Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

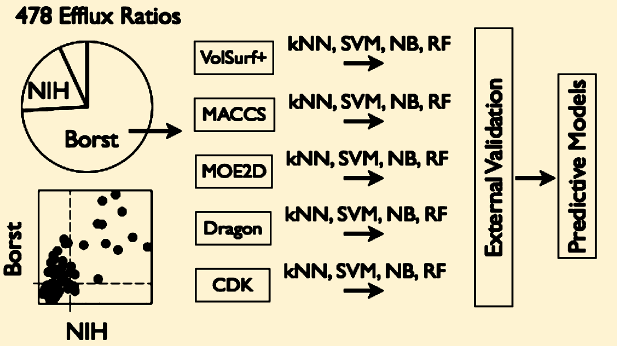
QSAR Models for P-Glycoprotein Transport Based on a Highly Consistent Data Set

Fabio Broccatelli*,[†]

[†]Laboratory of Chemometrics, Department of Chemistry, University of Perugia, Via Elce di Sotto 10, I-60123 Perugia, Italy

Supporting Information

ABSTRACT: P-Glycoprotein (Pgp) is involved in the elimination and in the disposition of a significant portion of marketed drugs. So far, publicly available data sets used for modeling Pgp transport included compounds tested in different assays, different cell lines, and different protocols. In this work, we present a collection of 478 Efflux Ratios (ERs) in MDCK-MDR1 cell lines, and from this collection we define a data set of 187 compounds that were tested in the Borst-derived MDCK-MDR1 cell lines. Of the 23 models resulting from the use of different descriptors, classification algorithms, and variable selection techniques, the 4 most accurate in external validation (~0.86) are based on VolSurf+ (VS+) descriptors. Two of these models are Naïve Bayes (NB) classifiers using 4 descriptors that were selected through a new technique hereby first time extensively described.



INTRODUCTION

ADME/Tox modeling is currently suffering the lack of accuracy in predicting the interactions of small molecules with transporters. Transporters are the key to anticipate the major route of elimination,[1−3] the brain penetration,[4−7] the bioavailability,[8,9] and the drug−drug and food−drug interactions of drugs and new molecular entities (NMEs).[10−12] The scientific community has invested a great deal of efforts in the transporters field; yet, data produced need to be rationalized and properly integrated to fulfill the transition from information to knowledge. Pgp (ABCB1) is easily the most studied human transporter.[13] Pgp decreases the brain and intestinal absorption of its substrates,[4−6,9] promotes their elimination unchanged in the urine and bile,[2,10] and regulates their intestinal metabolism.[14] In the oncologic field it is of interest to avoid Pgp transport, since its overexpression can lead to multidrug resistance.[15] Predicting the effect of Pgp in drug disposition is not straightforward, since its influence varies depending on the permeability, the solubility, and the metabolic profile of its substrates.[10] However, progress is being made in this field, due to a knowledge integration process involving different disciplines: PKPD modeling, chemistry, cheminformatics, and data mining.[1,3,4,16] Hence, it is anticipated that reliable models for Pgp transport will have an increasing impact in the near future.

The portion of the chemical space occupied by Pgp substrates ranges from the 28% to the 59% of the drug-like space,[6,7,17] depending on the chemical class considered. Wager et al.[6] reported that Pgp substrates are the ~46% of centrally acting compounds undergoing clinical trials. A similar percentage of Pgp substrates was observed among NMEs.[18] The high promiscuity of Pgp suggests a high importance of the entropic component of the energy of binding, which in turn is symptomatic of shape and hydrophobicity driven nonspecific interactions. We previously presented a composite model to predict Pgp inhibition that was used to classify over 500 external compounds with accuracy ~0.86.[19] Most of the predictions were based on the VS+[20,21] physicochemical holistic descriptors, while a minor fraction was based on the electrostatic and shape similarity to a minimalistic pharmacophore having a single polar region.[22] Based on past experience, in this work, Pgp transport is modeled using physicochemical descriptors. This is in line with a number of previous works that exploited a data set for Pgp transport presented by Penzotti and co-workers.[23] Starting from Seelig's[24] study, Penzotti assembled a literature collection, comprehensive of data derived from different typology of experiments, cell lines, and protocols. QSAR models based on Penzotti's data set, and trained with physicochemical descriptors, showed superior accuracy with respect to pharmacophoric and structure based approaches.[23,25−29] de Cerqueira Lima and co-workers started from Penzotti's data set to perform an exhaustive study resulting in several predictive models having accuracy in external validation up to 0.81.[25] Cabrera and co-workers integrated Penzotti's data set with other data and excluded from the collection molecules for which ambiguities were observed. Even if the removal of these compounds can be justified by the interpretation of the *in vitro* assays, it is the product of arbitrary judgments lacking mathematical criteria. In fact, one could

Received: June 18, 2012
Published: August 19, 2012

dx.doi.org/10.1021/ci3002809 | *J. Chem. Inf. Model.* 2012, 52, 2462−2470

argue that a number of borderline Pgp substrates (ER 1.8−2.5), were included in the data set; other investigators used Cabreara's data set and gained accuracy in external validation up to 0.9.[26,28] QSAR models based on proprietary data sets and using machine learning approaches to model holistic physicochemical descriptors typically reached accuracy in external validation below 0.87.[18,30] Interestingly, Crivori et al.[31] showed that publicly available data can be used to train models successful in predicting the Pgp profiles of new molecular libraries.

Recently we provided an extensive collection of ER data (the gold standard protocol for Pgp transport) in MDCK-MDR1 cell lines.[4] In several cases we noted that the same compound tested in MDCK-MDR1 cell lines retrieved from different sources (e.g., NIH or Borst) showed very different results. In this work, the collection previously presented is improved: 478 ER values in MDCK-MDR1 cell lines were analyzed and used to create a data set of 187 compounds tested in Borst MDCK-MDR1 cell lines. The data set was modeled using different descriptors and classification algorithms. Two modeling strategies were used: one based on a new feature selection technique and one based on the Chembench approach.[32−34] Finally, mispredictions and relevant descriptors are discussed.

■ METHODS

Data Set. Small molecules can be characterized as Pgp substrates or nonsubstrates based on the results of the ER assay. The ER is the ratio between the apparent permeability from basolateral to apical direction (secretory) and the apparent permeability from apical to basolateral direction (absorptive) in a cell line overexpressing Pgp. Compounds having ER equal or above 2 are usually assumed to be Pgp substrates, even though a number of Pgp substrates were found having ER under 2.[17] A further major limitation of the ER assay is that it is prone to produce false negatives when compounds tested have high passive permeability: those substrates are defined as "nontransported substrates"[17] and are potentially a confounding factor for QSAR models. Alternative assays to characterize the interaction of small molecules with Pgp are the ATPase assay and the competitive inhibition assays, and these assays match the outcomes of the ER assay in only the ~50% of the cases.[17] The three assays used in parallel allow to fully characterize the interaction of Pgp with small molecules and can be used to distinguish nonsubstrates from nontransported substrates. Unfortunately, there is limited availability of these kinds of data in the literature. Therefore, we defined Pgp substrates exclusively based on the ER assay, including the nontransported substrates in the data set as Pgp nonsubstrates. The systematic error introduced by this choice is addressed by estimating the maximal theoretical accuracy for this data set (see Results and Discussion).

In order to minimize the variability, only ERs produced in MDCK-MDR1 cell lines were collected. This effort resulted in 478 data (414 quantitative, 65 qualitative) extracted from 30 different references.[6,7,17,30,35−60] Most of the data (354) were originated in Netherlands Cancer Institute (Borst) MDCK-MDR1 cell lines; 91 data were produced in National Institutes of Health (NIH) MDCK-MDR1 cell lines, while the remaining 33 data were originated form other sources. Borst and NIH data were separately elaborated. When only one numeric value was available, that value was used to determine the Pgp class, and when multiple values were available, the values were averaged. For 9 compounds the numeric ER value was not available, thus the qualitative value was adopted. Compounds having ERs equal or above 2 were assigned to the substrates class, and the others were considered nonsubstrates. The Borst data set included 187 compounds (110 nonsubstrates and 77 substrates). The NIH data set included 84 compounds, 56 of which had Borst ER available. For these compounds, the NIH ER was not a good predictor of the Pgp class as defined by the Borst ER (see Result and Discussion). Hence, the data set was only inclusive of data produced in the Borst cell lines. SMILES structures were produced using MarvinSketch[61] and were double-checked using Chemspider,[62] when this was possible.

A Principal Component Analysis (PCA) based on 128 VS+ descriptors was computed for the 187 compounds in the data set. The validation set was extracted by selecting 20% of the substrates (15 compounds) and 20% of the nonsubstrates (22 compounds), by means of the most descriptive compounds subset selection algorithm (available in VS+) applied to the first five PCs. The algorithm was forced to exclude the 16 compounds for which there is lack of agreement in the Pgp class from the validation set (see Results and Discussion). The final training set included 150 compounds (62 substrates and 88 nonsubstrates).

Descriptors. Conformation, protonation dependent chemical descriptors were computed using VS+. The suitability of VS + descriptors in transporters related issues was demonstrated in a number of works.[19,25,31,63] Most of the VS+ descriptors result from the elaboration and the condensation of the GRID Molecular Interaction Fields (MIFs).[64,65] Protonation dependent descriptors are calculated exploiting MoKa pK_a calculations.[66,67] The most stable conformation for each compound was produced starting from the SMILES structures, using the VS+ minimizer with default options. In total, 128 descriptors were derived from the 187 neutralized structures. Using Chembench it was possible to calculate several other 2D descriptors: MOE2D, MACCS keys, DRAGON, and CDK.[68−71]

Modeling. The resulting matrices were reduced in dimension and modeled using widgets and classifiers available in Orange Canvas and in Chembench.[32−34,72] Orange provides tools for each step of the data mining process including the following: a) evaluation of the impact of each descriptor on the entropy of the data set, b) linear and nonlinear classifiers and c) test learner, for internal n-fold cross-validation (CV) and external validation. In the Orange approach, Information Gain (IG) was adopted to reduce and prioritize the variables considered during the variables selection process. IG provides a measure of the reduction of the uncertainty (entropy) associated with the presence of a variable in the data set. The quality of the models was evaluated in terms of Classification Accuracy (CA), Sensitivity (Se), Specificity (Sp), Matthews Correlation Coefficient (MCC), and Area under the Curve (AUC)

$$\mathrm{CA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{MCC} = \sqrt{\frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{(\mathrm{TP} + \mathrm{FP}) + (\mathrm{TP} + \mathrm{FN}) + (\mathrm{TN} + \mathrm{FP}) + (\mathrm{TN} + \mathrm{FN})}}$$

where TP are true positives predictions, TN are true negatives predictions, FP are false positives predictions, and FN are false negatives predictions.

In MCC both accuracy and precision (reliability of a prediction) are important, thus the uneven distribution of the objects in the two classes is less likely to result in biased models. AUC is the measure of the area under the ROC curve. Unlike CA and MCC, AUC is not dependent on the threshold used.

VS+ descriptors were modeled using the following algorithms implemented in Orange: NB, k Nearest Neighbors (kNN), and Support Vector Machines (SVM). The NB classifier was used with the following options: Laplace function for probability estimation, LOESS window equal to 0.5, size equal to 100, adjusted threshold for binary classes. In the Orange kNN models, neighbors distances were weighted by rank, the distance metrics was euclidean (with continuous attributes normalized). *k* was equal to 6 in the initial stages of the variable selection process, and it was optimized ($1 \leq k \leq 6$) at the end of the process. In the Orange SVM models, the parameters were optimized using the automatic parameter selection, and after computing SVM models with each available kernel function, the best model was selected.

Eighteen models were derived using the cheminformatic Web tool Chembench. Designed by leaders in the QSAR field, Chembench includes 4 different classifiers (of which one was not used due the long computational time required) and 6 sets of descriptors (of which one was not used due to technical issues). Furthermore, it was possible to upload and use VS+ descriptors. This provided means to evaluate the suitability of the Orange-VS+ approach. Chembench models were computed using the following classifiers with default options: Random Forest (RF), SVM, and GA-kNN (kNN exploiting a genetic algorithm for the variable selection process). The 3 classifiers were used with 6 sets of descriptors (VS+, Dragon with Hydrogens, Dragon without Hydrogens, CDK, MOE2D, and MACCS).

Models were internally validated using 5-fold CV. Predicted activities for the training set were analyzed using the R package ROCR,[73] in order to find the predicted activity threshold that maximized the MCC. Only models having both specificity and sensitivity above 0.7 in external validation were accepted.

Feature Selection. The feature selection procedure used is an evolution of the one reported in our recent work and involves the use of one classifier (either NB or kNN) and one score to optimize (either AUC or MCC).[3] The choice of the classifiers was done to minimize the computation time. AUC and MCC were used as scores since one is independent of the threshold used (AUC), and the other is a balanced measure of robustness (MCC). The variable selection process is schematized in Figure 1 and can be described as follows: descriptors are ranked based on their IG value, and the descriptor having the highest IG value is initially used alone in the classifier. The score value of the resulting model is stored, and descriptors are iteratively added in decreasing IG order. For each new combination, a model is calculated and a corresponding score value is stored. If the score value of the new model is higher than the score value of the previous model, the new model is adopted. For a new combination of N variables, N leave-one-descriptor-out models are computed, and the new model having highest score is adopted. The loop restarts with the addition of the following descriptor and terminates when all the descriptors considered have been evaluated. When models are equivalent in terms of score, a secondary score (e.g., CA) is considered. When the first model having both specificity and sensitivity over 0.7 is found, new models are accepted only if they have both higher score value and specificity and sensitivity above 0.7.

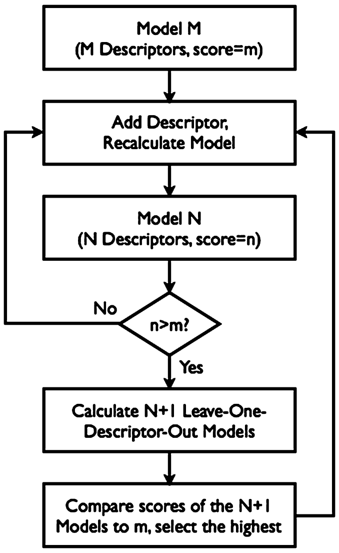

Model M
(M Descriptors, score=m)
Add Descriptor, Recalculate Model
Model N
(N Descriptors, score=n)
No
n>m?
Yes
Calculate N+1 Leave-One-Descriptor-Out Models
Compare scores of the N+1 Models to m, select the highest

Figure 1. Flowchart of the Information Gain based feature selection technique presented in this study.

In order to reduce the calculation time and the noise derived by unuseful descriptors, only the 30 descriptors having highest IG were considered. In total 6 models were performed using Orange Canvas; 4 are a combination of the two classifiers (NB and kNN) and the two scores (AUC and MCC). The descriptors selected in these models were used to train SVMs, and only the two SVM models maximizing AUC and MCC were selected. In this specific case, they coincided.

■ RESULTS AND DISCUSSION

In this work, we present several classification models for predicting Pgp transport based on a new and highly consistent data set including 187 compounds tested in MDCK-MDR1 Borst cell lines. We recently presented an extensive collection of ERs (316 data-points).[4] We also noted that it is not infrequent to find very different results for the same compounds, depending on the source of the MDCK-MDR1 cell line. Here this observation was supported by a more comprehensive data collection (476 data-points): for the 56 molecules having ER in both Borst and NIH MDCK-MDR1 cell lines, no correlation was observed ($r^2 = 0.07$) (Figure 2a). Furthermore, NIH ER poorly predicted Pgp class defined by Borst ERs (CA = 0.6, AUC = 0.79) (Figure 2b).

Thus, the final data set was derived by ERs produced in Borst cell lines only (Figure 3). For the 91 compounds included in the Borst data set that had multiple ERs available, the "agreement class" ("Yes", "No", or "borderline") was determined (Figure 4). For 16 compounds lack of agreement (in terms of Pgp class) was observed between different references. Ten of them (chlorpromazine, doxorubicin, fluoxetine, fluvoxamine, loratadine, neostigmine, ranitidine, risperidone, trimethoprim, verapamil) were assigned to the "No agreement" class, since they have at least one ER above 2.5 and one ER under 1.8. The remaining 6 (bromocriptine, cyclobenzaprine, methysergide, morphine, nalbuphine, parox-

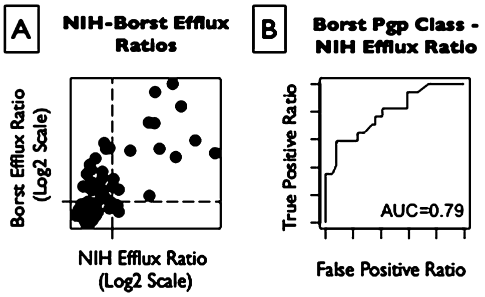


Figure 2. Comparison of the ERs produced in MDCK-MDR1 Borst and in MDCK-MDR1 NIH cell lines. a) Scatter plot of Borst and NIH ERs in logarithmic scale; the 2nd and the 4th quadrants contain mispredictions if a threshold value of ER = 2 is adopted for categorizing Pgp substrates. b) Pgp classes defined based on Borst ER are predicted using NIH ER.

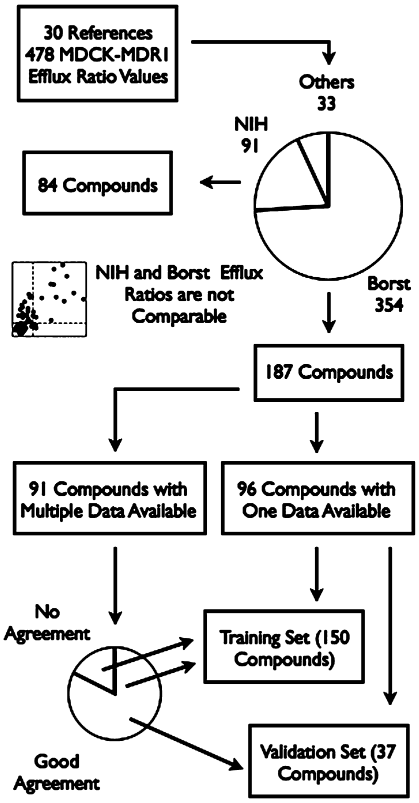


Figure 3. Training set and validation set definition.

etine) were assigned to the "borderline" class. Assuming that a theoretically perfect model would correctly classify all the compounds in the "Yes" class, incorrectly classify all the compounds in the "No" class, and correctly classify half of the compounds in the "borderline" class, the maximal theoretical accuracy for this data set can be esteemed to be ~0.86. However, the choice of not including ambiguous data in the validations set leads to an uneven distribution of the experimental noise between internal and external validation.

Shown in Table 1 are the statistics in 5-fold CV and in external validation for the Orange-VS+ models accepted and the names of the variables used. Variables were selected based on a new procedure, aiming to the identification of few highly informative descriptors. These processes offered different results based on the classifier used and the score optimized. When the only difference between the models was the classifier,

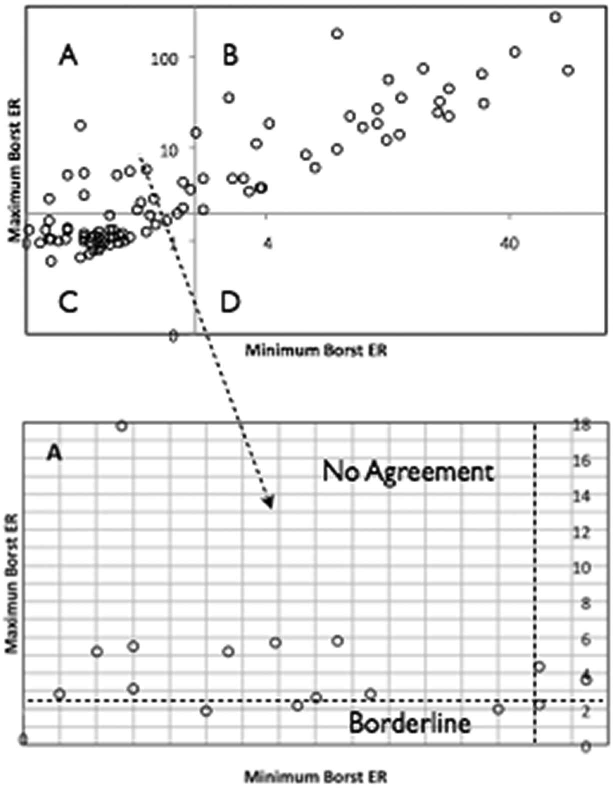


Figure 4. Comparison of the maximum ER and the minimum ER in Borst cel lines for the 91 compounds having multiple ERs available. The plot is divided in 4 quadrants identified by the threshold ER = 2. Quadrant A, containing compounds for which lack of agreement was observed, is enlarged in order to better distinguish compounds in the "No Agreement" class and in the "borderline" class.

NB always outperformed kNN models. When the only difference was the score used during the optimization process, models using MCC always outperformed those using AUC. The best model, in terms of accuracy in external validation (NB-MCC), employed 4 VolSurf+ descriptors (CA = 0.86). The second best Orange-VS+ model in external validation (NB-AUC) was also a NB classifier employing 4 descriptors (CA = 0.81). The two other Orange-VS+ models accepted were a kNN using 7 variables and a SVM using 11 variables, both having accuracy ~0.78. In order to reduce the computational burden, the exhaustive search for the best possible combination of descriptors was only performed for NB and kNN models; SVM models employed all the descriptors resulting from these searches. Thus it is possible that the descriptors used in the SVM models significantly differ in terms of names and quantity from the best possible combination. This is likely why better results were obtained with NB and kNN models. By definition the accuracy depends on the threshold value used for the classifications, while AUC does not; not surprisingly, the most accurate Orange-VS+ model (NB-MCC) was not the one having highest AUC. By changing the threshold, two Orange-VS+ models with highest AUC (SVM and NB-AUC) achieved accuracy ~0.86. Due to the surprisingly high accuracy gained, the possibility of assembling further composite models was not explored.

Orange-VS+ models were compared to 18 Chembench models (Table 2). In this context, VS+ models always outperformed (or equaled) the other models using same classifier and different descriptors. For this data set, it appears

Table 1. Predictive Orange Models

	5-fold CV			external validation					
model	Ac	AUC	MCC	Ac	AUC	MCC	Se	Sp	VS+ descriptors
NB-AUC	0.81	0.86	0.62	0.81	0.90	0.60	0.73	0.86	DRDRDO, S, POL, W6
NB-MCC	0.81	0.86	0.62	0.86	0.85	0.72	0.73	0.95	DRDRDO, DRACDO, S, W6
kNN-MCC	0.83	0.86	0.64	0.78	0.79	0.58	0.87	0.73	DRDRDO, DIFF, V, W1, W2, W4, WN4
SVM-RBF	0.81	0.85	0.63	0.78	0.87	0.55	0.73	0.82	DRDRDO, DRACDO, S, POL, DIFF, V, W1, W2, W6, W4, WN4

Table 2. Chembench Models (External Validation)

	RF		SVM		GA-kNN		mean	
	Ac	AUC	Ac	AUC	Ac	AUC	Ac	AUC
VS+	0.84	0.88	0.73	0.85	0.81	0.89	0.79	0.87
CDK	0.78	0.83	0.65	0.70	0.78	0.87	0.74	0.80
MOE2D	0.68	0.71	0.68	0.78	0.70	0.78	0.69	0.76
DRH	0.65	0.84	0.73	0.76	0.78	0.85	0.72	0.82
DR	0.78	0.85	0.73	0.81	0.78	0.85	0.76	0.84
MACCS	0.70	0.71	0.62	0.69	0.65	0.71	0.66	0.70
Mean	0.74	0.80	0.69	0.76	0.75	0.82		

Table 3. Predictive Chembench Models

	5-fold CV			external validation					
model	Ac	AUC	MCC	Ac	AUC	MCC	Se	Sp	3 most used descriptors
RF-VS+	0.81	0.87	0.62	0.84	0.88	0.66	0.80	0.86	DRDRDO, VD, DRACDO
GAkNN-VS+	0.83	0.90	0.65	0.81	0.89	0.63	0.86	0.81	DRDRDO, DRACDO, ACACDO

that RF and GA-kNN perform better than SVM classifiers; it is not entirely clear if this is due to the variable selection procedure adopted or to other factors (e.g., uneven distribution of the objects in the 2 classes). Two Chembench models were accepted based on the external predictions: RF-VS+ (CA = 0.84) and GA-kNN-VS+ (CA = 0.81) (Table 3). Also in this case, by changing threshold it was possible to achieve accuracy ~0.86.

For this data set high accuracy was reached by several models using different classifiers (NB, RF, SVM, and kNN) and different variable selection procedures. Yet, the use of 3D descriptors appears to be a very relevant factor. Concerning the feature selection, the approach presented in this work has two major advantages: a) the short calculation time (<5 min) and b) the low number of variables selected (4 descriptors, one every 37.5 objects in the training set). The latter point could improve the interpretation of the results and the rules-of-thumb based design of novel compounds; however, such a low number of variables could also lead to unstable models. Particularly in this case, multiple external validation sets are necessary to assess the real accuracy of the model.[74]

The misclassifications of the 4 most accurate models (NB-MCC-VS+, NB-AUC-VS+, GAkNN-VS+, and RF-VS+) were analyzed by dividing the data set in in four classes: a) “inconsistent” class (including 16 compounds having different Pgp substrate class depending on the reference considered), b) “borderline” class (including 18 compounds having ER in the 1.8−2.5 range), c) “nontransported substrates” class (only the few reported on by Polli et al.[17] are considered), and d) “reliable” class (including the compounds not belonging to the other classes).

Table 4. Mispredictions Analysis

class	validation	N comp	N comp misc	tot misc	mean Ac
reliable	5-fold-CV	123	35	63	0.87
NontransSubs	5-fold-CV	7	5	15	0.46
borderline	5-fold-CV	13	9	30	0.42
inconsistent	5-fold-CV	16	9	21	0.67
reliable	external	31	8	17	0.86
NontransSubs	external	1	0	0	1
borderline	external	5	3	7	0.65
inconsistent	external	0	0	0	0

Results of this analysis are shown in Table 4. Interestingly, the difference in CA between the internal validation (average accuracy 0.81) and the external validation (average accuracy 0.83) disappears when only molecules belonging to the “reliable” class are considered, implying that this discrepancy results from the exclusion of the “inconsistent data” from the validation set. In 5-fold CV the CA for these 16 compounds was, on average, 0.7; for borderline compounds predictions were random (average CA = 0.5). Also in external validation, borderline compounds were the main sources of misclassification: 3 out of 5 were misclassified by at least one of the 4 most predictive models. Pravastatin acid (ER = 1.87) was incorrectly predicted as Pgp substrate by all of them. If the cutoff ER value to categorize Pgp substrates is set to ER = 2, borderline substrates and inconsistent data mostly belong to the Pgp substrates class. This explains why for this work, as well as for most of the Pgp models presented so far, the specificity is generally lower than the sensitivity.[25,26,28,30] Recently Gupta et al.[18] presented a model for Pgp transport based on 43408 compounds. In this data set, borderline substrates were included in the Pgp nonsubstrates class, and, consistently, models derived had high specificity and lower sensitivity.

The most accurate models presented in this work are based on VolSurf+ descriptors. Main features of VolSurf+ descriptors are the 3-Dimensionality, the use GRID MIFs, and the use of pK_a dependent descriptors. The 11 descriptors deemed relevant based on the variable selection process described above approximately belong to 3 families: size descriptors, hydrophilic surface descriptors, and TOPP descriptors. None of these descriptors is pK_a dependent. The TOPP descriptor DRDRDO (highest IG and most used descriptor) is the measure of the largest triangle individuated by two hydrophobic regions and a hydrogen bond donor. This descriptor does not exploit information derived by GRID calculations but requires an energy minimization and a conformational analysis to individuate the most suitable conformation. Except for W2, the relevant size descriptors (S, POL, DIFF, V, W1) are highly correlated with the MW ($r^2 > 0.89$); among them POL is not a 3D GRID based descriptor and can be rapidly calculated. The descriptor W6 (most frequently used hydrophilic surface descriptor) is correlated with the PSA ($r^2 > 0.8$) that could also be derived by using 2D approaches. Taken together these considerations suggest that the 3-Dimensionality of VolSurf+ descriptors is the main reason for their superiority with this data set.

Boxplots in Figure 5 represent the distribution of Pgp substrates and nonsubstrates with respect to the above-discussed descriptors and LogP n-Oct. DRDRDO provides very straightforward information that could be used for rule-of-thumb purposes: the 33 compounds having DRDRDO > 28.7 are all Pgp substrates, and of the 37 compounds having DRDRDO = 0, only 1 has ER > 2.5. Similarly, the 24 compounds having S < 400 are all nonsubstrates, while 35 out of the 38 compounds with S > 665 are Pgp substrates; in the latter case, 2 of the three outliers are nontransported substrates.[17] Using the rules defined above, 97 of the 187 compounds (51.4%) can be classified with very high accuracy (93.7%). These criteria could be used for fast database screening.

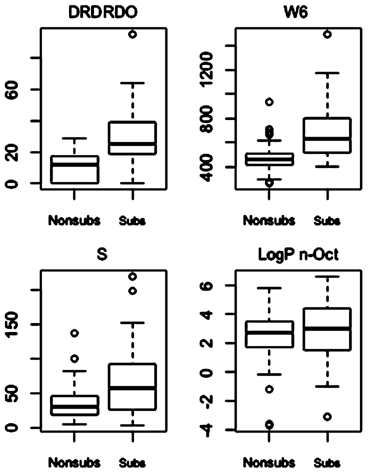


Figure 5. Boxplots for 3 relevant descriptors and LogP n-Oct. The box edges represent the 1st and the 3rd quartiles, the whiskers represent minimum and maximum values, dots are outliers, and the line in the middle of the box is the median.

Results of the models described in this work were compared with the results of the Pgp inhibition model that we previously presented.[19] Consistent with what was observed in vitro by Polli and co-workers,[17] the relevant descriptors of the two models suggest that high passive permeability promote Pgp inhibition but not Pgp transport. LogP n-Oct, a descriptor that estimates the tendency of drugs to distribute in hydrophobic environments, does not offer any separation between Pgp substrates and nonsubstrates (Figure 5). On the contrary, LogP n-Oct was one of the most relevant descriptors in the Pgp inhibition model. Furthermore, Pgp transport appears to be promoted by the presence of hydrogen bond donors, while for Pgp inhibition the only polar requirement is a hydrogen bond acceptor. Not surprisingly, hydrogen bond donors are more detrimental for permeability in comparison to hydrogen bond acceptors.[75]

We[4] previously reported on the correlation between Lipinski Rule of five[75] violations and Pgp transport. Based on this data set and on the "LipinskiFailures" CDK descriptor, 83% of the compounds with at least one violation are Pgp substrates. This is in good agreement with the detoxification role that Pgp is generally believed to play.

■ CONCLUSIONS

In this work, we presented several QSAR models for Pgp transport. The models exploit a novel data set composed by 187 publicly available compounds, classified based on Efflux Ratio data in MDCK-MDR1 Borst cell line. The data set stems from a rigorous analysis of over 450 ER data. The analyses confirm the big experimental variability for ER data associated with different MDCK-MDR1 cell lines (particularly Borst vs NIH).[4] In total, 23 models were computed using various combinations of descriptors, classification algorithms, and variable selection techniques, one of which is here for the first time extensively described. The four best models all employ VS+ descriptors: two of them are NB classifiers using 4 descriptors, and the other two are Chembench models (GA-kNN and RF classifiers). In external validation, accuracy up to 0.86 was achieved. In 5-fold CV lower accuracies were observed, and a thorough analysis of the mispedicition demonstrates that this difference is the consequence of the exclusion of compounds with ambiguous class from the validation set. Further sources of misclassifications were compounds that have borderline values and nontransported substrates. Due to the activity threshold adopted, the experimental noise derived by the use of these compounds affects the specificity more than the sensitivity.

The suitability of VS+ for modeling this data set is a consequence of the 3-Dimensionality of its descriptors. Size, shape, and hydrophobic features result as the driving force of Pgp transport, in agreement with the high promiscuity observed for this transporter. Not surprisingly, pharmacophoric approaches generally failed to discriminate Pgp substrates from nonsubstrates using data sets with similar composition to the one presented in this study. It would also not be surprising to observe an accuracy drop when applying our models to an unbiased space containing a significant number of Pgp substrates having intermediated size, polarity, and flexibility (e.g., CNS candidate drugs) or having passive permeability high enough to saturate efflux transporters (nontransported substrates). In this context, we anticipate that a) shape based pharmacophoric hypotheses and b) prefiltering based on passive permeability modeling could effectively integrate our work. In conclusion, the accuracy of Pgp transport modeling is limited by the quality of the data set, not by the computational approaches used. Substrates, nonsubstrates, and nontransported substrates should be modeled as 3 different classes, and, to achieve this aim, the compounds in the data set should be tested in parallel in ER, ATPase, and competitive inhibition assays. Also, the standard deviations of the secretory and absorptive measured permeabilities should be addressed by excluding borderline compounds (ER 1.8−2.5). Lastly, a more densely populated set could allow uncovering the pharmacophoric and shape requirements underlying the Pgp transport.

■ ASSOCIATED CONTENT

S Supporting Information

An xls file is available containing all the data collected (SI-1), the data set used (SI-2), the VS+ descriptors (SI-3), the comparison between NIH and Borst MDCK-MDR1 Efflux Ratios (SI-4), and the predictions in internal and external validation for the 23 models computed (SI-5, SI-6, SI-7). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fabio@chemiome.chm.unipg.it. Corresponding address: University of Perugia, Via Elce di Sotto 10, I-60123 Perugia, Italy.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

I express my thanks to Prof. Tropsha and to Prof. Pozefsky for the assistance with the use of Chembench and to Dr. Anita Grover for the help with the revision of this manuscript.

■ ABBREVIATIONS USED

Pgp, P-Glycoprotein; VS+, VolSurf+; NIH, National Institutes of Health; NB, Naïve Bayes; SVM, Support Vector Machine; RF, Random Forest; MDCK, Madin-Darbey Canine Kideny Cell; kNN, k Nearest Neighbors; GA, Genetic Algorithm; CV, cross-validation; IG, Information Gain; CA, Classification Accuracy; Sp, Specificity; Se, Sensitivity; TP, True Positive; TN, True Negative; FN, False Negative; FP, False Positive; AUC, Area Under the Receiver Operating Characteristic Curve; MCC, Matthews Correlation Coefficient; MIF, Molecular Intaraction Field

■ REFERENCES

(1) Varma, M. V.; Gardner, I.; Steyn, S. J.; Nkansah, P.; Rotter, C. J.; Whitney-Pickett, C.; Zhang, H.; Di, L.; Cram, M.; Fenner, K. S.; El-Kattan, A. F. pH-Dependent Solubility and Permeability Criteria for Provisional Biopharmaceutics Classification (BCS and BDDCS) in Early Drug Discovery. *Mol. Pharmaceutics* **2012**, *9*, 1199−1212.

(2) Varma, M. V. S.; Feng, B.; Obach, R. S.; Troutman, M. D.; Chupka, J.; Miller, H. R.; El-Kattan, A. Physicochemical Determinants of Human Renal Clearance. *J. Med. Chem.* **2009**, *52*, 4844−4852.

(3) Broccatelli, F.; Cruciani, G.; Benet, L. Z.; Oprea, T. I. BDDCS Class Prediction for New Molecular Entities. *Mol. Pharmaceutics* **2012**, *9*, 570−580.

(4) Broccatelli, F.; Larregieu, C. A.; Cruciani, G.; Oprea, T. I.; Benet, L. Z. Improving the Prediction of the Brain Disposition of Orally Administered Drugs using BDDCS. *Adv. Drug Delivery Rev.* **2012**, *64*, 95−109.

(5) Broccatelli, F.; Carosati, E.; Cruciani, G.; Oprea, T. I. Transporter-Mediated Efflux Influences CNS Side Effects: ABCB1, from Antitarget to Target. *Mol. Inf.* **2010**, *29*, 16−26.

(6) Wager, T.; Chandrasekaran, R.; Hou, X.; Troutman, M.; Verhoest, P.; Villalobos, A.; Will, Y. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chem. Neurosci.* **2010**, *1*, 420−434.

(7) Mahar Doan, K.; Humphreys, J.; Webster, L.; Wring, S.; Shampine, L.; Serabjit-Singh, C.; Adkison, K.; Polli, J. J. Passive Permeability and P-Glycoprotein-Mediated Efflux Differentiate Central Nervous System (CNS) and non-CNS Marketed Drugs. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 1029−1037.

(8) Varma, M.; Obach, R.; Rotter, C.; Miller, H.; Chang, G.; Steyn, S.; El-Kattan, A.; Troutman, M. Physicochemical Space for Optimum Oral Bioavailability: Contribution of Human Intestinal Absorption and First-Pass Elimination. *J. Med. Chem.* **2010**, *53*, 1098−1108.

(9) Varma, M. V. S.; Sateesh, K.; Panchagnula, R. Functional Role of P-Glycoprotein in Limiting Intestinal Absorption of Drugs: Contribution of Passive Permeability to P-Glycoprotein Mediated Efflux Transport. *Mol. Pharmaceutics* **2004**, *2*, 12−21.

(10) Shugarts, S.; Benet, L. Z. The Role of Transporters in the Pharmacokinetics of Orally Administered Drugs. *Pharm. Res.* **2009**, *26*, 2039−2054.

(11) Custodio, J. M.; Wu, C. Y.; Benet, L. Z. Predicting Drug Disposition, Absorption/Elimination/Transporter Interplay and the Role of Food on Drug Absorption. *Adv. Drug Delivery Rev.* **2008**, *60*, 717−733.

(12) Wu, C. Y.; Benet, L. Z. Predicting Drug Disposition via Application of BCS: Transport/Absorption/ Elimination Interplay and Development of a Biopharmaceutics Drug Disposition Classification System. *Pharm. Res.* **2005**, *22*, 11−23.

(13) Juliano, R. L.; Ling, V. A Surface Glycoprotein Modulating Drug Permeability in Chinese Hamster Ovary Cell Mutants. *Biochim. Biophys. Acta* **1976**, *455*, 152−162.

(14) Benet, L. Z.; Cummins, C. L.; Wu, C. Y. Unmasking the Dynamic Interplay between Efflux Transporters and Metabolic Enzymes. *Int. J. Pharm.* **2004**, *277*, 3−9.

(15) Choi, C. ABC Transporters as Multidrug Resistance Mechanism and the Development of Chemosensitizers for Their Reversal. *Cancer Cell Int.* **2005**, *5*, 30.

(16) Benet, L. Z.; Broccatelli, F.; Oprea, T. I. BDDCS Applied to Over 900 Drugs. *AAPS J.* **2011**, *13*, 519−547.

(17) Polli, J.; Wring, S.; Humphreys, J.; Huang, L.; Morgan, J.; Webster, L.; Serabjit-Singh, C. J. Rational Use of in Vitro P-Glycoprotein Assays in Drug Discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620−628.

(18) Gupta, R. R.; Gifford, E. M.; Liston, T.; Waller, C. L.; Hohman, M.; Bunin, B. A.; Ekins, S. Using Open Source Computational Tools for Predicting Human Metabolic Stability and Additional Absorption, Distribution, Metabolism, Excretion, and Toxicity Properties. *Drug Metab. Dispos.* **2010**, *38*, 2083−2090.

(19) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A Novel Approach for Predicting P-Glycoprotein (ABCB1) Inhibition Using Molecular Interaction Fields. *J. Med. Chem.* **2011**, *54*, 1740−1751.

(20) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular Fields in Quantitative Structure−Permeation Relationships: The VolSurf Approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17−30.

(21) VolSurf+, version 1.0.4; Molecular Discovery Ltd.: London, UK, 2000.

(22) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2009**, 1−16.

(23) Penzotti, J.; Lamb, M.; Eversen, E.; Grootenhuis, P. D. J. A Computational Ensemble Pharmacophores Model for Identifying Substrates of P-Glyciprotein. *J. Med. Chem.* **2002**, *45*, 1−4.

(24) Seelig, A. A General Pattern for Substrate Recognition by P-Glycoprotein. *Eur. J. Biochem.* **1998**, 1−10.

(25) de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245−54.

(26) Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-Glycoprotein Substrates Using a Support Vector Machine Optimized by a Particle Swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638−1647.

(27) Xue, Y.; Yap, C.; Sun, L.; Cao, W.; Wang, J.; Chen, Y. Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1497−1505.

(28) Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R. C.; Yan, A. P-Glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Data Set. *J. Chem. Inf. Model.* **2011**, *27*, 1447−1456.

(29) Cabrera, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. A Topological Substructural Approach for the Prediction of P-Glycoprotein Substrates. *J. Pharm. Sci.* **2006**, *95*, 589−606.

(30) Gombar, V.; Polli, J.; Humphreys, J.; Wring, S.; Serabjit-Singh, C. J. Predicting P-Glycoprotein Substrates by a Quantitative Structure-Activity Relationship Model. *Pharm. Sci.* **2004**, *93*, 957−968.

(31) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggesi, I. Computational Models for Identifying Potential P-Glycoprotein Substrates and Inhibitors. *Mol. Pharmaceutics* **2005**, *3*, 33−44.

(32) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(33) Chembench. http://chembench.mml.unc.edu/ (accessed March 2012).

(34) Walker, T.; Grulke, C. M.; Pozefsky, D.; Tropsha, A. Chembench: A Cheminformatics Workbench. *Bioinformatics* **2010**, *26*, 3000−3001.

(35) Baltes, S.; Gastens, A.; Fedrowitz, M.; Potschka, H.; Kaever, V.; Löscher, W. Differences in the Transport of the Antiepileptic Drugs Phenytoin, Levetiracetam and Carbamazepine by Human and Mouse P-Glycoprotein. *Neuropharmacology* **2007**, *52*, 333−346.

(36) Bertelsen, K. M.; Greenblatt, D. J.; von Moltke, L. L. Apparent Active Transport of MDMA Is Not Mediated by P-Glycoprotein: A Comparison with MDCK and Caco-2 Monolayers. *Biopharm. Drug Dispos.* **2006**, *27*, 219−227.

(37) Callegari, E.; Malhotra, B.; Bungay, P. J.; Webster, R.; Fenner, K. S.; Kempshall, S.; LaPerle, J. L.; Michel, M. C.; Kay, G. G. A Comprehensive Non-Clinical Evaluation of the CNS Penetration Potential of Antimuscarinic Agents for the Treatment of Overactive Bladder. *Br. J. Clin. Pharmacol.* **2011**, *72*, 235−46.

(38) Carrara, S.; Reali, V.; Misiano, P.; Dondio, G.; Bigogno, C. Evaluation of in Vitro Brain Penetration: Optimized PAMPA and MDCKII-MDR1 Assay Comparison. *Int. J. Pharm.* **2007**, *345*, 125−133.

(39) Chang, C.; Bahadduri, P.; Polli, J.; Swaan, P.; Ekins, S. Pharmacophore-Based Discovery of Ligands for Drug Transporters. *Drug Metab. Dispos.* **2006**, *34*, 1976−1984.

(40) Chen, C.; Hanson, E.; Watson, J.; Lee, J. P-Glycoprotein Limits the Brain Penetration of Nonsedating but Not Sedating H1-Antagonists. *Drug Metab. Dispos.* **2003**, *31*, 312−318.

(41) Cummins, C. L.; Jacobsen, W.; Christians, U.; Benet, L. Z. CYP3A4-Transfected Caco-2 Cells as a Tool for Understanding Biochemical Absorption Barriers: Studies with Sirolimus and Midazolam. *J. Pharmacol. Exp. Ther.* **2004**, *308*, 143−155.

(42) de Souza, J.; Benet, L. Z.; Huang, Y.; Storpirtis, S. Comparison of Bidirectional Lamivudine and Zidovudine Transport Using MDCK, MDCK-MDR1, and Caco-2 Cell Monolayers. *J. Pharm. Sci.* **2009**, *98*, 4413−4419.

(43) Eriksson, U.; Dorani, H.; Karlsson, J.; Fritsch, H.; Hoffmann, K.; Olsson, L.; Sarich, T.; Wall, U.; Schutzer, K. Influence of Erythromycin on the Pharmacokinetics of Ximelagatran May Involve Inhibition of P-Glycoprotein-Mediated Excretion. *Drug Metab. Dispos.* **2006**, *34*, 775−782.

(44) Feng, B.; Mills, J.; Davidson, R.; Mireles, R.; Janiszewski, J.; Troutman, M.; De Morais, S. In Vitro P-glycoprotein Assays To Predict the in Vivo Interactions of P-Glycoprotein with Drugs in the Central Nervous System. *Drug Metab. Dispos.* **2008**, *36*, 268−275.

(45) Gertz, M.; Harrison, A.; Houston, J. B.; Galetin, A. Prediction of Human Intestinal First-Pass Metabolism of 25 CYP3A Substrates from In Vitro Clearance and Permeability Data. *Drug Metab. Dispos.* **2010**, *38*, 1147−1158.

(46) Huang, L.; Wang, Y.; Grimm, S. ATP-Dependent Transport of Rosuvastatin in Membrane Vesicles Expressing Breast Cancer Resistance Protein. *Drug Metab. Dispos.* **2006**, *34*, 738−742.

(47) Kim, W.; Benet, L. P-Glycoprotein (P-gp/MDR1)-Mediated Efflux of Sex-Steroid Hormones and Modulation of P-gp Expression in Vitro. *Pharm. Res.* **2004**, *21*, 1284−1293.

(48) Liu, W.; Okochi, H.; Benet, L. Z.; Zhai, S. D. Sotalol Permeability in Cultured-Cell, Rat Intestine, and PAMPA System. *Pharm. Res.* **2012**, *29*, 1768−1774.

(49) Luo, S.; Pal, D.; Shah, S. J.; Kwatra, D.; Paturi, K. D.; Mitra, A. K. Effect of HEPES Buffer on the Uptake and Transport of P-Glycoprotein Substrates and Large Neutral Amino Acids. *Mol. Pharmaceutics* **2010**, *7*, 412−420.

(50) Navarro, C.; González-Álvarez, I.; González-Álvarez, M.; Manku, M.; Merino, V.; Casabó, V. G.; Bermejo, M. Influence of Polyunsaturated Fatty Acids on Cortisol Transport through MDCK and MDCK-MDR1 Cells As Blood-Brain Barrier in Vitro Model. *Eur. J. Pharm. Sci.* **2011**, *42*, 290−299.

(51) Obradovic, T.; Dobson, G.; Shingaki, T.; Kungu, T.; Hidalgo, I. Assessment of the First and Second Generation Antihistamines Brain Penetration and Role of P-Glycoprotein. *Pharm. Res.* **2007**, *24*, 318−327.

(52) Ouyang, H.; Andersen, T.; Chen, W. J. A Comparison of the Effects of P-Glycoprotein Inhibitors on the Blood − Brain Barrier Permeation of Cyclic Prodrugs of an Opioid Peptide (DADLE). *Pharm. Sci* **2009**, *98*, 2227−2236.

(53) Park, M.; Okochi, H.; Benet, L. Active Transport of the Angiotensin-II Antagonist Losartan and Its Main Metabolite EXP 3174 Across MDCK-MDR1 and caco-2 Cell Monolayers. *Arch. Drug Inf.* **2011**, *4*, 1−9.

(54) Soldner, A.; Benet, L. Z.; Mutschler, E. Christians, U. Is Ciprofloxacin a Substrate of P-glycoprotein? *Br. J. Pharmacol.* **2000**, *129*, 1235−43.

(55) Summerfield, S. G.; Read, K.; Begley, D. J.; Obradovic, T.; Hidalgo, I.; Coggon, S.; Lewis, A. V.; Porter, R. A.; Jeffrey, P. Central Nervous System Drug Disposition: The Relationship between in Situ Brain Permeability and Brain Free Fraction. *J. Pharmacol. Exp. Ther.* **2007**, *322*, 205−213.

(56) Tang, F.; Borchardt, R. T. Characterization of the Efflux Transporter(s) Responsible for Restricting Intestinal Mucosa Permeation of an Acyloxyalkoxy-Based Cyclic Prodrug of the Opioid Peptide DADLE. *Pharm. Res.* **2002**, *19*, 780−6.

(57) Tang, F.; Ouyang, H.; Yang, J. Bidirectional Transport of Rhodamine 123 and Hoechst 33342, Fluorescence Probes of the Binding Sites on P-Glycoprotein, Across MDCK-MDR1 Cell Monolayers. *Pharm. Sci.* **2004**, *93*, 1185−94.

(58) Troutman, M.; Thakker, D. R. Novel Experimental Parameters To Quantify the Modulation of Absorptive and Secretory Transport of Compounds by P-Glycoprotein in Cell Culture Models of Intestinal Epithelium. *Pharm. Res.* **2003**, *20*, 1210−1224.

(59) Wang, Q.; Rager, J. D.; Weinstein, K.; Kardos, P. S.; Dobson, G. L.; Li, J.; Hidalgo, I. Evaluation of the MDR-MDCK Cell Line As a Permeability Screen for the Blood-Brain Barrier. *Int. J. Pharm.* **2005**, *288*, 349−359.

(60) Zhang, C.; Kwan, P.; Zuo, Z.; Baum, L. In Vitro Concentration Dependent Transport of Phenytoin and Phenobarbital, but Not Ethosuximide, by Human P-Glycoprotein. *Life Sci.* **2010**, *86*, 899−905.

(61) MarvinSketch, version 5.2.4; ChemAxon: 2009.

(62) Chemspider. http://www.chemspider.com/ (accessed January 2012).

(63) Brincat, J.; Broccatelli, F.; Sabatini, S.; Frosini, M.; Kaatz, G. W.; Cruciani, G.; Carosati, E. Ligand Promiscuity between the Efflux Pumps Human P-Glycoprotein and S. aureus NorA. *ACS Med. Chem. Lett.* **2012**, *3*, 248−251.

(64) Goodford, P. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(65) Cross, S.; Cruciani, G. Molecular Fields in Drug Discovery: Getting Old or Reaching Maturity? *Drug Discovery Today* **2010**, *15*, 23−32.

(66) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original pKa Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172−81.

(67) MoKa, version 1.1.0.; Molecular Discovery Ltd.: London, UK, 2007.

(68) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.

(69) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual Computational Chemistry Laboratory - Design and Description. *J. Comput.-Aided Mol. Des* **2005**, *19*, 453−463.

(70) QuaSAR; Chemical Computing Group: Montreal, Quebec, Canada, 2000.

(71) MACCS Structural Keys; MDL Information System Inc.: San Ramon, CA, 2005.

(72) Curk, T.; Demšar, J.; Xu, Q.; Leban, G.; Petrovič, U.; Bratko, I.; Shaulsky, G.; Zupan, B. Microarray Data Mining with Visual Programming. *Bioinformatics* **2005**, *21*, 396−398.

(73) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing Classifier Performance in R. *Bioinformatics* **2005**, *21*, 3940−3941.

(74) Broccatelli, F.; Mannhold, R.; Moriconi, A.; Giuli, S.; Carosati, E. QSAR Modeling and Data Mining Link Torsades de Pointes Risk to the Interplay of Extent of Metabolism, Active Transport, and hERG Liability. *Mol. Pharmaceutics* **2012**, *9*, 2290−2301.

(75) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches To Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.