinal epithelial cells. J Pharm Sci 85(6):577-579.

60. Collett A, Sims E, Walker D, He YL, Ayrton J, Rowland M, Warhurst G 1996. Comparison of HT29-18-C1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. Pharm Res 13(2):216-221.
61. Schipper NG, Osterberg T, Wrange U, Westberg C, Sokolowski A, Rai R, Young W, Sjostrom B 2001. In vitro intestinal permeability of factor Xa inhibitors: influence of chemical structure on passive transport and susceptibility to efflux. Pharm Res 18(12):1735-1741.
62. Zhu C, Jiang L, Chen TM, Hwang KK 2002. A comparative study of artificial membrane permeability assay for high throughput profiling of drug absorption potential. Eur J Med Chem 37(5):399-407.
63. Aungst BJ, Nguyen NH, Bulgarelli JP, Oates-Lenz K 2000. The influence of donor and reservoir additives on Caco-2 permeability and secretory transport of HIV protease inhibitors and other lipophilic compounds. Pharm Res 17(10):1175-1180.
64. Ruiz-Garcia A, Lin H, Pla-Delfina JM, Hu M 2002. Kinetic characterization of secretory transport of a new ciprofloxacin derivative (CNV97100) across Caco-2 cell monolayers. J Pharm Sci 91(12):2511-2519.
65. Wu X, Whitfield LR, Stewart BH 2000. Atorvastatin transport in the Caco-2 cell model: contributions of P-glycoprotein and the proton-monocarboxylic acid co-transporter. Pharm Res 17(2):209-215.
66. Liang E, Proudfoot J, Yazdanian M 2000. Mechanisms of transport and structure-permeability relationship of sulfasalazine and its analogs in Caco-2 cell monolayers. Pharm Res 17(10):1168-1174.
67. Lentz KA, Polli JW, Wring SA, Humphreys JE, Polli JE 2000. Influence of passive permeability on apparent P-glycoprotein kinetics. Pharm Res 17(12):1456-1460.
68. Saha P, Kou JH 2002. Effect of bovine serum albumin on drug permeability estimation across Caco-2 monolayers. Eur J Pharm Biopharm 54(3):319-324.
69. Stenberg P, Norinder U, Luthman K, Artursson P 2001. Experimental and computational screening models for the prediction of intestinal drug absorption. J Med Chem 44(12):1927-1937.
70. Artursson P, Magnusson C 1990. Epithelial transport of drugs in cell culture. II: Effect of extracellular calcium concentration on the paracellular transport of drugs of different lipophilicities across monolayers of intestinal epithelial (Caco-2) cells. J Pharm Sci 79(7):595-600.

71. Hilgendorf C, Spahn-Langguth H, Regardh CG, Lipka E, Amidon GL, Langguth P 2000. Caco-2 versus Caco-2/HT29-MTX co-cultured cell lines: permeabilities via diffusion, inside- and outside-directed carrier-mediated transport. J Pharm Sci 89(1):63-75.
72. Gres MC, Julian B, Bourrie M, Meunier V, Roques C, Berger M, Boulenc X, Berger Y, Fabre G 1998. Correlation between oral drug absorption in humans, and apparent drug permeability in TC-7 cells, a human epithelial intestinal cell line: comparison with the parental Caco-2 cell line. Pharm Res 15(5):726-733.
73. STATISTICA version. 6.0 Statsoft I.
74. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H 2000. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412-424.
75. Metz CE 1978. Basic principles of ROC analysis. Sem Nuc Med 8:283-298.
76. Belsey DA, Kuh E, Welsch RE. 1980. Regression Diagnostics. ed., New York: Wiley.
77. Golbraikh A, Tropsha A 2002. Beware of q2! J Mol Graph Model 20(4):269-276.
78. Randic M 1991. Generalized Molecular Descriptors. J Math Chem 7:155-168.
79. Cramer RD, III. 1980. BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid State. J Am Chem Soc 102:1837-1849.
80. Cramer RD, III. 1980. BC(DEF) Parameters. 2. An Empirical Structure Based Scheme for the Prediction of Some Physical Properties. J Am Chem Soc 102:1849-1859.
81. Needham DE, Wei IC, Seybold PG 1988. Molecular Modeling of the Physical Properties of Alkanes. . J Am Chem Soc 110:4186-4194.
82. Malinowski ER, Howery DG. 1980. Factor Analysis in Chemistry. ed., New York: Wiley-Interscience.
83. Franke R. 1984. Theoretical Drug Design Methods. ed., Amsterdam: Elsevier. p 197-188.
84. Estrada E, Gonzalez H 2003. What are the limits of applicability for graph theoretic descriptors in QSPR/QSAR? Modeling dipole moments of aromatic compounds with TOPS-MODE descriptors. J Chem Inf Comput Sci 43(1):75-84.
85. Estrada E 1999. Edge-Connectivity Indices in QSPR/QSAR Studies. 2. Accounting for Long-Range Bond Contributions. J Chem Inf Comput Sci 39:1042-1048.
86. Estrada E, Rodríguez L 1999. Edge-Connectivity Indices in QSPR/QSAR Studies. 1. Comparison to Other Topological Indices in QSPR Studies. J Chem Inf Comput Sci 39:1037-1041.
87. González-Díaz H, Marrero-Ponce Y, Hernández I, Bastida I, Tenorio E, Nasco O, Uriarte U, Castañedo N, Cabrera MA, Aguila E, Marrero O, Morales A, Pérez M 2003. 3D-MEDNEs:

An Alternative “In Silico” Technique for Chemical Research in Toxicology. 1. Prediction of Chemically Induced Agranulocytosis. . Chem Res Toxicol 16:1318-1327.

88. Penney KB, Smith CJ, Allen JC 1984. Depigmenting action of hydroquinone depends on disruption of fundamental cell processes. J Invest Dermatol 82(4):308-310.
89. Clark DE 1999. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. J Pharm Sci 88(8):807-814.
90. Palm K, Stenberg P, Luthman K, Artursson P 1997. Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm Res 14(5):568-571.
91. Chiou WL, Barve A 1998. Linear correlation of the fraction of oral dose absorbed of 64 drugs between humans and rats. Pharm Res 15(11):1792-1795.
92. Zhao YH, Le J, Abraham MH, Hersey A, Eddershaw PJ, Luscombe CN, Butina D, Beck G, Sherborne B, Cooper I, Platts JA, Boutina D 2001. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. J Pharm Sci 90(6):749-784.
93. Benet LZ, Øie S, Schwartz JB. 1996. Design and Optimisation of Dosage Regiments; Pharmacokinetic Data. In Hardman JG, Limbird LE, Gilman AG, editors. Pharmacological Basis of Terapeutics, 9th ed., New York: McGraw-Hill. p 1707-1793.
94. Egan WJ, Merz KM, Jr., Baldwin JJ 2000. Prediction of drug absorption using multivariate statistics. J Med Chem 43(21):3867-3877.
95. van de Waterbeemd H, Smith DA, Jones BC 2001. Lipophilicity in PK design: methyl, ethyl, futile. J Comput Aided Mol Des 15(3):273-286.