Prediction of ADME properties, Part 1: Classification models to predict Caco-2 cell permeability using atom-based bilinear indices.

Juan A. Castillo-Garit,[a,b,c,*] Yudith Cañizares-Carmenate,[b] Yovani Marrero-Ponce,[b,d] Francisco Torrens,[d] and Concepción Abad[c]

[a]Centro de Estudio de Química Aplicada, Facultad de Química-Farmacia, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba. [b]"Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Facultad de Química-Farmacia, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830, Villa Clara, Cuba. [c]Departament de Bioquímica i Biologia Molecular, Universitat de València, E-46100 Burjassot, Spain.[d]Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain.

Predicción de las propiedades de ADME , Parte 1 : modelos de clasificación para predecir Caco- 2 la permeabilidad celular utilizando índices bilineales basado en Atom

Predicció de les propietats d'ADME, Part 1: models de classificació per predir Caco-2 la permeabilitat cel·lular utilitzant índexs atòmics bilineals.

Recibido: 30 de mayo de 2013; revisado: 2 de diciembre de 2013; aceptado: 4 de diciembre de 2013

RESUMEN

Se lleva a cabo la predicción de la permeabilidad a través de células cultivadas de Caco-2 (de uso frecuente en el modelo *in vitro* para la absorción de fármacos), usando modelos teóricos. Se utilizan índices atómicos bilineales y análisis discriminante lineal (LDA) para obtener modelos cuantitativos, que discriminan entre compuestos con una elevada absorción y compuestos con absorción baja o moderada, que forman una base de datos de medida P_{Caco-2} a partir de un gran conjunto de 157 compuestos estructuralmente diversos. Desarrollamos dos modelos LDA, con más de un 90% de exactitud para los conjuntos de prueba y de ensayo; el mejor modelo presenta una precisión de 91,79 % y 91,30 %, respectivamente. Los resultados obtenidos en este trabajo se comparan favorablemente con otros métodos publicados anteriormente en la literatura técnica. El porcentaje de buena correlación fue del 80% en el cribado virtual de 241 fármacos con los valores reportados del porcentaje de absorción intestinal humana (EIS). Por último, podemos decir que en el presente *"in silico"* método podría ser una herramienta valiosa en el proceso de descubrimiento de fármacos con el fin de seleccionar las moléculas con las mayores posibilidades antes de la síntesis .

Palabras clave: Células Caco-2, índices atómicos bilineales; ADME computacional; modelización '*in silico*', cribado virtual, absorción intestinal humana

SUMMARY

The prediction of the permeability through cultured Caco-2 cells (an often-used *in vitro* model for drug absorption) is carried out using theoretical models. Atom-based bilinear indices and linear discriminant analysis (LDA) are used to obtain quantitative models, which discriminate between higher absorption and moderate-poorer absorption compounds, form a database of measured $\mathbf{P_{Caco-2}}$ from a large data set with 157 structurally diverse compounds. We develop two LDA models with more than 90% of accuracy for training and test set; the best model presents accuracy of 91.79% and 91.30%, respectively. The results achieved in this work compare favourably with other approaches previously published in the technical literature. The percentage of good correlation was of 80%, in the virtual screening of 241 drugs with the reported values of the percentage of human intestinal absorption (HIA). Finally, we can say that, the present *"in silico"* method would be a valuable tool in the drug discovery process in order to select the molecules with the greatest chance before synthesis.

Keywords: Caco-2 cell; atom-based bilinear indices; computational ADME; '*in silico*' modelling; virtual screening; human intestinal absorption.

*To whom correspondence should be addressed: Telephone: 53-42-281510; Fax: 53-42-281130; jacgarit@yahoo.es; juancg.22@gmail.com or juancg@uclv.edu.cu

RESUM

Es porta a terme la predicció de la permeabilitat a través de les cèl·lules cultivades de Caco-2 (d'ús freqüent en el models *in vitro* per a l'absorció de fàrmacs) utilitzant models teòrics. S'utilitzen índexs atòmics bilineals i l'anàlisi discriminant lineal (LDA) per obtenir models quantitatius que discriminen entre substàncies amb absorció elevada i substàncies d'absorció baixa o moderada, que formen una base de dades de valors de P_{Caco-2}, determinats a partir d'un gran conjunt de 157 compostos estructuralment diferents. Desenvolupem dos models LDA amb més d'un 90 % d'exactitud en els conjunts de prova i d'assaig; el millor model presenta una exactitud de 91,79 % i 91,30 %, respectivament. Els resultats obtinguts en aquest treball es comparen favorablement amb altres mètodes publicats anteriorment en la literatura tècnica. El percentatge de bona correlació va ser del 80% en el cribratge virtual de 241 fàrmacs amb els valors publicats del percentatge d'absorció intestinal humana (EIS) . Finalment, podem dir que aquest mètode "*in silico*" podria ser una eina valuosa en el procés de descobriment de fàrmacs per tal de seleccionar, abans de la síntesi, les molècules amb les millors possibilitats .

Paraules clau : Cèl·lules Caco-2, els índexs atòmics bilineals; ADME computacional; modelització *'in silico'*; cribratge virtual, absorció intestinal humana .

1. INTRODUCTION

Oral administration of drugs is the most appropriate route in many cases for its convenience, low cost and high patient compliance rates. In an organism, the drug is absorbed from the gastro-intestinal tract, and the transport across the intestinal epithelial cell barrier may occur by one or more of four different routes: the passive transcellular and paracellular routes, the carrier mediated route and by transcytosis.[1-5] The majority of the absorption in the gastro-intestinal tract occurs in the small intestine, whose epithelium contains a heterogeneous population of cells between which one can find enterocytes. Enterocytes are considered absorptive cells and are the most abundant cells (80-90%). The tight junctions present between these cells and lipophilic nature of the intestinal epithelium serve as a physical barrier to the absorption of orally administered drugs, whereas the metabolizing enzymes expressed by the enterocytes constitute a biochemical barrier.[6]

In order to obtain a rapid estimation of human absorption, many cells culture models have been investigated as potential *in vitro* models for drug absorption and metabolic studies. *In vitro* models of intestinal absorption generally focus on determining membrane permeability using Caco-2 cells, MDCK cells, artificial membranes, immobilized artificial membrane (IAM) columns and parallel artificial membrane permeation assay (PAMPA).[7]

Caco-2 monolayer is the most advanced *in vitro* model because of this cell line expressed several of the biological membrane properties.[8] These are well-differentiated intestinal cells derived from human colorectal carcinoma with 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.[3] The conventional Caco-2 cell monolayer suffers from the limitations of the 21-day-long culturing time, low levels/lack of cytochrome P450 3A4, lack of mucus layer, tighter junctions compared to the small intestine, nonspecific drug binding to plastic devices and/or cells (cacophilicity) and variable expression of transporters but can be considered a highly valuable tool for prediction of oral absorption potential of new drug candidates and optimization of drug delivery systems, provided that the shortcomings associated with the model are understood, appreciated and corrected.[6]

In recent years with the advanced of combinatorial chemistry it has been increased the synthesis of new chemical entities and, with it, the cost of compounds evaluation by traditional methods. Therefore, it should be also expected an increasing use of Quantitative Structure Activity/Property Relationship (QSAR/QSPR) methods in the estimation of absorption parameters from new potentiality active chemical entities during the drug discovery and development processes. These methods are quantitative approaches focused at finding relationships between molecular structure and properties/activities either measured or calculated.

Studies QSPR/QSAR date from 1868 and have been applied to the modeling of dissimilar physical, chemical and biological molecular properties.[9, 10] In these approaches, the descriptors or molecular indices play a fundamental role and are currently defined as a result of a logical-mathematical procedure, which transforms chemical information encoded within a symbolic representation of a molecule in a useful number.

In this context, our research group has recently developed a novel scheme to generate molecular fingerprints based on the application of discrete mathematics and linear algebra theory, known as ***TOMOCOMD*** (acronym of *TOpological MOlecular COMputational Design*).[11] This scheme has been successfully applied to the prediction of several physical, physicochemical, chemical, and biological properties.[12-15]

In the present report, atom-based non-stochastic and stochastic bilinear indices are used to find classification models that allow the discrimination of Caco-2 absorption compounds.

2. MATERIALS AND METHODS

2.1 Permeability data

In this study, we used a data set of measured $\mathbf{P_{Caco-2}}$ consisting of 157 structurally diverse compounds. It was compiled from several published works[16-39] 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 $\mathbf{P_{Caco-2}}$ (8×10^{-6} cm/s). This value of $\mathbf{P_{Caco-2}}$ was fixed taking into consideration the experimental results reported in the literature and the wide inter-laboratory variability.[40, 41] The molecular structures of these 157 compounds are given as Supplementary Material (see Table S1).

Experimental values of $\mathbf{P_{Caco-2}}$ (AP→BL) (APical→BasoLateral), for both training and test set, are also in the Supplementary Material with their references (see Table S2). The data set used for '*in silico*' permeability studies included compounds with a diverse molecular weight and charge. In addition, compounds with different absorption mechanisms were included in the model.

2.2 Computational strategies

The theory of the atom-based bilinear indices used in this study was discussed in detail in earlier publications.[42] Specifically, the *CARDD* (*C*omputed-*A*ided *R*ational *D*rug *D*esign) module implemented in the ***TOMOCOMD*** Software[11] was used in the calculation of atom-based non-stochastic and stochastic bilinear indices. In this study, the properties used to differentiate the molecular atoms are those previously proposed for the calculation of the DRAGON descriptors, i.e.,[43-45] atomic mass (M), atomic polarizability (P), atomic Mulliken electronegativity (K), van der Waals atomic volume (V), plus the atomic electronegativity in Pauling scale (G).[46]

The following descriptors were calculated in this work:

(I) the k^{th} non-stochastic total bilinear indices, not considering and considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b}_k(\bar{x}, \bar{y})$ and $\boldsymbol{b}_k^{H}(\bar{x}, \bar{y})$, respectively].

(II) the k^{th} non-stochastic local (atomic group = heteroatoms: S, N, O) bilinear indices, not considering and considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b}_{kL}(\bar{x}_{E}, \bar{y}_{E})$ and $\boldsymbol{b}_{kL}^{H}(\bar{x}_{E}, \bar{y}_{E})$, correspondingly]. These local descriptors denote putative H-bonding acceptors; in addition, they represent charge as well as dipole moment.

(III) the k^{th} non-stochastic local (atomic group = H-atoms bonding to heteroatoms: S, N, O) bilinear indices, considering H atoms in the molecular pseudograph (G) [$\boldsymbol{b}_{kL}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H})$]. These local descriptors denote putative H-bonding donors.

The k^{th} stochastic total [${}^{s}\boldsymbol{b}_k(\bar{x}, \bar{y})$ and ${}^{s}\boldsymbol{b}_k^{H}(\bar{x}, \bar{y})$] and local [${}^{s}\boldsymbol{b}_{kL}(\bar{x}_{E}, \bar{y}_{E})$, ${}^{s}\boldsymbol{b}_{kL}^{H}(\bar{x}_{E}, \bar{y}_{E})$ and ${}^{s}\boldsymbol{b}_{kL}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H})$] bilinear indices were also computed.

2.3 Chemometric analysis

The LDA was performed with the STATISTICA software package.[47] The quality of the models was determined by examining Wilks´ λ parameter (U-statistic), square Mahalanobis distance (D^2), Fisher ratio (F) and the corresponding p-level [*p*(F)], as well as the accuracy, Matthews' correlation coefficient, sensitivity, specificity and false positive rate (false alarm rate).[48]

One of the crucial steps consists in the statistical validation of the results to determine its reliability and significance, while providing an indication of how well the model can predict activity for new molecules. Several procedures are available for this task, and in the present work we carry out both internal and external ones.

As internal validation we performed two experiments: a *leave-group-out* (LGO) and a *randomization test (Y-scrambling)*.[49]

(I) Namely, the *Cross-validation (CV)* is the most commonly used statistical technique for internal validation in which different proportions of the TS are iteratively held-out; then a new model is developed and used to "predict" the held-out compound as new in order to the verify internal "predictability". This procedure is repeated for each set of modified data. The LGO experiment is made by introducing the large perturbation in the data set; the predictability estimated by LGO is more realistic than the one by leave-one-out (LOO). Here the LGO test is used in such a way that a fraction of the 5%, 10%, 15%, 20%, 25%, 30% and 35% of the training set and model predictions are made based on the reduced data. These compounds were removed in a *random* way. This process is repeated until each observation has been left out once. The accuracies for the reduced training set and test set are calculated and plotted. Usually, good results in this experiment are considered as a confirmation of the high predictive power of the models. However, this assumption is not always correct, and it can result that there is a poor correlation between the good LGO results and the high predictive ability of QSAR models. Thus, the good behavior of models in an LGO procedure appears to be a necessary but not sufficient condition for the models to show a high predictive power.

(II) *Y-scrambling*, or response permutation testing, is another technique widely use to check the robustness of a QSAR model and to identify models based on chance correlation. The set of activity values is reassigned randomly to different molecules and a new QSAR model is performed. This procedure is similar to cross-validation CV but, instead of leaving groups outside, it categorizes the assignments (1 for –1 and vice versa), for each group (active and inactive) of the database with 10%, 20%, 30% and 40% of the total. This process is repeated until each case has been inverted once. The accuracies for new models are calculated and plotted. If the quality of the random response models is comparable to the original one, the set of observations is not sufficient to support the model and the chance correlation is detected.

Nevertheless, the most important criterion, for the acceptance or not of a discriminant model, is based on statistics for the external prediction set (compounds that were not used for the development of the model) which is known as *the predictive power* of the model. The predictability of a model is estimated by comparing the predicted and observed classes of a representative test of compounds.

3. RESULTS AND DISCUSSION

3.1 Development of the linear discriminant analysis models.

Classification models obtained by using atom-based bilinear indices: In order to develop the LDA models, the data was conformed by 80 compounds with higher absorption ($P \geq 8x10^{-6}$cm/s) and 54 compounds with moderate-poorer absorption ($P < 8x10^{-6}$cm/s). The best obtained discrimination models are given below, together with the LDA statistical parameters:

$$\begin{aligned}\text{Class} = {} & 4.718 - 0.004\ {}^{MK}\boldsymbol{b}_{2L}^{H}(\bar{x}_{E}, \bar{y}_{E}) - 0.588\ {}^{MP}\boldsymbol{b}_{2L}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H}) \\ & + 0.088x10^{-6}\ {}^{MV}\boldsymbol{b}_{7}(\bar{x}, \bar{y}) + 0.006\ {}^{MV}\boldsymbol{b}_{1L}^{H}(\bar{x}_{E}, \bar{y}_{E}) - 0.010\ {}^{VK}\boldsymbol{b}_{1L}^{H}(\bar{x}_{E}, \bar{y}_{E}) \\ & + 0.017\ {}^{VK}\boldsymbol{b}_{3L}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H}) - 0.019\ {}^{VP}\boldsymbol{b}_{1L}(\bar{x}_{E}, \bar{y}_{E}) \qquad (1)\end{aligned}$$

$N = 134$ $\quad \lambda = 0.456$ $\quad D^2 = 4.88$ $\quad F = 21.45$ $\quad p < 0.001$

$Q_{Total} = 90.30\%$ $\quad MCC = 0.80$ $\quad Sen = 90.00\%$ $\quad Spec = 93.51\%$ $\quad FPR = 9.26\%$

$$\begin{aligned}\text{Class} = {} & 5.503 - 0.031\ {}^{MKs}\boldsymbol{b}_{2L}^{H}(\bar{x}_{E}, \bar{y}_{E}) + 0.035\ {}^{MKs}\boldsymbol{b}_{3L}(\bar{x}_{E}, \bar{y}_{E}) - 4.640\ {}^{MPs}\boldsymbol{b}_{2L}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H}) \\ & - 0.409\ {}^{PKs}\boldsymbol{b}_{1}(\bar{x}, \bar{y}) + 10.732\ {}^{PKs}\boldsymbol{b}_{12L}^{H}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H}) - 0.714\ {}^{VKs}\boldsymbol{b}_{15L}(\bar{x}_{E\text{-}H}, \bar{y}_{E\text{-}H}) \\ & + 0.053\ {}^{VPs}\boldsymbol{b}_{12}(\bar{x}, \bar{y}) \qquad (2)\end{aligned}$$

$N = 134$ $\quad \lambda = 0.398$ $\quad D^2 = 6.18$ $\quad F = 27.18$ $\quad p < 0.001$

$Q_{Total} = 91.79\%$ $\quad MCC = 0.83$ $\quad Sen = 93.75\%$ $\quad Spec = 92.59\%$ $\quad FPR = 11.11\%$

where N is the number of compounds, λ is the Wilks' statistic, D^2 is the square Mahalanobis distance, F is the

Table 1. *Results for the Classification of Compounds in Training and Test Sets thought the Discriminant Functions Obtained Using Non-stochastic and Stochastic Bilinear 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]	-68.70	-66.53	Methanol	92.87	90.87	Theophyline[c]	-15.12	58.62
Acetylsalicylic acid	77.69	59.64	Metoprolol	50.97	83.89	Guanoxan[c]	-79.15	17.53
Alprenolol	57.49	90.04	Naproxen	95.12	93.86	Lidocaine	84.05	90.85
Alprenolol ester	66.78	83.47	Nevirapine	73.13	82.36	Tiacrilast	79.21	54.89
Aminopyrine	94.59	98.01	Nicotine	97.51	99.41	Nitrendipine	68.99	80.49
Artemisin	88.31	82.68	Oxprenolol	51.50	88.67	Fleroxacin	75.69	75.80
Betaxolol	56.84	93.50	Oxprenolol ester	52.04	78.12	Verapamil	97.18	97.13
Betaxolol ester	57.33	85.49	Phencyclidine	98.45	99.80	Mibefradil	73.44	84.32
Bremazocine	94.60	98.33	Phenitoin[d]	3.95	-44.65	Naloxone	89.93	93.54
Caffeine	55.60	92.49	Pindolol	1.39	37.92	Taurocholic acid[c,d]	-98.60	-99.80
Chloranphenicol[c]	-33.99	98.62	Piroxicam	94.26	95.06	Tenidap	69.77	46.09
Chlorpromazine	99.42	99.95	Prazocin	52.41	72.46	Trovafloxacin	10.65	30.03
Clonidine	9.78	67.04	Progesterone	97.09	98.85	Acid valproic	87.70	47.82
Corticosterone	68.79	75.59	Propranolol	76.30	96.68	Ziprasidone	95.73	98.69
Desipramine	84.71	90.45	Propranolol ester	76.61	92.35	D-Glucose[c,d]	-68.61	-63.72
Dexamethasone	44.71	61.30	Quinidine	95.40	97.85	L-Phenylalanine	62.66	91.96
Diazepam	98.11	99.51	Salicylic acid	83.72	64.85	Ketoprofen	93.97	91.64
Dopamine	55.80	19.80	Scopolamine	76.54	67.29	SB 209670	66.64	27.96
Estradiol	94.83	96.76	Telmisartan	93.23	95.24	SB 217242	80.54	69.43
Felodipine	91.77	90.98	Testosterone	94.60	98.01	Sidenafil	95.40	86.44
Griseofulvin	96.96	96.43	Timolol	60.32	92.61	Oxazepam	75.08	78.18
Hydrocortisone	40.85	51.88	Timolol ester	60.79	86.16	Nordazepam	92.97	94.99
Ibuprophen	93.13	94.89	Warfarin	95.17	91.98	Alfentanil	48.98	82.27
Imipramine	98.83	99.78	Antipyrine	95.66	98.47	Glycine	18.06	74.27
Indomethacin	91.80	85.10	Diltiazen	93.83	98.70	Phe-Pro	10.46	32.50
Labetalol[c,d]	-42.30	-71.64	Guanabenz[c]	-52.18	90.04	Fluconazole	52.36	96.84
Meloxicam	88.84	87.45	Cumarin	97.19	98.51			
Test set								
DMP450	22.44	54.58	CNV97102	73.22	68.81	Nicardipine	72.34	91.75
DMP850[c]	-60.65	6.51	CNV97103	73.33	68.45	Sulfapyridine	26.73	34.90
Amprenavir	35.10	30.39	CNV97104	73.42	65.69	Descarboxysulfasalazine	29.09	58.89
CNV97101	74.59	69.01						
***Moderate-poor absorption group* (M-P)**								
Training set								
Acebutolol	-69.09	-45.97	Practolol	-58.05	-16.64	Azithromycin	-95.69	-96.04
Acyclovir	-93.75	-98.64	Ranitidine	-27.66	-90.50	Penicilin G	-31.53	-38.05
Artesunate[c,d]	79.15	55.03	Sucrose	-99.70	-99.93	H21644	-97.20	-99.22
Atenolol	-55.29	-26.19	Sulphasalazine	-53.33	-65.57	Sumatriptan	-35.38	-63.89
Chlorothiazide	-99.23	-97.99	Terbutaline[c,d]	6.43	5.10	Cephalexin	-75.40	-72.67
Cimetidine	-96.92	-99.44	Uracil	-59.66	-96.56	Gly-Pro	-39.10	-35.51
Dexamethasone-B-D-glucoside	-91.56	-89.95	Urea	-14.67	-76.41	Raffinose	-100.00	-100.0
Dexamethasone-B-D glucuronide	-95.24	-98.00	Ziduvudine	-99.14	-99.90	Metolazone	-93.71	-88.38
Doxorrubicin	-78.47	-94.20	Amoxicillin	-91.18	-96.72	Lactulose	-99.65	-99.93
Erythromycin	-96.21	-97.01	Enalapril	-50.09	-60.87	Foscarnet	-99.42	-95.71
Ganciclovir	-98.55	-99.78	Furosemide	-98.72	-99.50	Ciprofloxacin	-16.18	-52.23
Hydrochlorothiazide	-99.85	-99.93	Epinephrine	-21.48	-79.59	Amiloride	-99.19	-96.64
Mannitol	-87.05	-91.34	Sulpiride	-98.69	-99.27	BVARAU	-68.29	-91.70
Methotrexate	-99.55	-99.86	Bosentan	-51.54	-19.73	Lisinopril	-97.32	-99.12
Methylscopolamine[c,d]	75.83	61.18	Proscillaridin[d]	-35.20	44.29	SQ-29852	-94.37	-99.92
Nadolol[d]	-25.88	0.84	Ceftriaxone	-99.94	-99.82	L-Glutamine	-63.07	-76.22
Olsalazine	-13.56	-82.69	Remikiren	-97.14	-96.40	Pravastatin	-60.39	-98.88
Pirenzepine[c,d]	43.26	48.46	Saquinavir	-99.65	-99.76	Gabapentin[c]	35.61	-11.61
Test set								
PNU200603	-28.62	-47.09	DMP851	-72.96	-60.30	Nelfinavir[c]	1.90	-5.62
Cyclosporine	-100.0	-100.0	Losartan[d]	-17.71	54.52	Ritonavir	-99.66	-98.29
Homofasalazine	-55.86	-64.10	Lucifer Yellow	-100.0	-100.0	Vinblastine[d]	-4.53	25.59
Sulfasalamide	-68.19	-75.00	Indinavir	-85.19	-71.83	CNV97100	-1.62	-4.72
Artorvastatin	-8.97	-95.08						

[a,b]ΔP% = [P(Active) - P(Inactive)]x100.

[a]Classification of each compounds using the obtained model with non-stochastic bilinear indices.

[b]Classification of each compounds using the obtained model with stochastic bilinear indices.

[c]Compounds missclassified by Eq. 1

[d]Compounds missclassified by Eq. 2

Fisher ratio and p-value is its significance level, Q_{Total} is the accuracy (in percentage) of the model for the training set, MCC is the Matthews' correlation coefficient, *Sen* and *Spec* are the sensibility and specificity (in %) of the model, respectively, and FPR is the 'false positive rate' (in %).

The non-stochastic model (Eq.**1**) developed with non-stochastic indices, presents an accuracy of 90.30 % for the training set. This model showed a high MCC of 0.80; MCC quantifies the strength of the linear relationship between the molecular descriptors and the classifications; usually it may provide a much more balanced evaluation of the prediction than, for instance, the accuracy.[50] Together with the accuracy, sensitivity, specificity and false-positive rate (also known as 'false-alarm rate') appear among the most

commonly used parameters in medical statistics. While the sensitivity is the probability of correctly predicting a positive case, the specificity (also known as 'hit rate') is the probability that a positive prediction is correct [48]. This model showed, for the training set, a good sensitivity value of 90.00%, a specificity value of 93.51% and a false-positive rate of only 9.26%.

On the other hand, in the case of the stochastic bilinear indices (Eq. **2**), the model achieved a slightly greater accuracy of 91.79 % than the non-stochastic model; the MCC value was of 0.83. The reported values for sensitivity and specificity were 93.75% and 92.59%, respectively, as well as a false-positive rate of 11.11%; these values are similar to those obtained with the non-stochastic model. The results of classification and a posteriori probabilities for the compounds of the training set are shown in Table 1.

3.2 Validation

The *robustness* of the model refers to the stability of its parameters (predictor coefficients) and, consequently, to the stability of its predictions when a perturbation is applied to the training set and the model is regenerated from the "perturbed" training set. Here, we develop the leave-group-out (LGO) and *Y-scrambling* procedures [48, 49, 51] as very important tools in order to detect what is sometimes referred to as "internal predictability" and possible chance correlation in the obtained models, respectively.

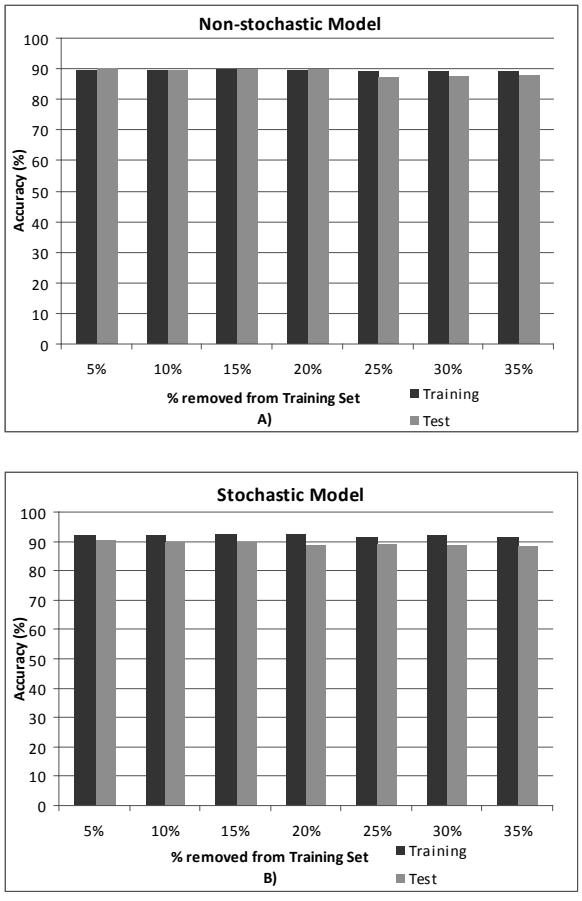

Figure 1. *Behaviour of the overall accuracy of the model in the LGO experiment: A) Non-stochastic model (Eq. 1); B) Stochastic model (Eq. 2)*

First, an LGO strategy was performed and the calculation of accuracies in the new training set and test set compounds permitted us to carry out the assessment of the models. The results of this validation process are illustrated in Figure 1. It can be observed from this plot that the models present a high stability to disturbances within the database. The results of the stochastic model were better than those obtained with the non-stochastic model.

After that, the *Y-scrambling* test was carried out. The results of our randomization experiments are shown in Figure 2 and indicate that, when the random group size is increased, the globally good accuracy of the model decayed gradually. This outcome indicates that the values of good overall classification are not because of chance correlation or structural redundancy in the training set.

However, the most important criterion, for the acceptance or not of a discriminant model, is based on the statistics for external prediction set. The non-stochastic model showed an accuracy of 91.30 % (MCC = 0.82) for the compounds in the test set. At the same time a good value of 90.00% of sensitivity and specificity was obtained and a false-positive rate of only 7.69%. The results of classification and a posteriori probabilities for the compounds of the test set are also shown in Table 1.

For the stochastic bilinear indices model, the results for the test set were an accuracy of 91.30%, MCC of 0.84 (quite similar to those obtained with the non-stochastic model); sensitivity of 100.00%, a specificity of 83.33% and false-positive rate of 15.38%; specificity value is somewhat smaller than that obtained with Eq. (**1**), but they still show an excellent predictive capacity.

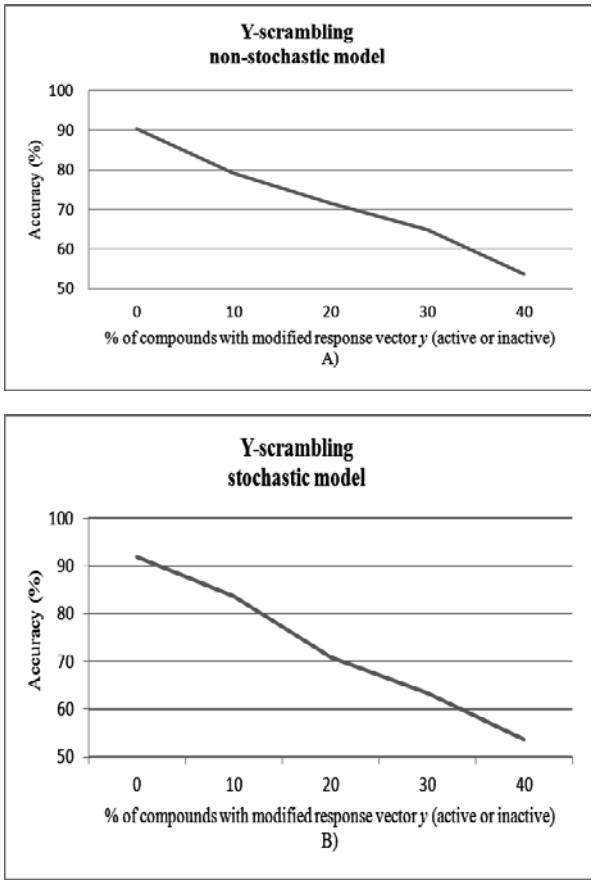

Figure 2. *Behaviour of the percentage of accuracy in the Y-scrambling analysis:* **A***) Non-stochastic model (Eq. 1);* **B***) Stochastic model (Eq. 2)*

3.3 Comparison with other approaches.

The use of atom-based non stochastic and stochastic bilinear indices, for the classification of molecules with poorer-moderate and higher absorption, was compared with other previously published methods.[52] All those models were also developed using LDA as statistical technique and several of them were carried out with a same data set size.

The first comparison was based on the quality of the statistical parameters of the discriminant function, as well as predictive capacity of the generated models. As can be seen, the present approach (non-stochastic atom-based bilinear indices) achieved the best values, for the statistical parameters of the developed QSAR models. Our models showed low values of the Wilks' lambda and high values of square Mahalanobis distance and Fisher ratio, similar to linear indices. All the models were significant, from the 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 stochastic bilinear indices (Q_{Total} = 91.79%), with non-stochastic linear indices (Q = 90.58%) and with non-stochastic bilinear indices and atom-based quadratic indices (Q = 90.30%), respectively. The best value of the sensitivity was obtained with quadratic indices (96.29%), but notice that our two models Eqs. **1** and **2** had a sensitivity of 90.00% and 93.75%, correspondingly. The specificity of our models reach values similar to the linear indices (greater than 90%), while the specificity of the rest of the approaches is between 70% and 89%. It is remarkable that both models showed a false alarm rate of 9.26% and 11.11%, correspondingly; these values are lower than those of the other models (between 14% and 38%) and are only improved by the value obtained with non-stochastic linear indices model.

An important point of view either to accept or reject a QSAR model is the statistics for the external prediction set.[53] For the test set, our models show an accuracy of 91.30%, which is the greatest value of accuracy (the accuracy of the other models was between 52.38% and 84.21%). The obtained MCC were 0.82 (for non-stochastic model) and 0.84 (for stochastic model), better than the results of other researches; the previously achieved greatest values were of 0.73 with the non-stochastic and stochastic linear indices. In addition, it should be noted that our test set (23 compounds) is larger than the other ones (11, 19 and 21, compounds). All these results are summarized in Table 2, where a comparison among different computational schemes can be performed more easily.

3.4 Virtual Screening.

The relevance of QSAR studies in the prediction of human intestinal absorption has been demonstrated in recent publications[54-58] and the so-called "*Rule-of-5*" has proved very popular as a rapid screen for compounds that are likely to be poorly absorbed.[59] In the present study, a virtual search was simulated to predict the absorption profile of 241 compounds,[58] using the obtained models with non-stochastic and stochastic bilinear indices. The aim of the present report is to evaluate the capacity of human absorption prediction from the classification models (Eqs. 1 and 2), into high and moderate-poor, for drug absorption in Caco-2 cells. Moreover, some compounds included in the obtained models (either 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*.[58] 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.[58] These experimental values and the evaluation results of these compounds are depicted in Table 3. 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 2. *Comparison between Atom-Based Bilinear Indices and Others Approaches Predicting Permeability across Caco-2 Cells.*

Models' features to be compared	**NS-BI[a] (Eq. 1)**	**St-BI[b] (Eq. 2)**	**NS-LI[c]**	**St-LI[d]**	**2D –Aut[e]**	**BCUT[f]**	**GCI[g]**	**TI[h]**	**MWC[i]**	**NS-QI[j]**
					Training set					
N	134	134	138	138	133	133	133	133	133	134
Wilks'λ (U-statistic)	0.456	0.398	0.435	0.436	0.568	0.891	0.743	0.543	0.936	0.480
F	21.45	27.18	20.94	20.86	13.55	2.56	7.25	15.00	1.44	16.88
D^2	4.88	6.18	5.31	5.29	3.12	0.50	1.42	3.45	0.28	4.52
p-level	0.001	0.001	0.001	0.001	0.001	0.023	0.001	0.001	0.204	0.001
Accuracy (%)	90.30	91.79	90.58	89.13	83.46	69.92	71.43	84.21	61.65	90.30
MCC	0.80	0.83	0.81	0.77	0.66	0.40	0.41	0.68	0.22	0.80
Sensitivity (%)	90.00	93.75	90.24	91.46	82.5	68.75	73.75	83.75	62.50	96.29
Specificity (%)	93.51	92.59	93.67	90.36	89.19	78.57	77.63	89.33	70.42	88.64
False Positive Rate (%)	9.26	11.11	8.93	14.29	15.09	28.30	32.07	15.09	39.62	18.86
					Test set					
N	23	23	19	19	21	21	21	21	21	11
Predictability (%)	91.30	91.30	84.21	84.21	52.38	57.14	71.43	52.38	61.90	83.33
MCC	0.82	0.84	0.73	0.73	0.07	0.15	0.43	0.06	0.24	0.71

Model developed with: [a]Non-stochastic Bilinear Indices (Eq. **1***), [b]Stochastic Bilinear Indices (Eq.* **2***), [c]Non-stochastic Linear Indices (Eq. 2 in [52]), [d]Stochastic Linear Indices (Eq. 3 in [52]), [e]2D autocorrelation indices (Eq. 4 in [52]), [f]BCUT indices (Eq. 5 in [52]), [g]Gálvez charge Indices (Eq. 6 in [52]), [h]Topological indices (Eq. 7 in [52]), [i]Walk count indices (Eq. 8 in [52]), [j]Quadratic Indices (Eq. 12 in [8])*

Initially, we analyze the first 145 compounds (data of the best quality), the correctness of good extrapolation (from "*in vitro*" to "*in vivo*") achieved by our models was 79.31% and 80.00%, correspondingly, when non-stochastic linear indices and stochastic linear indices were used. These values are similar to that previously obtained in some reports,[8, 52] for data of best quality. After that, we consider as a unique great group the rest of the experimental data (compounds 146-241), where lesser realistic data of Abs% are reported, the percentages of correct correspondence between "*in vitro"* permeability data (Caco-2 cells, predicted by bilinear indices) and the human absorption were 71.58% with non-stochastic and 73.68% with stochastic bilinear indices. This group presents a lesser percentage of correspondence than the first 145 previously mentioned compounds.

After that, we analyze the last group (146-241) but divided into several subgroups according to the Zhao *et al.* classification.[58] Compounds from 146 to 172 were considered as uncertain and unchecked data. For these compounds, the global good classifications were of 76.92% for both models. In addition, it was analyzed the subgroup of twenty zwitterionic drugs (173-192) reported by Zhao *et al.*[58] For this kind of drugs, our models show only a 50% and 57.14 % for Eq. **1** and Eq. **2,** correspondingly. The result previously obtained with linear and quadratic indices was of 50%.[8, 52] The prediction of compounds 193-201 (subgroup 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 the criteria followed by Zhao *et al*. to classify these compounds.[58] For the subgroup with dose-limited, dose-dependent and formulation-dependent drugs the correspondence between "*in vitro"* permeabilities and the human absorption for our models were lesser than the results previously achieved with linear indices. Finally, for the analysis of drugs with expected higher absorption, according to Zhao *et al.*[58], 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 3 are considered, our two models explain 91.67% of the experimental variance.

Table 3. *Results for the Virtual Screening of 241 Drugs; Permeability Coefficient from Models (Eq. 1 and Eq.2) and Observed Human Absorption and Bioavailability from Literature.*

Compounds	ΔP[a]	ΔP[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
1-Cisapride	0.54	0.14	100		100	100
2-Valproic acid	0.88	0.48	100	~100	90(68-100)	100
3-Salicylic acid	0.84	0.65	100	100		100
4-Diazepam	0.97	0.99	97-100	100		100
5-Sudoxicam	0.89	0.86		100		100
6-Glyburide	-0.97	-1.00				100
7-Gallopamil	0.97	0.95		~100	15	100
8-Mexiletine	0.93	1.00		100	88	100
9-Nefazodone	0.77	0.98		100	15-23	100
10-Naproxen	0.95	0.94	94-99	100	99	99
11-Lamotrigine	0.79	0.97	70		98	98
12-Tolmesoxide	1.00	1.00	100		85	98
13-Disulfiran	0.99	1.00		91		97
14-Torasemide	-0.92	-0.86			96	96
15-Metoprolol	0.34	0.86	95-100	>90	50	95
16-Naloxone	0.91	0.93				91
17-Terazocin	0.47	0.41	91	~100	90	90
18-Sulindac	0.93	0.70		90		90
19-Sultopride	0.93	0.90	100	~100		89
20-Topiramate	-0.99	-1.00			81-95	86
21-Tolbutamide	-0.78	-0.86				85
22-Propiverine	0.97	0.99		84		84
23-Digoxin	-0.97	-1.00			67	81
24-Mercapto ethane sulfonic acid	0.39	-0.92				77
25-Cimetidine	-0.97	-0.99	62-98		60	64
26-Furosemide	-0.99	-1.00	61	61	61	61
27-Metformin	-0.98	-0.99			50-60	53
28-Rimiterol	0.09	-0.19				48
29-Cymarin	0.04	-0.49		47		47
30-Ascorbic Acid	-0.02	-0.14				35
31-Fosfomycin	-0.96	-0.97				31
32-Fosmidomycin	-0.99	-0.99		30		30
33-k-Strophanthoside	0.97	0.95		16		16
34-Adefovir	-1.00	-1.00	12		12	16
35-Acarbose	-1.00	-1.00		1-2.		2
36-Ouabain	0.93	0.97		1.4		1.4
37-Kanamycin	-1.00	-1.00				1
38-Lactulose	-1.00	-1.00	0.6	0.6		0.6
39-Camazepan	0.89	0.98	99		100	100
40-Indomethacin	0.92	0.85	100		100	100
41-Levomorgestrel	0.97	0.99			100	100
42-Tenoxicam	0.99	0.98			100	100
43-Theophyline	-0.15	0.59	96		100	100
44-Oxatomide	0.77	0.95	100			100
45-Desipramine	0.85	0.90	95-100	>95	40	100
46-Fenclofenac	0.96	0.96	100			100
47-Imipramine	0.99	1.00	95-100	>95	22-67	100
48-Lormetazepan	0.94	0.98	100	100	80	100
49-Diclofenac	0.90	0.91	100		90	100
50-Granisetron	0.69	0.82	100	100		100
51-Testosterone	0.95	0.98	100	100		100
52-Caffeine	0.56	0.92	100	100		100
53-Corticosterone	0.69	0.76	100	100		100
54-Ethinyl stradiol	0.97	0.98	100	~100	59	100
55-Isoxicam	0.90	0.92		100		100
56-Lornoxicam	0.99	0.97		100		100
57-Nicotine	0.98	0.99	100	100		100
58-Ondansetron	0.95	0.99	100	100	60	100
59-Piroxicam	0.94	0.95	100	100		100
60-Verapamil	0.97	0.97	100	>90	10-52.	100
61-Progesterone	0.97	0.99	91-100	91		100
62-Stavudine	-0.57	-0.93			100	100
63-Toremifene	1.00	1.00			100	100
64-Cyproterone acet.	0.97	0.97			100	100
65-Praziquantel	0.92	0.97		100		100
66-Cicaprost	0.63	-0.08		100		100
67-Aminopyrine	0.95	0.98	100			100
68-Nordazepam	0.81	0.89	99		99	99
69-Carfecillin	-0.86	-0.97	100			99
70-Prednisolone	0.48	0.59	99		70-100	99
71-Propranolol	0.76	0.97	90-100	>90	30	99
72-Viloxazine	0.69	0.88	100	~100	61-98	98
73-Warfarin	0.95	0.92	98	~100	93-98	98
74-Atropine	0.82	0.70		90		98
75-Minoxidil	0.21	0.47		95		98
76-Clofibrate	0.96	0.98	96		95-99	97
77-Trimethoprim	0.52	0.58	97		92-102	97
78-Venlafaxine	0.95	0.97	92			97
79-Antipyrine	0.96	0.98	100	~100	97	97
80-Bumetanide	-0.99	-0.99	100	100	~100	96
81-Trapidil	0.74	0.95			96	96
82-Fluconazole	0.52	0.97	95-100		>90	95
83-Sotalol	-0.80	-0.39	95	~100	90-100	95
84-Codeine	0.95	0.96	95		91	95
85-Flumazenil	0.81	0.94	95	>95	16	95
86-Ibuprophen	0.93	0.95	100			95
87-Labetalol	-0.42	-0.72	90-95	>90	33	95
88-Oxprenolol	0.52	0.89	97	90	50	95
89-Practolol	-0.58	-0.17	95	~100		95
90-Timolol	0.60	0.93	72	>90	75	95
91-Alprenolol	0.57	0.90	93-96	>93		93

Compounds	ΔP[a]	ΔP[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
92-Amrinone	0.59	0.72		93		93
93-Ketoprofen	0.94	0.92	100	~100	>92	92
94-Hydrocortisone	0.41	0.52	89-95	84-95		91
95-Betaxolol	0.57	0.93	90	90	80-89	90
96-Ketorolac	0.87	0.82	100	Well	80-100	90
97-Meloxicam	0.89	0.87	90		90	90
98-Phenytoin	0.04	-0.45	90	90	90	90
99-Amphetamine	0.94	0.99				90
100-Chloramphenicol	-0.34	0.99	90		80	90
101-Felbamate	-0.73	-0.93		90-95	102	90
102-Nizatidine	-0.42	-0.77	99		>90	90
103-Alprazolam	0.97	1.00			80-100	90
104-Tramadol	0.97	0.98			65-75	90
105-Nisoldipine	0.84	0.64				89
106-Oxazepam	0.42	0.68	97	~100	92.8	89
107-Tenidap	0.70	0.46	90		89	89
108-Dihydrocodeine	0.95	0.96			20	88
109-Felodipine	0.92	0.91	100	100	16	88
110-Nitrendipine	0.69	0.80			23	88
111-Saccharin	0.48	0.41	97	88		88
112-Mononidine	-0.53	0.27			88	87
113-Bupropion	0.34	0.47	87		87	87
114-Pindolol	0.01	0.38	92-95	>90	87	87
115-Lamivudine	-0.93	-1.00			86-88	85
116-Morphine	0.91	0.90	100	~100	20-30	85
117-Lansoprazole	0.96	0.85			85	85
118-Oxyfedrine	0.65	0.71			85	84
119-Captopril	0.45	-0.90	77	71	62	84
120-Bromazepam	0.97	0.97	84		84	84
121-Acetylsalicylic acid	0.78	0.60				82
122-Sorivudine	-0.68	-0.92	82	82	61	82
123-Methylprednisolone	0.50	0.61	82		82	82
124-Mifobate	1.00	1.00				81
125-Flecainide	-0.90	-0.96			81	81
126-Quinidine	0.93	0.97	80	81	81	81
127-Piroximone	-0.12	-0.10			81	80
128-Acebutolol	-0.61	-0.23	90	90	50	80
129-Ethambutol	-0.90	-0.96		75-80		80
130-Acetaminophen	0.71	0.74	80-100	80	68.95	80
131-Dexamethasone	0.45	0.61	92-100		80	80
132-Guanabenz	-0.52	0.90	75			80
133-Isoniazid	-0.79	-0.90				80
134-Omeprazole	0.96	0.94				80
135-Methadone	0.99	1.00			80	80
136-Fanciclovir	-0.50	-0.51			77	77
137-Metolazone	-0.94	-0.88	64	62-64		64
138-Fenoterol	-0.11	-0.64		60		60
139-Nadolol	-0.26	0.01	20-35	34	34	57
140-Atenolol	-0.55	-0.26	50-54	50	50	50
141-Sulpiride	-0.99	-0.99	36		30	44
142-Metaproterenol	0.06	0.05		44	10	44
143-Famotidine	-1.00	-1.00			37-45	28
144-Foscarnet	-0.99	-0.96	17	17(12-22)		17
145-Cidofovir	-1.00	-1.00			<5	3
146-Isradipine	0.67	0.64	92	90-95	17	92
147-Terbutaline	0.06	0.05	60-73	50-73	16	62
148-Reproterol	-0.95	-0.96		60		60
149-Lincomycin	-0.90	-0.92		20-35		28
150-Streptomycin	-1.00	-1.00		poor		1
151-Fluvastatin	0.69	0.05	100	>90	19-29	100
152-Urapidil	-0.26	-0.07			68	78
153-Propylthiouracil	-0.15	-1.00	75		76(53-88)	76
154-Recainam	-0.83	-0.87				71
155-Cycloserine	-0.47	0.19				73
156Hydrochlorothiazide	-1.00	-1.00	67-90	65-72		69(65-72)
157-Pirbuterol	-0.46	-0.51				60
158-Sumatriptan	-0.35	-0.64	55-75	>57	14	57
159-Amiloride	-0.99	-0.97				50

Compounds	ΔP[a]	ΔP[b]	%Abs.[c]	%Abs.[d]	%Bio.[e]	%Abs.[f]
160-Mannitol	-0.87	-0.91	16-26			16
161-Ganciclovir	-0.99	-1.00	3-3.8	3	3	3
162-Neomycin	-1.00	-1.00				1
163-Raffinose	-1.00	-1.00	0.3			0.3
164-Phenglutarimide	0.61	0.65	100			100
165-Bornaprine	0.96	0.98	100			100
166-D-Phe-L-Pro	0.14	0.44	100			100
167-Scopolamine	0.77	0.67	90-100			95
168-Naloxone	0.90	0.94	91			91
169-Ziprasidone	0.96	0.99	60			60
170-Guanoxan	-0.79	0.18		50		50
171-Netivudine	-0.91	-0.90		28		28
172-Gentamicin-C1	-1.00	-1.00	0	poor		poor
173-Cefadroxil	-0.89	-0.94			100	100
174-Ofloxacin	0.72	0.63			100	100
175-Pefloxacin	0.76	0.91			100	100
176-Cephalexin	-0.75	-0.73	98	100		100
177-Loracarbef	-0.79	-0.84	100	100		100
178-Glycine	0.32	0.45	100			100
179-Amoxicillin	-0.91	-0.97	94		93	93
180-Tiagabine	0.97	0.98			90	90
181-Telmisartan	0.93	0.95	90	rapid	43	90
182-Trovafloxacin	0.11	0.30	88		88	88
183-Acrivastine	0.94	0.96	88			88
184-Nicotinic acid	0.88	0.86				88
185-Levodopa	-0.05	0.09	100	80-90	86	86
186-Cefatrizine	-0.99	-0.98			75	75
187-Ampicillin	-0.81	-0.86				62
188-Vigabatrin	0.48	0.20				58
189-Tranexamic acid	0.22	-0.37	55			55
190-Eflurnithine	-0.46	-0.48				55
191-Metyldopa	0.27	0.59		41		41
192-Ceftriaxone	-1.00	-1.00	1	1		1
193-Distigminebromide	0.22	0.40			47	8
194-Ziduvudine	-0.99	-1.00	100	100	63	100
195-Ximoprofen	0.65	0.89	100		98	98
196-Clonidine	0.10	0.67	85-100	100	75-95	95
197-Viomycin	-1.00	-1.00				85
198-Ceftizoxime	-0.98	-1.00				72
199-Capreomycin	-1.00	-1.00				50
200-AAFC	-0.80	-0.77		32		32
201-Bretylium tosylate	1.00	1.00	23		23	23
Dose-limited, dose-dependent, and formulation-dependent drugs						
202-Spironolactone	0.90	0.93		>73		73
203-Etoposide	-0.33	-0.47	50		50(25-75)	50(25-75)
204-Cefetamet pivoxil	-0.93	-0.97			47	47
205-Cefuroximeaxetil	-0.97	-0.99	36		36-58	44(36-52)
206-Azithromycin	-0.87	-0.93	35-37		37	37
207-Fosinopril	0.84	0.96		36	25-29	36
208-Pravastatin	-0.60	-0.99	34	34	18	34
209-Cyclosporin	-1.00	-1.00	35		10-60	28(10-65)
210-Bromocriptine	-0.41	-0.99	28	28	6	28
211-Doxorrubicin	-0.78	-0.94	5	trace	5	12(0.7-23)
212-Cefuroxime	-0.98	-1.00				1
213-Iothalamate sodium	1.00	0.99	1.9	1.9		1.9
214-Sulphasalazine	-0.53	-0.66	12-13.			59(56-61)
215-Benazepril	-0.30	0.27	37	>37		≥37
216-Lisinopril	-0.97	-0.99	25	25	25-50	28(25-50)
217-Esalaprilat	-0.78	-0.85	9-10.	10-40.		25(10-40)
218-Anfotericina	-1.00	-1.00	5	poor		3(2-5)
219-Aztreonam	-1.00	-1.00		<1	<1	1
220-Mibefradil	0.73	0.84			37-109	69(37-100)
221-Ranitidine	-0.28	-0.91	50-61		50(39-88)	64(39-88)
222-Chlorotiazide	-0.99	-0.98	13-56			49(36-61)
223-Acyclovir	-0.94	-0.99	20-30		15-30	23(15-30)
224-Norfloxacin	-0.16	-0.51	35	30-40	~70	71
225-Methotrexate	-1.00	-1.00	20-100	100		70(53-83)
226-Gabapentin	0.36	-0.12	50	well	60A(36-64)	59(43-64)
227-Prazocin	0.52	0.72	100		44-69	86(77-95)

228-Olsalazine	-0.14	-0.83	2.3		2.3	24(17-31)
Drugs expected to have higher absorption						
229-Ciprofloxacin	-0.16	-0.52	69-100		69	≥69
230-Ribavirin	-0.94	-0.91			33	≥33
231-Pafenolol	-0.98	-0.93			28	≥29
232-Azosemide	-0.99	-0.98			10	≤10
233-Xamoterol	-0.97	-0.99			5	≥5
234-Enalapril	-0.53	-0.50	66	60-70	29-50	66(61-71)
235-Phenoxy-methyl penicilin	-0.68	-0.83	45	45(31-60)		59(49-68)
236-Gliclazide	-0.82	-0.86				≥65
237-Benzylpenicilin	-0.32	-0.38	30	15-30		≥30
238-Thiacetazone	-0.63	-0.83				≥20
239-Lovastatin	0.70	0.32	30	31		≥10
240-Cromolyn sodium	-0.44	-0.71				≥0.4
241-Erythromycin	-0.95	-0.98	35		35	≥35

ΔP = [P(High absorption group) - P(moderate-poor group)]. [a]Results of the classification of compounds obtained from Eq.1. [b]Results of the classification of compounds obtained from Eq.2. [c]The data used for QSAR studies was taken from Clark [55], Wessel [54], Palm [56], Yazdanian [24], Yee [23] and Chiou [57]. [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 [58] based on the analysis of literature.

If we take into consideration the full set of 241 compounds, both models developed in this work with the bilinear indices showed a 76.76% (185/241) and 77.59% (187/241) of the explanation of the human absorption values, for equation 1 an 2, respectively. This is a logic result if we consider the structural variability and the biological property. Notice that these values are similar to those obtained by other researchers.[8, 52]

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.[55] In addition, cefadroxil, digoxin and cepahalexin were compounds with known active transport.[60] 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 viomycin, with an appropriate intestinal absorption (Abs % = 85) and a molecular weight value of 685 $g.mol^{-1}$ (>500), similar to 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*.[61]

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

4. CONCLUSIONS

In this work, we have develop LDA 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. We developed, based on a large set of drug or drug-like molecules, two discriminant functions that permit us the classification of molecules between poorer-moderate and higher absorption, in accordance with their molecular structure.

These results demonstrated that non-stochastic and stochastic atom-based bilinear indices are a useful approach to generate adequate models for the correct classification of the intestinal permeability for structurally diverse drugs. The models are robust and stable, also acceptable efficiency and a fairly good predictability were found for an external test set. Our method achieved positive results in the comparison with other previously published approaches. Furthermore, rather satisfactory results were obtained by evaluating the capacity of prediction of human absorption for the obtained classification models. This approach could be applied to larger sets of new chemical entities synthesized via a combinatorial chemistry approach. Finally, we can say that, the present *"in silico"* method would be a valuable tool in the drug discovery process to select the molecules with the greatest chance before synthesis.

ACKNOWLEDGEMENT:

Castillo-Garit, J.A. and Marrero-Ponce, Y. thanks the program 'Estades Temporals per a Investigadors Convidats' for a fellowship to work at Valencia University in 2011. F.T. acknowledges financial support from the Spanish Minister de Ciencia e Innovación (Project No. BFU 2010-19118).

REFERENCES

1. Artursson, P.; Palm, K.; Luthman, K. *Adv Drug Deliv Rev* **2001,** *46*, 27.
2. Ren, S.; Lien, E. J. *Prog. Drug Res.* **2000,** *54*, 1.
3. Guangli, M.; Yiyu, C. *J. Pharm. Pharmaceut. Sci* **2006,** *9*, 210.
4. Fujiwara, S.; Yamashita, F.; Hashida, M. *Int. J. Pharm.* **2002,** *237*, 95.
5. Fossati, L.; Dechaume, R.; Hardillier, E.; Chevillon, D.; Prevost, C.; Bolze, S.; Maubon, N. *Int. J. Pharm.* **2008,** *360*, 148.
6. Shah, P.; Jogani, V.; Bagchi, T.; Misra, A. *Biotechnol. Prog.* **2006,** *22*, 186.
7. Fujikawa, M.; Nakao, K.; Shimizu, R.; Akamatsu, M. *Bioorg. Med. Chem.* **2007,** *15*, 3756.
8. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *Int. Electron. J. Mol. Des.* **2005,** *4* 124.
9. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Weinheim (Ger.): Wiley-VCH, **2000**.
10. Crum-Brown, A.; Fraser, T., R. *Trans. R. Soc. Edinburg.* **1868-9,** *25*, 151.
11. Marrero-Ponce, Y.; Romero, V.: Central University of Las Villas., **2002**.
12. Castillo-Garit, J. A.; Vega, M. C.; Rolon, M.; Marrero-Ponce, Y.; Kouznetsov, V.; Torres, D. F.; Gómez-Ba-

rrio, A.; Alvarez, A.; Montero, A.; Torrens, F.; Pérez-Giménez, F. *Eur. J. Pharm. Sci.* **2010,** *39*, 30.
13. Castillo-Garit, J. A.; Vega, M. C.; Rolón, M.; Marrero-Ponce, Y.; Gómez-Barrio, A.; Escario, J. A.; Bello, A. A.; Montero, A.; Torrens, F.; Pérez-Giménez, F.; Arán, V. J.; Abad, C. *Europ. J. Med. Chem.* **2011,** *46*, 3324.
14. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R.; Rodríguez-Borges, J. E. *QSAR Comb. Sci.* **2009,** *28*, 1465.
15. Castillo-Garit, J. A.; Martinez-Santiago, O.; Marrero Ponce, Y.; Casañola-Martin, G. M.; Torrens, F. *Chem. Phys. Lett.* **2008,** *464*, 107.
16. Artursson, P. *J. Pharm. Sci.* **1990,** *79*, 476.
17. Artursson, P.; Karlsson, J. *Biochem. Biophys. Res. Commun.* **1991,** *175*, 880.
18. Haeberlin, B.; Rubas, W.; Nolen, H. W., 3rd; Friend, D. R. *Pharm. Res.* **1993,** *10*, 1553.
19. Rubas, W.; Jezyk, N.; Grass, G. M. *Pharm. Res.* **1993,** *10*, 113.
20. Hovgaard, L.; Brondsted, H.; Buur, A.; Bundgaard, H. *Pharm. Res.* **1995,** *12*, 387.
21. Augustijns, P.; D'Hulst, A.; Van Daele, J.; Kinget, R. *J. Pharm. Sci.* **1996,** *85*, 577.
22. Collett, A.; Sims, E.; Walker, D.; He, Y. L.; Ayrton, J.; Rowland, M.; Warhurst, G. *Pharm. Res.* **1996,** *13*, 216.
23. Yee, S. *Pharm. Res.* **1997,** *14*, 763.
24. Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. *Pharm. Res.* **1998,** *15*, 1490.
25. Schipper, N. G.; Osterberg, T.; Wrange, U.; Westberg, C.; Sokolowski, A.; Rai, R.; Young, W.; Sjostrom, B. *Pharm. Res.* **2001,** *18*, 1735.
26. Zhu, C.; Jiang, L.; Chen, T. M.; Hwang, K. K. *Eur. J. Med. Chem.* **2002,** *37*, 399.
27. Camenisch, G.; Alsenz, J.; van de Waterbeemd, H.; Folkers, G. *Eur. J. Pharm. Sci.* **1998,** *6*, 317.
28. Chong, S.; Dando, S. A.; Morrison, R. A. *Pharm. Res.* **1997,** *14*, 1835.
29. Aungst, B. J.; Nguyen, N. H.; Bulgarelli, J. P.; Oates-Lenz, K. *Pharm. Res.* **2000,** *17*, 1175.
30. Ruiz-Garcia, A.; Lin, H.; Pla-Delfina, J. M.; Hu, M. *J. Pharm. Sci.* **2002,** *91*, 2511.
31. Wu, X.; Whitfield, L. R.; Stewart, B. H. *Pharm. Res.* **2000,** *17*, 209.
32. Liang, E.; Proudfoot, J.; Yazdanian, M. *Pharm. Res.* **2000,** *17*, 1168.
33. Lentz, K. A.; Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Polli, J. E. *Pharm. Res.* **2000,** *17*, 1456.
34. Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1585.
35. Saha, P.; Kou, J. H. *Eur J Pharm Biopharm* **2002,** *54*, 319.
36. Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. *J. Med. Chem.* **2001,** *44*, 1927.
37. Artursson, P.; Magnusson, C. *J. Pharm. Sci.* **1990,** *79*, 595.
38. Hilgendorf, C.; Spahn-Langguth, H.; Regardh, C. G.; Lipka, E.; Amidon, G. L.; Langguth, P. *J. Pharm. Sci.* **2000,** *89*, 63.
39. Gres, M. C.; Julian, B.; Bourrie, M.; Meunier, V.; Roques, C.; Berger, M.; Boulenc, X.; Berger, Y.; Fabre, G. *Pharm. Res.* **1998,** *15*, 726.
40. Artursson, P.; Palm, K.; Luthman, K. *Adv. Drug Deliv. Rev.* **1996,** *22*, 67.
41. Delie, F.; Rubas, W. *Crit. Rev. Ther. Drug. Carrier. Syst.* **1997,** *14*, 221.
42. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; Rotondo, R. *J. Mol. Graphics Model.* **2007,** *26*, 32.
43. Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Letchworth, U. K: Research Studies Press, **1986**.
44. Todeschini, R.; Gramatica, P. *Perspect. Drug Disc. Des.* **1998,** *9-11*, 355.
45. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 682.
46. Pauling, L. *The Nature of Chemical Bond*; Ithaca (New York): Cornell University Press **1939**.
47. STATISTICA version. 6.0. **2001,** *StatSoft,*, Tulsa.
48. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics* **2000,** *16*, 412.
49. Wold, S.; Erikson, L. In *Chemometric Methods in Molecular Design;* van de Waterbeemd, H., Ed.; VCH Publishers: Weinheim (Ger.), **1995**, 309.
50. Penney, K. B.; Smith, C. J.; Allen, J. C. *J. Invest. Dermatol.* **1984,** *82*, 308.
51. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR & Comb. Sci.* **2003,** *22*, 69.
52. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R. *J. Pharm. Sci.* **2008,** *97*, 1946.
53. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model.* **2002,** *20*, 269.
54. Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. *J. Chem. Inf. Comput. Sci.* **1998,** *38*, 726.
55. Clark, D. E. *J. Pharm. Sci.* **1999,** *88*, 807.
56. Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. *Pharm. Res.* **1997,** *14*, 568.
57. Chiou, W. L.; Barve, A. *Pharm. Res.* **1998,** *15*, 1792.
58. Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Butina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A.; Boutina, D. *J. Pharm. Sci.* **2001,** *90*, 749.
59. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997,** *23*, 3.
60. Benet, L. Z.; Øie, S.; Schwartz, J. B. In *Pharmacological Basis of Terapeutics;* Hardman, J. G., Limbird, L. E., Gilman, A. G., Eds.; McGraw-Hill: New York, **1996**, 1707.
61. Egan, W. J.; Merz, K. M., Jr.; Baldwin, J. J. *J. Med. Chem.* **2000,** *43*, 3867.