Bioorganic & Medicinal Chemistry 19 (2011) 2615–2624

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

QSAR-based permeability model for drug-like compounds

Rafael Gozalbes [a,*], Mary Jacewicz [b], Robert Annand [b], Katya Tsaioun [b], Antonio Pineda-Lucena [a,*]

[a] *Structural Biochemistry Laboratory, Department of Medicinal Chemistry, Centro de Investigación Príncipe Felipe (CIPF). Avda. Autopista del Saler 16, 46012 Valencia, Spain*
[b] *APREDICA, 313 Pleasant Street, Watertown, MA 02472, USA*

ARTICLE INFO

Article history:
Received 29 November 2010
Revised 4 March 2011
Accepted 5 March 2011
Available online 12 March 2011

Keywords:
Chemical space
Computational chemistry
Drug discovery
Quantitative structure–activity relationships (QSAR)
Permeability

ABSTRACT

A QSAR model was developed for predicting intestinal drug permeability, one of the most important parameters when evaluating compounds in drug discovery projects. First, a set of relevant properties for establishing a drug-like chemical space was applied to a database of compounds with Caco-2 permeability values obtained from previous studies. Several QSAR regression models were then developed from this set of drug-like structures. The best model was selected based on the accuracy of correct classifications obtained for training and validation subsets previously defined, including 17 structures from the FDA Biopharmaceutics Classification System (BCS). Further validation of the QSAR model was performed by applying it to 21 drugs for which Caco-2 permeability values were experimentally determined by us. The good agreement between predictions and experimental values in all cases confirmed the reliability of the equation. Since the model was developed with very simple descriptors, easy to calculate, its applicability to large collections of in silico chemicals is guaranteed.

1. Introduction

Choices about the adequacy of candidate compounds usually rely on the early screening of ADME-T properties.[1] Cellular permeability of chemicals is one of the key physicochemical parameters to be considered when selecting compounds. To be absorbed, a drug has to diffuse across a series of separate compartments and, amongst them, the single layer of epithelial cells is the most significant barrier to drug absorption. Almost all drugs that are given orally are absorbed across the intestinal mucosa by passive diffusion.[2]

In vitro assays based on Caco-2 cell monolayer models have been used since the early 1990s to predict the absorption rate of candidate drug compounds across the intestinal epithelial cell barrier. The Caco-2 cell line is an immortalized line of heterogeneous human epithelial colorectal adenocarcinoma cells. When Caco-2 cells are cultured under specific conditions, they form confluent monolayers that resemble morphologically and functionally the intestine enterocytes.[3] Furthermore, in vitro apparent permeability (P_{app}) on Caco-2 cells and in vivo absorbed fraction are well correlated.[4] This fact, linked to the significant advances in automation technology, has favoured the application of these assays in the pharmaceutical industry to measure intestinal absorption. However, this assay has two main drawbacks: (i) drug absorption rates are determined several days after Caco-2 cell seeding to allow for monolayer formation and cell differentiation, that resulting in a very high cost, and (ii) over the years, cells with different characteristics have been employed by different laboratories, making it difficult to compare results.[5] Parallel artificial membrane permeability assays (PAMPA) are also a useful tool for the prediction of oral absorption and for the development of structure–permeability relationships. Measurement of PAMPA is much more rapid and cost effective than Caco-2 assays, but it has other disadvantages: it is not recommended to use PAMPA to predict absorption involving transporter-mediated processes, and very careful data interpretation is needed.[6] In this context, alternative methods such as in silico predictive models could have a tremendous impact.

The most widely used computational approach for the coarse estimation of passive intestinal absorption is the Lipinski's 'rule of five'.[7] It states that poor absorption or permeation is more probable for chemical compounds with molecular weight >500 Da, $c\,\mathrm{Log}\,P$ >5, number of hydrogen bond donors >5 and number of hydrogen bond acceptors >10. The main advantage of the rule of five is its simplicity and rapidity to compute, although several authors have pointed out significant limitations in its application to permeability evaluation,[8,9] thereby prompting the need for more elaborated models. Several QSAR approaches have been proposed to estimate Caco-2 cell permeability but (i) most of the models reported so far have been based on small sets of molecules,[10,11] (ii) the experimental Caco-2 permeability data have been highly influenced by inter-laboratory differences,[5] and/or (iii) the models have not been fully validated by external test sets to guarantee their predictive performance.[11] QSAR equations for predicting PAMPA permeability coefficients on structurally diverse compounds have

[*] Corresponding authors. Tel.: +34 96 328 96 80; fax: +34 96 328 97 01.
E-mail addresses: rgozalbes@cipf.es (R. Gozalbes), apineda@cipf.es (A. Pineda-Lucena).

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.03.011

also been published.[12] Some reviews are available that summarize the different computational strategies proposed until now, their comparative advantages and disadvantages, and the need for more robust and reliable predictors.[10,11,13]

Our objective in the present study has been the development of a computational model able to predict the permeability of chemicals, especially drug-like compounds. The model was intended to be used, in the early phases of drug discovery projects, as a decision tool to discard those compounds presenting permeability problems. To accomplish this objective we have developed a QSAR model based on publicly available information and experimental Caco-2 cell permeability data. Trying to overcome the limitations of previously published predictors, we have compiled permeability data from an extensive number of drugs, one of the biggest to the best of our knowledge. The model was validated with compounds characterized by a high structural and therapeutic diversity, including an external data set composed by 21 drugs for which Caco-2 permeability data were experimentally determined by us, perhaps representing one of the most homogeneous validation sets reported so far.

2. Results and discussion

2.1. QSAR model

Following the protocol described in Section 4, permeability values corresponding to 138 drug-like structures were obtained from the scientific literature for the development of the QSAR model. To the best of our knowledge, this database represents one of the largest collections of data used so far for the development of Caco-2 permeability predictive models,[11] thereby circumventing previous pitfalls associated to models based on small datasets. Furthermore, the selection of compounds was based on the satisfaction of drug-like features[14], thus guaranteeing its applicability to other drug-like structures. The database consisted of a training and a validation set (Tset and Vset, respectively) (Tables 1 and 2). Some of

Table 1
List of compounds included in the Tset and their permeability values (Log P_{app})[a]

Drug	Log P_{app}^{EXP}	Log P_{app}^{PRED}	Drug	Log P_{app}^{EXP}	Log P_{app}^{PRED}
Gabapentin	**−7.58**	**−6.68**	Hydrocortisone	−4.68	−4.83
Olsalazine	**−7.38**	**−6.92**	Clonidine	−4.66	*−5.06*
Ouabain	**−6.96**	**−6.91**	Danazol	−4.66	−4.46
Ceftriaxone	**−6.89**	**−6.68**	Felodipine	−4.64	−4.65
Argipressin	**−6.85**	**−6.71**	Desipramine	−4.64	−3.92
Penicillin V	**−6.77**	**−6.26**	Ibuproxam	−4.63	−4.43
Lisinopril	**−6.66**	**−6.78**	Phencyclidine	−4.61	−4.83
Loracarbef	**−6.62**	**−6.49**	Methylprednisolone	−4.60	−4.72
Sulfasalazine	**−6.61**	**−6.12**	Nicotine	−4.58	−4.56
Cephalexin	**−6.57**	**−6.51**	Naloxone	−4.57	−4.70
Lactulose	**−6.52**	**−6.90**	Imipramine	−4.56	−4.02
Cefuroxime	**−6.42**	**−6.99**	Labetalol	−4.53	*−5.36*
Sulpiride	**−6.40**	*−5.73*	Indomethacin	−4.53	−4.37
Pirenzepine	**−6.36**	*−5.92*	Fluconazole	−4.53	−4.84
Ganciclovir	**−6.32**	**−6.18**	Nevirapine	−4.52	−4.65
Proscillaridin	**−6.20**	*−5.66*	Trovafloxacin	−4.52	−4.83
Doxorubicin	**−6.15**	*−5.96*	Cortisone	−4.51	−4.62
Acyclovir	**−6.15**	−4.64	Flurbiprofen	−4.47	−4.43
Cefatrizine	**−6.12**	**−6.22**	Taurocholic acid	−4.46	*−5.33*
Methylscopolamine	**−6.12**	*−5.87*	Piroxicam	−4.45	−4.65
Nadolol	**−6.11**	*−5.41*	Aminopyrine	−4.44	−4.59
Benzylpenicillin	**−6.11**	**−6.36**	Griseofulvin	−4.44	−4.02
Tranexamic acid	**−6.04**	**−6.03**	Timolol	−4.43	*−5.07*
Epinephrine	**−6.02**	*−5.40*	Pindolol	−4.41	−4.76
Acebutolol	*−5.90*	−4.74	Aceclofenac	−4.41	−4.55
L-Dopa	*−5.82*	*−5.99*	Tiaprofenic acid	−4.41	−4.07
Ciprofloxacin	*−5.77*	*−5.53*	Phenytoin	−4.40	−4.26
Acrivastine	*−5.72*	*−5.85*	Zaltoprofen	−4.40	−4.15
Practolol	*−5.72*	*−5.18*	Pranoprofen	−4.39	−4.74
Fexofenadine	*−5.70*	*−5.27*	Indobufen	−4.39	−4.72
Sumatriptan	*−5.68*	*−5.20*	Testosterone	−4.37	−4.54
Sotalol	*−5.38*	*−5.65*	Warfarin	−4.36	−4.47
Ketorolac	*−5.08*	*−5.12*	Loxoprofen	−4.35	−4.97
Dopamine	*−5.03*	*−5.21*	Diltiazem	−4.34	−4.58
Olopatadine	*−5.01*	−4.69	Sildenafil	−4.32	*−5.30*
Metolazone	−5.00	−4.33	Tenidap	−4.29	−4.73
Morphine	−5.00	−4.93	Ibuprofen	−4.28	−4.47
Scopolamine	−4.93	*−5.15*	Diclofenac	−4.27	−4.53
Chloramphenicol	−4.90	−4.93	Diazepam	−4.22	−4.28
Digoxin	−4.89	*−5.06*	Oxprenolol	−4.21	−4.73
Mibefradil	−4.87	−4.59	Lidocaine	−4.21	−4.67
Zidovudine	−4.84	*−5.09*	Alprenolol	−4.20	−4.51
Telmisartan	−4.82	−4.72	Alfentanil	−4.08	−3.98
Fleroxacin	−4.81	*−5.88*	Guanabenz	−−4.07	−4.01
Estradiol	−4.77	−4.63	Trimethoprim	−4.06	−4.77
Nitrendipine	−4.77	−4.60	Nordazepam	−4.02	−4.30
Dexamethasone	−4.74	−4.48	Acetaminophen	−4.00	−4.66
Guanoxan	−4.71	*−5.16*	Bupropion	−3.82	−4.28
Chlorpromazine	−4.70	−4.15			

[A] Bold values: poor permeability compounds; italicized values: moderate permeability compounds; other values: high permeability compounds

Table 2
Predicted versus experimental Log P_{app} values for the Vset ranked by predicted values[a]

ID	Log P_{app}^{PRED}	Log P_{app}^{EXP}	Source
Amoxicillin	**−6.78**	**−6.73**	BCS (Low)
Desmopressin	**−6.49**	**−6.89**	Randomly chosen
Mannitol	**−6.47**	**−6.35**	BCS (Low)
Terbutaline	**−6.33**	**−6.38**	Randomly chosen
Saquinavir	**−6.10**	**−6.26**	Randomly chosen
Ranitidine	*−5.84*	**−6.20**	BCS (Low)
Methyldopa	*−5.81*	**−6.62**	BCS (Low)
Remikiren	*−5.77*	**−6.13**	Randomly chosen
Methotrexate	*−5.74*	*−5.92*	Randomly chosen
Tiotidine	*−5.65*	*−5.66*	Randomly chosen
Alminoprofen	*−5.53*	−4.34	Randomly chosen
Furosemide	*−5.35*	**−6.55**	BCS (Low)
Sulindac	*−5.24*	−4.66	Randomly chosen
Atenolol	*−5.14*	*−5.92*	BCS (Low)
Acetylsalicylic acid	*−5.14*	*−5.35*	Randomly chosen
Propylthiouracil	*−5.12*	−4.02	Randomly chosen
Fenoprofen	*−5.11*	−4.38	Randomly chosen
Netivudine	*−5.09*	**−6.17**	Randomly chosen
Tiacrilast	−4.96	−4.90	Randomly chosen
Metoprolol	−4.94	−4.53	BCS (High)
Cimetidine	−4.90	*−5.88*	Randomly chosen
Ziprasidone	−4.85	−4.91	Randomly chosen
Corticosterone	−4.81	−4.33	Randomly chosen
Bremazocine	-4.76	−4.73	Randomly chosen
Verapamil	−4.75	−4.34	BCS (High)
Naproxen	−4.73	−4.33	BCS (High)
Antipyrine	−4.72	−4.78	BCS (High)
Meloxicam	−4.71	−4.71	Randomly chosen
Hydrochlorothiazide	−4.67	*−5.39*	BCS (Low)
Oxazepam	−4.62	−4.04	Randomly chosen
Ketoprofen	−4.61	−4.33	BCS (High)
Ondansetron	−4.56	−3.96	Randomly chosen
Propranolol	−4.49	−4.40	BCS (High)
Progesterone	−4.47	−4.55	Randomly chosen
Bosentan	−4.37	*−5.98*	Randomly chosen
Amfenac	−4.29	−4.52	Randomly chosen
Fluvastatin	−4.22	−4.33	BCS (High)
Caffeine	−4.02	−4.47	BCS (High)
Carbamazepine	−4.00	*−5.30*	BCS (High)
Fluparoxan	−3.76	−3.70	Randomly chosen
Theophylline	−2.83	−4.40	BCS (High)

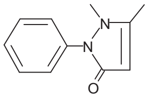
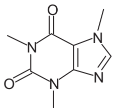
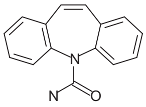
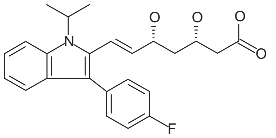
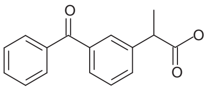
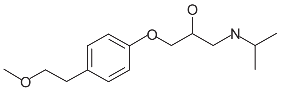
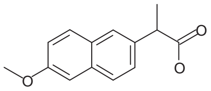
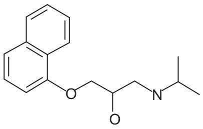
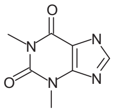
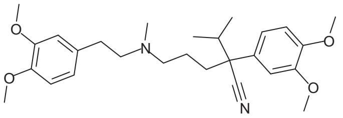
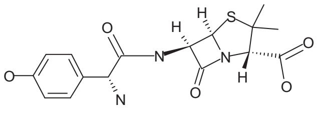
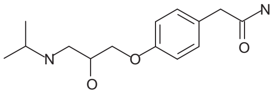
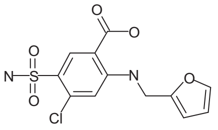
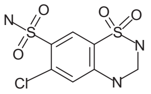
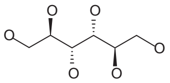
[A] Bold values: poor permeability compounds; italicized values: moderate permeability compounds; other values: high permeability compounds.

the structures included in the Vset correspond to 'model drugs' (BioPharmaceutics Classification System, BCS) for which a classification as having high or low permeability has been provided by the US Food & Drug Administration (FDA) (Table 3). The other compounds included in the Vset were randomly chosen from the initial dataset (Table 2). The distribution of compounds in the Tset and the Vset pursued a homogeneous distribution of permeability Caco-2 values and therapeutic activities in the two sets (Figs. 1 and 2).

Several multilinear regression QSAR models were developed based on the 97 compounds from the Tset and their corresponding Caco-2 permeability values, expressed as the logarithm of the apparent permeability coefficients in cm/s, Log P_{app}. Different combinations of descriptors were used to build the models and the performance of the different equations was evaluated by comparing their ability to properly classify the Log P_{app} of compounds from the Tset and Vset (41 structures).

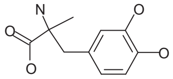
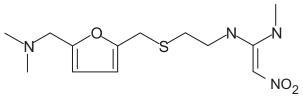
The best model turned out to be one developed with 14 descriptors belonging to different categories (Table 4): MACCS keys (9 descriptors), properties-based (1 descriptor), and 2D-based (2 Autocorrelation indices and 2 Topological indices). No conformational-dependent descriptors were selected since they did not

Table 3
'Model drugs' from the FDA-BCS Guidance included in the Vset

Drug	Structures	Permeability class
Antipyrine		High
Caffeine		High
Carbamazepine		High
Fluvastatin		High
Ketoprofen		High
Metoprolol		High
Naproxen		High
Propranolol		High
Theophylline		High
Verapamil		High
Amoxicillin		Low
Atenolol		Low
Furosemide		Low
Hydrochlorothiazide		Low
Mannitol		Low

(continued on next page)

Table 3 (*continued*)

Drug	Structures	Permeability class
α-Methyldopa		Low
Ranitidine		Low

improve the performance of the predictions. The simplicity of the selected descriptors represents an important advantage in the application of the model to large databases. Statistically, the QSAR predictor is comparable to the best models published so far and has been developed using a significant number of drugs ($n = 97$, $r^2 = 0.77$, LOO $q^2 = 0.69$, $S = 0.48$, Fisher-statistic = 19.29, p-value <0.001).[10,11,15]

Figure 3a shows the good agreement between experimental and calculated Log P_{app} values for the Tset. Furthermore, it was observed that most of the compounds exhibited very small variations between the experimental and the predicted Log P_{app} values (Fig. 3b): 96.9% of the Tset compounds (94/97) less than one Log unit and 76.3% (74/97) less than 0.5 Log units. Based on the criteria proposed by Chaturveldi et al.,[16] compounds were classified as being poorly (0–20%), moderately (20–70%) or well (70–100%) absorbed drugs according to their apparent permeability coefficients: $<1 \times 10^{-6}$ cm/s, $1–10 \times 10^{-6}$ cm/s, $>10 \times 10^{-6}$ cm/s, respectively. Table 5a summarizes the results when comparing experimental and predicted Log P_{app} values for the Tset.

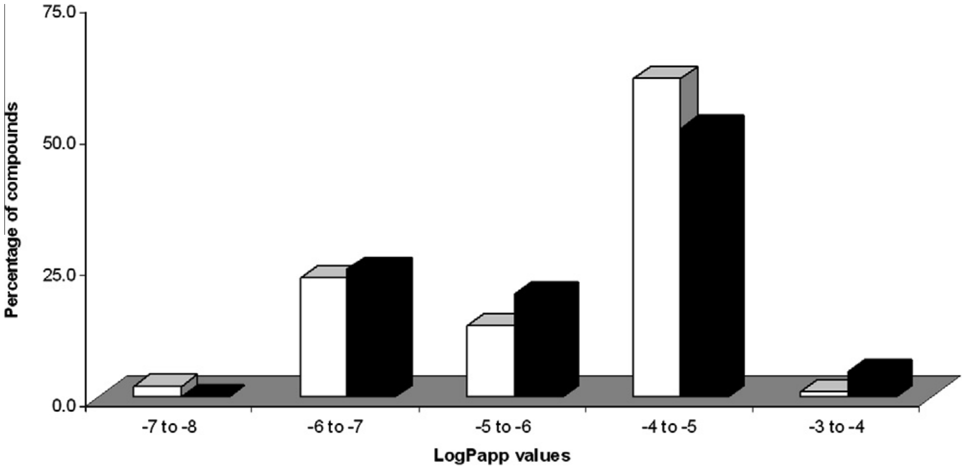


Figure 1. Percentage distribution of the Log P_{app} values in the Tset (white bars) and Vset (black bars).

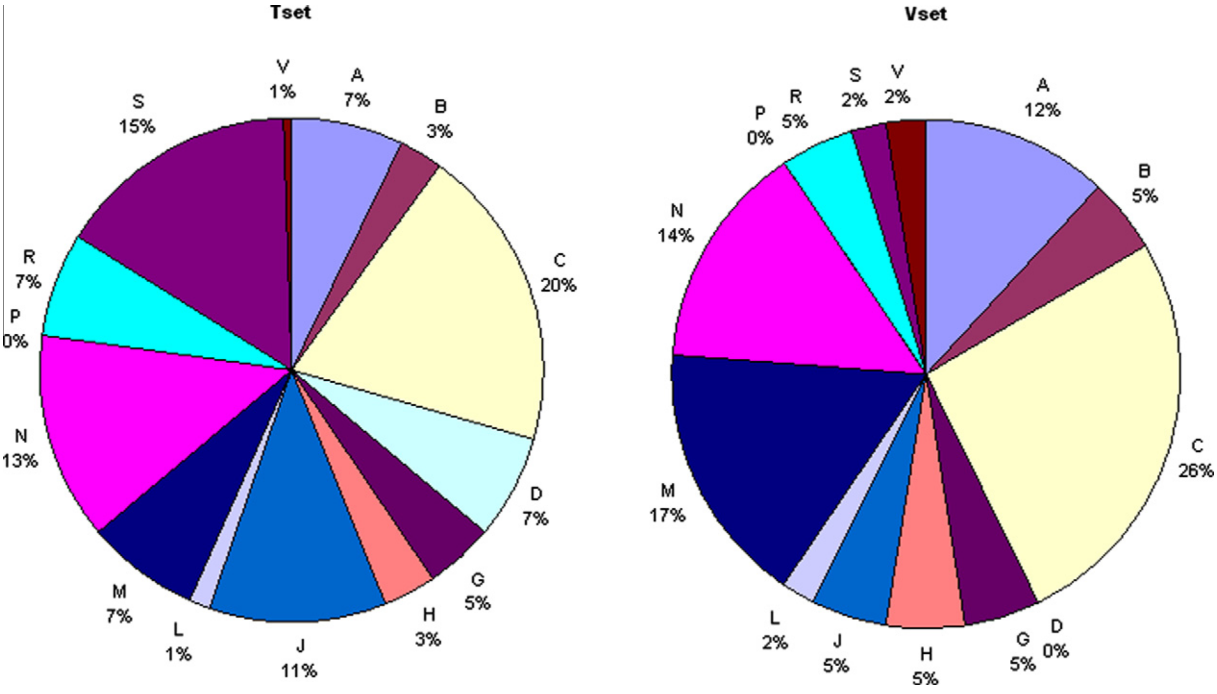


Figure 2. Distribution of therapeutic activities in the Tset and the Vset. Codes for the ATC system classification of drugs: A: alimentary tract and metabolism; B: blood and blood forming organs; C: cardiovascular system; D: dermatologicals; G: genito-urinary system and sex hormones; H: systemic hormonal preparations, excluding sex hormones and insulins; J: antiinfectives for systemic use; L: antineoplastic and immunomodulating agents; M: musculo-skeletal system; N: nervous system; P: antiparasitic products, insecticides and repellents; R: respiratory system; S: sensory organs; V: various.

Table 4
Parameters (descriptors and coefficients) defining the best QSAR predictive model

Abbreviation	Descriptor definition	Coefficients
MACCS8	Heteroatoms in four-membered rings	−1.25
MACCS36	S atoms in rings	0.31
MACCS50	C in C=C bonded to >=3C	−0.27
MACCS70	N bonded to two non-C heavy atoms	0.46
MACCS100	N attached to CH_2	−0.23
MACCS119	N in double bonds	−0.28
MACCS129	CH_2s separated by 3 bonds	−0.11
MACCS132	O 2 bonds from CH_2	0.21
MACCS139	OH groups	−0.28
M Log *P*	Moriguchi octanol–water partition coefficient (Log *P*)	0.19
GATS8e	Geary autocorrelation–lag 8/weighted by atomic Sanderson electronegativities	0.11
GATS5p	Geary autocorrelation—lag 5/weighted by atomic polarizabilities	1.02
JGI1	Mean topological charge index of order 1	−4.23
JGI2	Mean topological charge index of order 2	7.43
Constant	—	−7.13

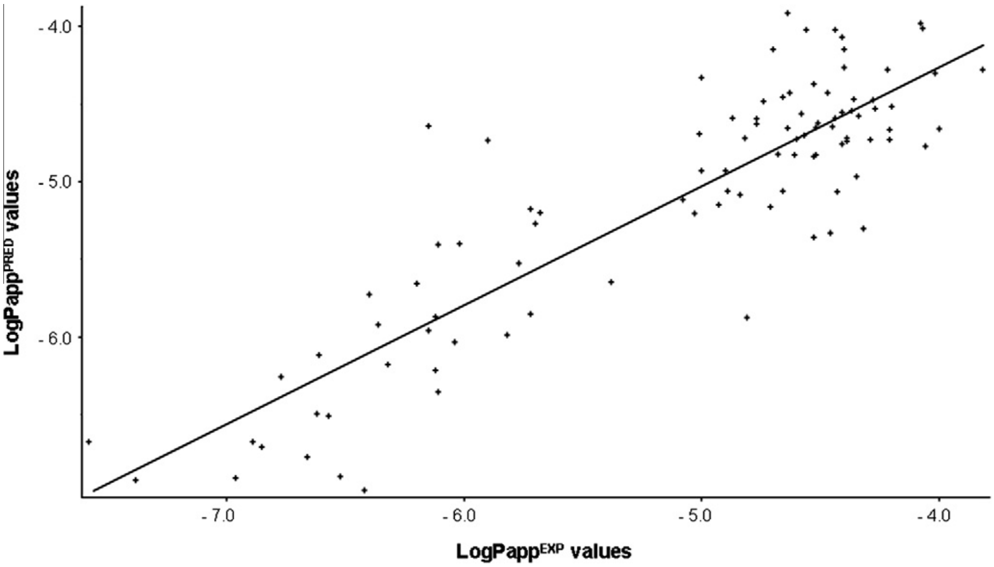


Figure 3a. Scatter plot obtained from the comparison of experimental (Log P_{app}^{EXP}) and calculated (Log P_{app}^{PRED}) permeability values for the Tset.

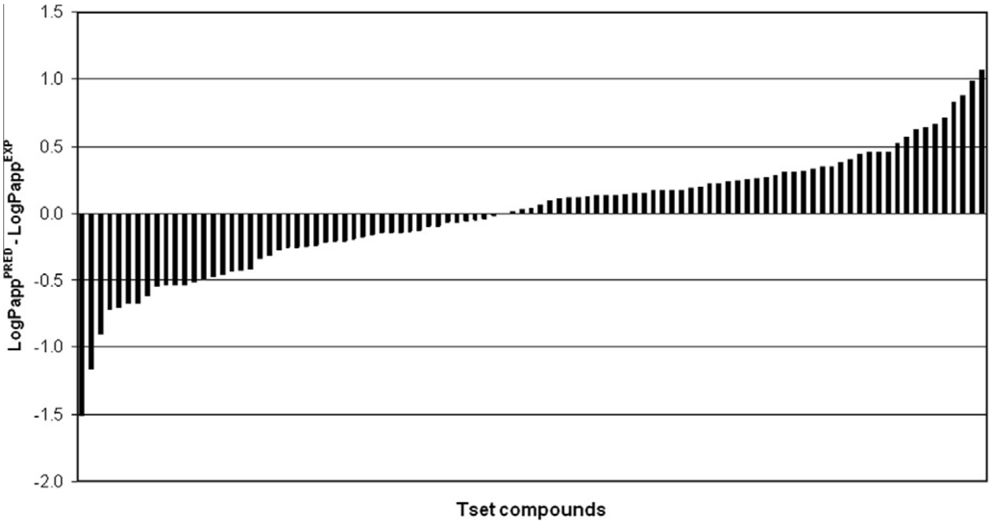


Figure 3b. Variations observed between the predicted (Log P_{app}^{PRED}) and the experimental (Log P_{app}^{EXP}) permeability values for the Tset.

The descriptors selected by the model are in good agreement with our actual understanding on permeability mechanisms. For example, the M Log *P* descriptor was found to have a positive influence in the permeability model, which is in good agreement with the fact that octanol/water partitioning is the most frequent physicochemical parameter used for the prediction of cellular

Table 5a
Confusion matrix summarizing the number and percentages of Tset compounds correctly classified by the QSAR model

		$\text{Log}\,P_{\text{app}}^{EXP}$ values			Recall
		<−6	>−6 and <−5	>−5	
$\text{Log}\,P_{\text{app}}^{PRED}$	<−6	16	0	0	100.0%
	>−6 and <−5	7	9	10	34.6%
	>−5	1	2	52	94.5%
	Precision	66.7%	81.8%	83.9%	79.4%

Table 5b
Confusion matrix summarizing the number and percentages of Vset compounds correctly classified by the QSAR model

		$\text{Log}\,P_{\text{app}}^{EXP}$ values			Recall
		<−6	>−6 and <−5	>−5	
$\text{Log}\,P_{\text{app}}^{PRED}$	<−6	5	0	0	100.0%
	>−6 and <−5	5	4	4	30.8%
	>−5	0	4	19	82.6%
	Precision	50.0%	50.0%	82.6%	68.3%

permeability.[17] Moreover, it is well known that highly hydrophilic compounds are unable for any distribution in the membrane.[17] Accordingly, it was found that the MACCS139 descriptor (i.e., number of –OH groups in a molecule) has a negative contribution to the QSAR permeability model. The influence of charges on permeability could also been identified by the presence of JGI1 and JGI2, mean topological charge indices of order 1 and 2 respectively.[18] Finally, a descriptor related to the size of the molecules, MACCS8 (i.e., number of heteroatoms in four-membered rings) was found to significantly influence drug permeability, thus reflecting the inverse relationship between number of heteroatoms and permeability.

2.2. Internal validation set

Similar results to those obtained with the Tset were observed when the model was applied to the Vset (Table 2): 83% (34/41) of the compounds exhibited differences between predicted and experimental values smaller than one Log unit, and 66% (27/41) of them less than 0.5 Log units. Extremely precise predictions were obtained for 27% (11/41) of the structures with differences under 0.1 Log units. A confusion matrix summarizing these results is shown in Table 5b.

With regard to the predictions on the FDA-BCS structures (Tables 2 and 3), it was found that 94% of the compounds (16/17) were correctly assigned to the 'low' or 'high' permeability groups considering the criteria suggested by Chaturveldi et al.,[16] with the exception of hydrochlorothiazide. In fact, although this diuretic drug is low-permeable, it is still the most permeable one on its own category based on the scientific literature[17,19–21] (experimental average Log $P_{\text{app}} = -5.39$).

2.3. External validation set

A good agreement was also found for the external validation set composed by 21 drugs (Table 6) whose experimental permeability values had been determined by us. Thus, for 57% (12/21) of the drugs, the differences between predicted and experimental values

Table 6
Predicted and experimental Log P_{app} values for the 21 drug-like compounds included in the external set[a]

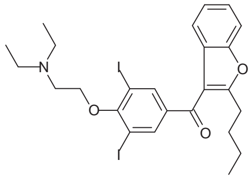
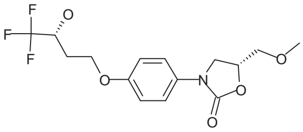
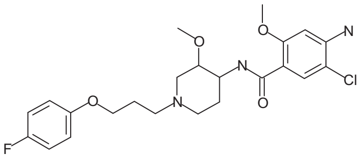
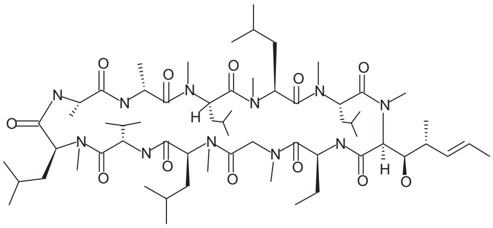
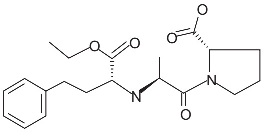
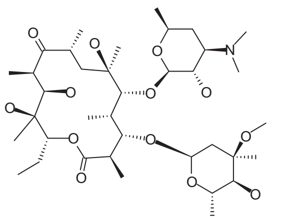
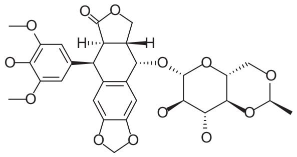
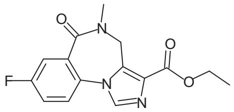
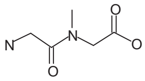
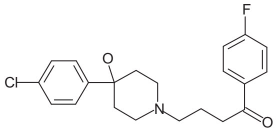
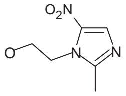
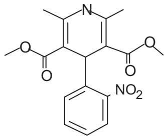
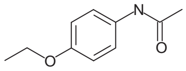
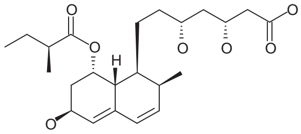
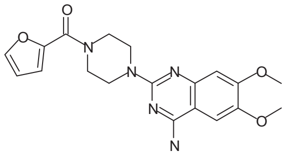
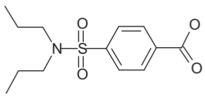
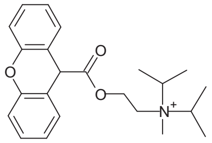
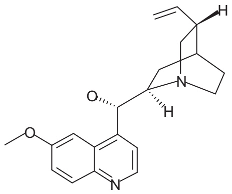
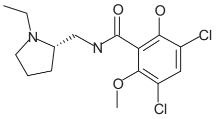
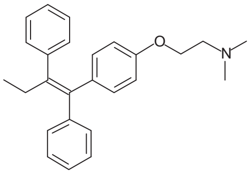
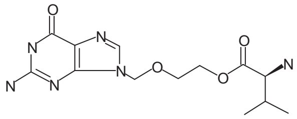
ID	Structures	Therapeutical activity	$\text{Log}\,P_{\text{app}}^{EXP}AB$	$\text{Log}\,P_{\text{app}}^{PRED}$
Amiodarone		Antiarrhythmic	−4.16 (−4.16)[b]	−3.84
Befloxatone		Antidepressant	−4.55 (−4.72)	−4.81
Cisapride		Gastroprokinetic	*−5.15* (−4.17)	*−5.00*
Cyclosporin A		Immunosupressant	**<−6.00** (<−6.00)	−**6.09**
Enalapril		Antihypertensive	−**6.10** (−5.70)	*−5.31*

Table 6 (*continued*)

ID	Structures	Therapeutical activity	Log $P_{app}^{EXP}AB$	Log P_{app}^{PRED}
Erythromycin		Antibacterial	*−5.49* (−5.15)	−4.54
Etoposide		Antineoplastic	*−6.00* (−5.30)	*−5.30*
Flumazenil		Benzodiazepine antagonist	−4.53 (−4.82)	−4.67
GlySar		(substrate for the transporter hPEPT1)	**<−6.00** (<−6.00)	−**6.67**
Haloperidol		Antipsychotic	−4.80 (−4.61)	−4.79
Metronidazole		Antiprotozoal	−4.41 (−4.62)	−4.86
Nifedipine		Antianginal	−4.49 (−4.61)	−4.42
Phenacetin		Analgesic	−4.59 (−4.49)	−4.25
Pravastatin		Antilipemic	−**6.56** (−5.64)	*−5.18*
Prazosin		Antihypertensive	*−5.39* (−4.88)	*−5.76*
Probenecid		Uricosuric	−4.78 (−4.84)	−4.94
Propantheline		Antispasmodic	*−5.92* (−5.18)	*−5.01*
Quinidine		Antiarrhythmic	−4.73 (−4.25)	*−5.32*

(*continued on next page*)

Table 6 (*continued*)

ID	Structures	Therapeutical activity	Log $P_{app}^{EXP}AB$	Log P_{app}^{PRED}
Raclopride		Antipsychotic	−4.61 (−4.85)	*−5.36*
Tamoxifen		Antineoplastic	<−**6.00** (−6.30)	−4.37
Valacyclovir		Antiviral	−**6.52** (−6.30)	*−5.04*

[a] Bold values: poor permeability compounds; italicized values: moderate permeability compounds; other values: high permeability compounds.

[b] In parenthesis: Log $P_{app}BA^{EXP}$.

Table 7
Confusion matrix summarizing the number and percentages of compounds included in the external set correctly classified by the QSAR model

		Log P_{app}^{EXP} values			Recall
		<−6	>−6 and <−5	>−5	
Log P_{app}^{PRED}	<−6	2	0	0	100.0%
	>−6 and <−5	3	4	2	44.4%
	>−5	1	1	8	80.0%
	Precision	33.3%	80.0%	80.0%	66.7%

were smaller than 0.5 Log units, and for 86% (18/21) of the compounds that difference was smaller than 1.0 Log unit. A tendency to overestimate the permeability values was observed in some cases, particularly for those compounds with small experimental values. The explanation for very insoluble compounds, like tamoxifen and erythromycin, could be that their permeability is affected by their poor solubility in real experimental situations. Other compounds, like enalapril and valacyclovir, are ester prodrugs that exhibit two peaks in the HPLC chromatograms, perhaps due to transesterification reactions, and this might affect the measured permeability. Furthermore, deesterification by Caco-2 cells could result in underestimation of the amount of transported material that leading to low experimentally determined P_{app}. On the contrary, the predicted permeability of quinidine was smaller than the measured one, perhaps reflecting the fact that this drug is a P-gp substrate subjected to active efflux.

Table 7 summarizes the results obtained for this external validation set based on the same criteria[16] previously described. Interestingly, similar percentages of correct predictions to that obtained for the Tset and Vset are obtained, especially for low and high permeable compounds (2/2 and 8/10, respectively). It is worth noting that the percentage of experimental low permeable compounds predicted as being highly permeable or vice versa is very low. In fact, there is only one compound (1/21, 4.8% of the total set), tamoxifen, that was predicted to be highly permeable despite the fact of being a low permeable drug.

3. Conclusions

Computational approaches for the prediction of cell permeability offer two significant advantages over the experimental determination of this parameter. First, these methods do not require the chemical synthesis of compounds and secondly, they allow the analysis, in a high-throughput manner, of thousands of compounds in a very short time. Moreover, it has been demonstrated in previous comparative studies than computationally derived mathematical models can predict with a reasonable accuracy human passive intestinal absorption.[22]

This paper describes a QSAR permeability model based on freely available information obtained from scientific literature. The model was developed based on information obtained from drug-like compounds, and thus it is expected that it will be applicable to the prediction of passive permeability of other drug-like structures. We considered this to be a critical step, as inadequate applicability domains of QSAR models is a problem often faced by QSAR practitioners.[23,24] A previously described set of filters (physicochemical parameters and cut-off values)[14] was applied to define the drug-likeness of chemicals included in this study.

The QSAR model is robust enough to correctly identify permeable drug-like compounds in relatively large chemical databases. Furthermore, the model does not require any experimentally determined value for the calculation of the permeability, and it is characterized by its simplicity and rapidity of calculation since it relies on 1D and 2D descriptors easy to compute.

Despite the fact that the QSAR model was based on information obtained from different bibliographic sources, only permeability data matching homogeneity criteria were retained for its development. The model exhibited an excellent performance when applied to internal and external validation sets, including drugs from the FDA-BCS Guidance that were used as reference compounds. The predictions obtained for 21 drugs whose permeability was experimentally determined by us were particularly interesting. These results showed that the inter-laboratory differences inherent to the data used to develop the model were not an obstacle when trying to predict permeabilities on a homogeneous dataset.

Overall, the results obtained for the Tset, Vset and the external validation sets confirmed the reliability of this QSAR permeability predictor as an efficient tool for selecting permeable compounds in the early stages of drug screening campaigns relying on high-throughput virtual screening.

4. Experimental section

4.1. Data set

The QSAR permeability model was based on drug absorption potentials, measured using Caco-2 cell assays, obtained from the scientific literature. The cell permeability values were expressed as the logarithm of the apparent permeability coefficients in cm/s (Log P_{app}). Trying to reduce inter-laboratory differences, a restrictive strategy, consisting in a systematic comparison of all the permeability values reported for a particular compound with respect to the calculated average value, was followed. Bibliographic references including at least one compound with a Log P_{app} differing in more than 1 Log unit (absolute value) from the average were discarded. Using this approach, nine bibliographic references[4,17,19–21,25–28] were finally selected for the development of the QSAR model.

4.2. Selection of drug-like compounds

Given the fact that the model was intended to be applied to drug-like compounds, and considering the importance of working in a well defined chemical space,[23,24] the compounds that could not be considered drug-like structures were discarded. The selection of compounds was thus based on the application of a set of previously described drug-like parameters (chemical features and cut-off values that characterize drug-like compounds).[14] All the structures were identified using different sets of properties (i.e., physicochemical values, atom and bond counts, adjacency and distance matrix descriptors) calculated with the VIDA module (version 3.0.0) from OpenEye Scientific Software (http://www.eyesopen.com) and the QuaSAR-descriptor tool from Molecular Operating Environment (MOE, http://www.chemcomp.com; version 2008.10). Finally, a set of 138 chemicals satisfying these filters was retained for the study (Tables 1 and 2).

4.3. Characterization of the compounds by chemical descriptors

For the development of the QSAR model, all the compounds were further characterized using several subsets of numerical identifiers implemented in the SArchitect™ Designer program (Strand Life Sciences, http://www.strandls.com/sarchitect; version 2.5.0). The identifiers belonged to three main groups: constitutional descriptors—1D (i.e., counts and property-based parameters), 2D descriptors reflecting the connection of atoms in the molecule, and 3D descriptors containing three-dimensional information about each molecule. The standard list of 166 MACCS keys (http://www.symyx.com) reporting the presence/absence of specific atoms or chemical groups was also calculated. At the end of this process, each compound was characterized by more than 1200 descriptors.

4.4. Distribution of compounds in training and validation sets

The development of the QSAR models was preceded by the distribution of the initial database (138 structures) in a training set (Tset) containing 97 structures and a validation set (Vset, 41 structures), thus representing a good balance between the two groups (70% and 30% of the full set, respectively). Seventeen out of the 41 compounds included in the Vset were 'model drugs' from the FDA-Biopharmaceutics classification system (BCS) guidance (Table 3) (www.fda.gov/cder/guidance/index.htm). The BCS is a scientific framework for qualitatively classifying drug substances based on their aqueous solubility and intestinal permeability. Model drugs included in this classification are very useful for validation purposes as their permeability is based on information provided by the FDA. In that sense, they are ideal for establishing the suitability of permeability methods.

The other 24 compounds of the Vset were randomly chosen from the initial database. The distribution of permeability Caco-2 values (Fig. 1) was quite similar in both sets. Furthermore, a similar distribution of therapeutic activities, as defined by the Anatomical Therapeutic Chemical classification system (ATC; http://www.whocc.no/), was also maintained for the two sets (Fig. 2).

4.5. QSAR development

Several QSAR models were obtained using different combinations of descriptors (1D, 2D, 3D, MACCS keys) and the Multiple Linear Regression tool implemented in SArchitect™. Previous to the development of the models the list of descriptors was further refined by discarding those with a low variance (standard deviation $\leqslant 0.1$), as well as those highly correlated among them (correlation coefficient $\geqslant 0.9$). To ensure that the descriptors selected by each model were meaningful and that the models were not obtained by chance, validation of the models was obtained using the Y-scrambling or Response Permutation Testing method.[29] The best model was selected based on the accuracy of correct classifications obtained for both set of compounds, the Tset and the Vset.

4.6. Cell culture

The evaluation of the Caco-2 permeability of the 21 drugs from the external set was performed in Apredica laboratories (www.apredica.com). Caco-2 cells were obtained from ATCC (Manassas, VA) and maintained in Dulbecco's Modifed Eagle Medium (HyClone, Logan, UT) supplemented with 10% fetal bovine serum (Sigma–Aldrich Corp, St Louis, MO) and 1% penicillin/streptomycin mixture (HyClone, Logan, UT). For transport experiments, cells were trypsinized and 24,500 cells/well were grown on fibrillar collagen coated PET membrane inserts in 24-well plates (BD Biosciences, San Jose, CA). The cells were grown at 37 °C and 5% CO_2 for 21 days, fed at 2–3 day intervals, before being used in transport experiments.

4.7. Transport experiments

50 mM stock solutions in DMSO of the compounds were diluted to 10 μM solutions in transport buffer (Hank's Buffered Salt Solution, HyClone, Logan, UT) containing 25 mM HEPES (HyClone, Logan, Utah) and 0.4% glucose (TEKnova, Hollister, CA). Apical side (A side) buffer was transport buffer adjusted to pH 6.5 and contained 100 μM Lucifer Yellow (Sigma–Aldrich Corp, St. Louis, MO); basolateral (B side) buffer was transport buffer adjusted to pH 7.4. The final concentration of DMSO in the test solutions was 1%. The cells were washed, test solutions were added in triplicate to the inserts (A side) or to the wells (B side), with the corresponding buffer on the other side. The plates were incubated for 2 h at 37 °C with shaking (100 rpm), then the two parts of the plate were separated and the test solutions were added to 96-well plates (Agilent Technologies, Santa Clara, CA) containing 10% acetonitrile and 1 μM propanolol/diclofenac internal standard. The amount of compound was quantified by LC/MS/MS in an Agilent 6140 Triple Quad LC/MS (Agilent, Santa Clara, CA). The integrity of the cell monolayer was assessed by measuring Lucifer Yellow transport by fluorometric determination at 385/435 nm (Victor2 Multilabel Counter, Perkin Elmer, Waltham, MA).

Permeability of the compounds was determined by calculating transport across the monolayer using the formula:

$$P_{app} = \frac{dQ/dt}{C_0 A}$$

where P_{app} is the permeability, dQ/dt is the rate of permeation, C_0 is the initial concentration of test agent, and A is the area of the monolayer.

Differences between Log P_{app}(AB) and Log P_{app}(BA) were calculated, and those compounds for which it was observed a difference over 1 Log unit between both directions were not retained for the final set, as they would be compounds involved in active transport. The final list (21 drugs) shows a high diversity of chemical structures, therapeutic activities and permeability values, thus constituting a good validation dataset (Table 6).

Acknowledgement

The authors wish to thank the Spanish Ministerio de Ciencia e Innovación (MCINN, SAF2008-01845) for its economic support.

References

1. Tsaioun, K.; Jacewicz, M. *Altern. Lab. Anim.* **2009**, *37*, 47–55.
2. Camenisch, G.; Folkers, G.; van de Waterbeemd, H. *Pharm. Acta Helv.* **1996**, *71*, 309–327.
3. Hidalgo, I. J.; Raub, T. J.; Borchardt, R. T. *Gastroenterology* **1989**, *96*, 736–749.
4. Artursson, P.; Karlsson, J. *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880–885.
5. Sambuy, Y.; De Angelis, I.; Ranaldi, G.; Scarino, M. L.; Stammati, A.; Zucco, F. *Cell. Biol. Toxicol.* **2005**, *21*, 1–26.
6. Avdeef, A.; Bendels, S.; Di, L.; Faller, B.; Kansy, M.; Sugano, K.; Yamauchi, Y. *J. Pharm. Sci.* **2007**, *96*, 2893–2909.
7. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25.
8. Frimurer, T. M.; Bywater, R.; Naerum, L.; Lauritsen, L. N.; Brunak, S. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315–1324.
9. Oprea, T. I. *J. Comput. Aided Mol. Des.* **2000**, *14*, 251–264.
10. Egan, W. J.; Lauri, G. *Adv. Drug. Deliv. Rev.* **2002**, *54*, 273–289.
11. Hou, T.; Wang, J.; Zhang, W.; Wang, W.; Xu, X. *Curr. Med. Chem.* **2006**, *13*, 2653–2667.
12. Akamatsu, M.; Fujikawaa, M.; Nakaob, K.; Shimizub, R. *Chem. Biodivers.* **2009**, *6*, 1845–1866.
13. Metcalfe, P. D.; Thomas, S. *Curr. Opin. Drug Discovery Dev.* **2010**, *13*, 104–110.
14. Gozalbes, R.; Pineda-Lucena, A. *Bioorg. Med. Chem.* **2010**, *18*, 7078–7084.
15. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R. *J. Pharm. Sci.* **2008**, *97*, 1946–1976.
16. Chaturveldi, P. R.; Deker, C. J.; Odinecs, A. *Curr. Opin. Chem. Biol.* **2001**, *5*, 452–463.
17. Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. *Pharm. Res.* **1998**, *15*, 1490–1494.
18. Gálvez, J.; García-Domenech, R.; Salabert, M.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520–525.
19. Irvine, J. D.; Takahashi, L.; Lockhart, K.; Cheong, J.; Tolan, J. W.; Selick, H. E.; Grove, J. R. *J. Pharm. Sci.* **1999**, *88*, 28–33.
20. Nordqvist, A.; Nilsson, J.; Lindmark, T.; Eriksson, A.; Garberg, P.; Kihlén, M. *QSAR Comb. Sci.* **2004**, *23*, 303–310.
21. Jung, S. J.; Choi, S. O.; Um, S. Y.; Kim, J. I.; Park Choo, H. Y.; Choi, S. Y.; Chung, S. Y. *J. Pharm. Biomed. Anal.* **2006**, *41*, 469–475.
22. Linnankoski, J.; Ranta, V. P.; Yliperttula, M.; Urtti, A. *Eur. J. Pharm. Sci.* **2008**, *34*, 129–139.
23. Dearden, J. C.; Cronin, M. T.; Kaiser, K. L. *SAR QSAR Environ. Res.* **2009**, *20*, 241–266.
24. Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
25. Yee, S. *Pharm. Res.* **1997**, *14*, 763–766.
26. Camenisch, G.; Alsenz, J.; van de Waterbeemd, H.; Folkers, G. *Eur. J. Pharm. Sci.* **1998**, *6*, 313–319.
27. Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. *J. Med. Chem.* **2001**, *44*, 1927–1937.
28. Pontier, C.; Pachot, J.; Botham, R.; Lenfant, B.; Arnaud, P. *J. Pharm. Sci.* **2001**, *90*, 1608–1619.
29. Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694–701.