JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

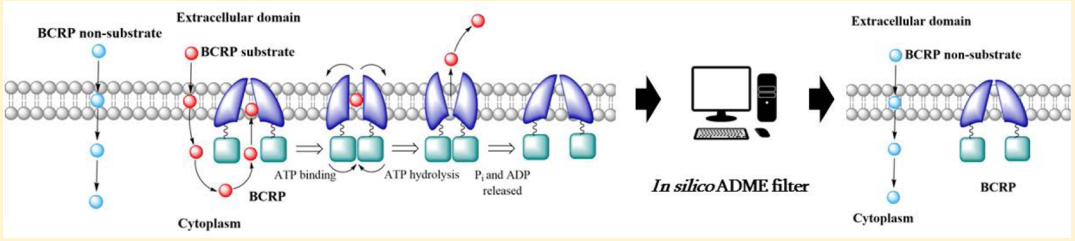
Development and Validation of a Computational Model Ensemble for the Early Detection of BCRP/ABCG2 Substrates during the Drug Design Stage

Melisa E. Gantner,*[,†] Roxana N. Peroni,[‡] Juan F. Morales,[†] María L. Villalba,[†] María E. Ruiz,[†] and Alan Talevi[†]

[†]Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB), Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata (UNLP) − Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), La Plata, B1900AJI Buenos Aires, Argentina

[‡]Instituto de Investigaciones Farmacológicas (ININFA UBA-CONICET), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956 5°, 1113 Ciudad Autónoma de Buenos Aires, Argentina

Supporting Information



ABSTRACT: Breast Cancer Resistance Protein (BCRP) is an ATP-dependent efflux transporter linked to the multidrug resistance phenomenon in many diseases such as epilepsy and cancer and a potential source of drug interactions. For these reasons, the early identification of substrates and nonsubstrates of this transporter during the drug discovery stage is of great interest. We have developed a computational nonlinear model ensemble based on conformational independent molecular descriptors using a combined strategy of genetic algorithms, J48 decision tree classifiers, and data fusion. The best model ensemble consists in averaging the ranking of the 12 decision trees that showed the best performance on the training set, which also demonstrated a good performance for the test set. It was experimentally validated using the *ex vivo* everted rat intestinal sac model. Five anticonvulsant drugs classified as nonsubstrates for BRCP by the model ensemble were experimentally evaluated, and none of them proved to be a BCRP substrate under the experimental conditions used, thus confirming the predictive ability of the model ensemble. The model ensemble reported here is a potentially valuable tool to be used as an *in silico* ADME filter in computer-aided drug discovery campaigns intended to overcome BCRP-mediated multidrug resistance issues and to prevent drug−drug interactions.

INTRODUCTION

Multidrug resistance (MDR) can be defined as the ability of a living cell to show resistance against a broad spectrum of structurally and functionally unrelated drugs.[1] A specific form of MDR is mediated by some members of the ATP-binding cassette (ABC) efflux transporters, which are integral membrane proteins whose main function is to actively translocate ligands across the plasmatic membrane. In eukaryotes, the transport always occurs in the inside-out direction, removing the substrates from the cell or organelle where they are expressed. These efflux transporters are characterized by a broad substrate specificity, and they have an important role in the traffic of a wide and structurally heterogeneous spectrum of endobiotics (e.g., lipids and other physiological compounds) and xenobiotics (e.g., waste products, drugs, and toxic agents).[2] They contribute to protect the body (and particularly sensitive tissues such as the brain), avoiding the entry of possibly toxic compounds or facilitating their elimination in bile, urine, and other fluids.[3] ABC transporters are grouped into seven families (ABCA to ABCG).[4] The ABCB1 protein (also known as MDR1 or P-glycoprotein - Pgp) was the first reported and thus most studied ABC efflux transporter so far; the ABCG2 protein (or Breast Cancer Resistance Protein - BCRP) and various members of the ABCC family (also known as Multidrug Resistance-Associated Proteins - MRPs) have also been associated with MDR in a number of pathologies.

Epilepsy is the most common chronic brain disorder affecting more than 50 million people worldwide.[5,6] Drug therapy is successful in controlling seizures in about 70% of the patients.[6] The remaining 30% suffer from refractory epilepsy[7] that may be

Received: January 9, 2017
Published: July 14, 2017

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

defined as the failure of at least two regimens of properly selected, well tolerated, and commonly used antiepileptic drugs (AEDs) to achieve sustained seizure freedom.[8,9] Among different hypotheses that provide possible explanations to the phenomenon of refractoriness in epilepsy, the transporter hypothesis holds that the pharmacoresistance is a result of the local seizure- or drug-induced overexpression and hyperactivity of MDR-associated ABC transporters at the blood-brain barrier (BBB) and/or the epileptic foci.[10−12] Refractoriness has been mainly attributed to the Pgp overexpression, since various studies indicate that several AEDs are substrates of this transporter.[13−15] Nakanishi et al.[16] demonstrated the involvement of BCRP in reducing brain bioavailability of phenobarbital, clobazam, zonisamide, gabapentin, tiagabine, and levetiracetam in genetically modified mice that lack either Pgp or Pgp and Bcrp. Later, Römermann et al.,[17] using the highly sensitive *in vitro* concentration equilibrium transport assay (CETA) with murine Bcrp1 and human BCRP transfected MDCKII cells, found that lamotrigine is a dual Pgp/BCRP substrate. Interestingly, a number of reports indicate that BCRP is the most abundantly expressed ABC efflux transporter throughout the intestine[18] and the BBB[19,20] of healthy subjects; its involvement in drug interactions explains why the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have recommended evaluating investigational drugs as substrates and/or inhibitors of Pgp and BCRP.[21−23]

While the search of specific inhibitors for MDR-ABC transporters arouses much interest,[21,24−29] previous clinical studies indicate that the inhibition of these transporters may lead to significant adverse reactions,[30−32] a fact that underlines the important physiological role of the ABC transporters, which compromises the potential of their inhibitors as add-on treatments in a long-term therapy scenario. Accordingly, the early recognition of BCRP substrates during the drug design stage is a viable strategy to design novel therapeutics for the treatment of refractory epilepsy and other diseases linked to ABC-mediated MDR issues.

The predictive *in silico* models for BCRP substrates are somewhat limited. Briefly, Hazai et al.[33] used support vector machines (SVM) to build a model to predict wild-type BCRP substrates; Zhong et al.[34] employed genetic algorithm-conjugate gradient-SVM (GA-CG-SVM) to discriminate substrates and nonsubstrates of BCRP; Sedykh et al.[35] developed a set of QSAR models using SVM, random forests (RF), and k-nearest neighbors for identification of both substrates and inhibitors of 11 intestinal transporters, including BCRP; Erić et al.[36] reported artificial neural network- (ANN) and SVM-based model ensembles for the prediction of transport and inhibition of Pgp and BCRP; Garg et al.[37] reported an *in silico* SVM model for the classification of BCRP substrates and nonsubstrates, which can be used in tandem with a second one aimed to estimate the BBB permeability; Lee et al.[38] developed a linear QSAR model to establish the relationship between specificity of BCRP substrates and their uptake rates by BCRP polymorphs. Finally, Ose et al.[39] developed an SVM-based prediction system to predict substrates of 7 categories of drug transporters (among them, BCRP). Noteworthy, the polyspecificity of the ABC transporters makes computational recognition of their substrates quite a challenging task; accordingly, the general trend is to resort to flexible modeling approaches (e.g., nonlinear and locally weighted techniques) to enable more accurate predictions.[40,41]

We have previously reported two linear classifier ensembles for the early recognition of BCRP substrates using conformation independent molecular descriptors, the first one obtained through Stepwise Forward Linear Discriminant Analysis[42] and the second one using Enhanced Replacement Method for variable selection.[43] Here we report the development of an ensemble of nonlinear computational models capable of discriminating between BCRP substrates and nonsubstrates. The ensemble was applied to the classification of an in-house library of drug candidates exhibiting anticonvulsant activity. For experimental validation of the predictions, those drugs that were classified as nonsubstrates for BCRP were evaluated using the *ex vivo* everted rat intestinal sac assay. To our knowledge, this is the first report of an *in silico* predictive model of BCRP substrates with experimental validation.

MATERIALS AND METHODS

***In Silico* Modeling.** *Data Set.* A data set of 262 human wild-type BCRP substrates and nonsubstrates was compiled from the literature. A compound was considered as a substrate only if it is transported by the BCRP and as a nonsubstrate otherwise. Given the variability of experimental conditions used in the literature to establish if a particular compound is or is not a substrate of the BCRP, it was impossible to establish a single cutoff value for the efflux rate, which is why we decided to use the criterion established by the authors in each original study according to their particular experimental conditions. To deal with conflicting reports or ambiguous results, we decided to discard those compounds from the final data set to avoid introducing noise caused by wrongly classified compounds.

The data set was partitioned into a 164-compound training set (composed of 85 substrates and 79 nonsubstrates) and a 98-compound test set (71 substrates and 27 nonsubstrates) by two consecutively clustering algorithms. The Library MCS v0.7 (ChemAxon) hierarchical clustering approach[44,45] was first applied to obtain the seeds for the k-means clustering algorithm[46−48] (in Statistica 10, Statsoft Inc., 2011). With the aim to obtain a balanced and representative training set, after the clustering procedure 50% of each cluster from the substrate category and 75% of each cluster from the nonsubstrate category were randomly assigned to the training set. Using a balanced training set is essential to avoid potential bias toward the prediction of the over-represented category of training examples.[49] The remaining elements of each cluster constitute a representative, independent test set for validation purposes.

Molecular Descriptors Calculation. 867 0-2D Dragon 4.0 (Milano Chemometrics, 2003) molecular descriptors were calculated, and the initial filters provided by Dragon software were used to exclude molecular descriptors with constant or nearly constant values within the training set (identical values for all compounds of the training set, except one) and descriptors with standard deviation below 0.001.

Modeling Procedure. The J48 decision tree-inducing algorithm, the implementation of the C4.5 decision tree algorithm[50] in Weka 3.6,[51] was used to obtain the corresponding nonlinear classifiers. For this purpose, the 867 molecular descriptors were randomized and partitioned into 6 sets of around 150 descriptors each. A nominal binary variable representing the class labels ("substrate" and "nonsubstrate") was used as the dependent variable of the model. We applied genetic algorithms (GA) implemented in Weka 3.6 to preselect the descriptors of each set with best discriminating capacity.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

The GA parameters values used were the following: initial population 100, number of generations 50, probability of crossover 0.6, and probability of mutation 0.033. As the response evaluation function we used the classifier subset evaluator with J48 as the induction algorithm in order to estimate the "merit" of the subset of descriptors, that is, to evaluate the correspondence of a group of descriptors with the class. At this stage (response evaluation function), only online pruning was applied by varying systematically the *M* parameter (the minimum number of instances per leaf) from 2 to 20. We conducted consecutive runs where the result of one run was used as the input for the next run, until convergence was observed.

As a result, for each of the 6 sets of randomized descriptors we obtained 19 different solutions, i.e. 114 subsets of descriptors were selected by GA.

The final models were built by the application of the J48 algorithm on each of these 114 subsets of descriptors. The following methodology was applied:

1) The *M* parameter was systematically varied from 2 to 20. For each value of *M*, the *C* parameter (confidence factor) was systematically varied from 0.001 to 0.5 in 0.01 steps, and for each tree obtained the subtree raising was applied.

2) On the other hand, on each of the 114 sets the reduced-error pruning was applied where the *N* parameter (number of partitions of the training set, where one fold is used for pruning and the rest for growing the tree) was varied systematically from 5 to 15.

Models Performance Evaluation. We resorted to Receiver Operating Characteristic (ROC) curves analysis to evaluate and compare the models' performance.[52] The area under the ROC curve (AUC ROC) allows evaluating if the model performance differs from a random classification and to statistically compare the performances of different models:[52] an ideal model presents an AUC ROC of 1 (equivalent to a perfect classification), while a random classification is represented by a line of slope 1 and corresponds to an AUC of 0.5.

ROC curves were constructed using MedCalc (MedCalc Software, 2011), and for statistical comparison of two AUC ROC the nonparametric method developed by DeLong et al.[53] was used to calculate the standard error of each AUC; the Z-statistic was computed in order to obtain the corresponding p-value.[54]

In Silico Model Validation. For the *in silico* validation of the obtained models we used standard validation techniques to evaluate their robustness and predictive ability, as described next.

Internal Validation. Internal validation was performed through stratified leave-group-out (LGO) cross-validation and Y-randomization test.[55] For the LGO procedure, 10 compounds of the training set (5 substrates and 5 nonsubstrates) were randomly extracted, a process that was repeated 150 times for each model, checking that all the compounds of the training set were removed at least once. For the Y-randomization test, the values of the class label were scrambled across the training set, and 50 randomized models were constructed for each individual model.

External Validation. External validation was performed on the 98-compound test set specifically generated for this purpose during the data set partitioning step.

The experiments conducted by Truchon and Bayly[56] suggest that when enrichment metrics (like AUC ROC) are used to evaluate the models on a small set of compounds, the calculated values have a significant error, which decreases and converges to a constant value for larger databases. Another problem that occurs when working with a high proportion of positive instances is the "saturation effect": once the hit compounds saturate the early part of the ranking the enrichment metric cannot get any higher. This effect is attenuated when the positive/negative instances ratio of the data set is much less than 1. Considering that our *in silico* model will be applied to analyze large chemical libraries, the original 98-compound test set is then not sufficient to assess the actual behavior of the models on a virtual screening application. Therefore, we constructed two large chemical libraries, in which the nonsubstrates/substrates ratio is lower than 0.05.

First we built a pilot library where the original test set was dispersed among 479 putative BCRP substrates, i.e. substrates of nonhuman BCRP homologues or highly similar compounds to known human BCRP substrates which were retrieved from ZINC[57] and PubChem[58] databases through molecular similarity searches (similarity score >0.75 when compared to known substrates). The pilot library obtained in this way (that we will call "simulated library") contains 27 known non-substrates among 550 known or putative substrates, leading to a positive/negative instances ratio smaller than 0.05.

Second, as a final challenge for our models, a larger and structurally diverse library was constructed, using the DUD-E (Enhanced Directory of Useful Decoys[59,60]) resource. This second library, which will be called "DUD-E library" from now on, contains 1346 compounds (1248 decoys plus the original test set) where each decoy is physicochemically similar but topologically dissimilar from the corresponding nonsubstrate. To this end we used the automatic decoys generation tool publicly available online (http://dude.docking.org/generate). Briefly, the decoys are properly matched to the nonsubstrates using molecular weight, a theoretic log transformation of the octanol−water partition coefficient (miLogP), the number of rotatable bonds, hydrogen bond acceptors count, hydrogen bond donors count, and net molecular charge. About 50 decoys were generated for each nonsubstrate by selecting decoys from the ZINC database using in the first place a dynamic protocol that adapts to the local chemical space by narrowing or widening windows around the 6 matching properties; in a second stage, molecular similarity based on ECFP4 fingerprints is calculated. Finally, the decoys are sorted according to the maximum Tanimoto coefficient for each nonsubstrate, and the 25% decoys which are the most topologically dissimilar from the known nonsubstrates were retained.

Data Fusion. Given the broad substrate specificity of the BCRP, which probably arouses from the existence of multiple binding sites,[61−66] we resorted to data fusion to achieve, by consensus, more robust predictions.[67] Behind the use of data fusion underlies a statistical assumption according to which, the more times a molecule is recovered by independent methods or models, the greater the probability that it meets the characteristics of interest. Selective combination (which means the combination of a few but well-performing models) can provide better accuracy and generalization.[67−69] Here we used five data fusion schemes applied on the 12 individual models that presented the best performance on the training set and that also demonstrated a good performance for the test set: the maximum value (MAX operator), the minimum value (MIN operator), the average score, the average ranking, and the average voting between the scores of the models constituting the ensemble. The average ranking consists in generating a

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

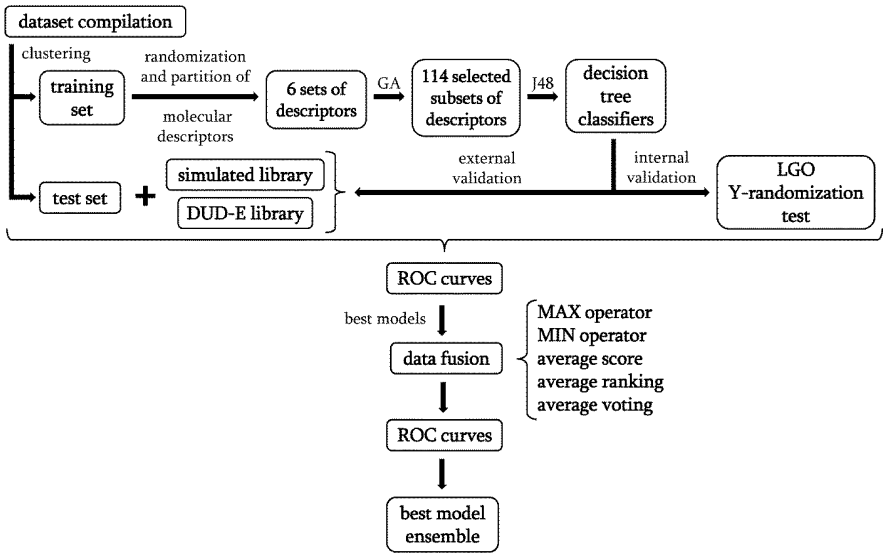


Figure 1. Flowchart of the methodology that has been applied to obtain the model ensemble.

ranking of the compounds based on each model score and then averaging it. The average voting is obtained by the calculation of the vote, which for the *j*-th compound in the *i*-th model is equal to max (0, int(11−rank$_{ij}$/0.02NDB)), where rank$_{ij}$ is the ranking of the *j*-th compound according to the *i*-th model, and NDB is the number of compounds in the entire tested database. The algorithm gives 10 votes in favor of the first 2% of the ranked compounds, 9 votes for the next 2%, and so on. For compounds ranked between 18 and 20% range, 1 vote is given. Compounds in the final 80% of the ranking list receive no votes.[70] The five combination schemes were analyzed and compared through ROC curves.

Figure 1 shows a schematic representation of the modeling procedure followed in the present work.

Experimental Validation. *Selection of the Compounds.* The best ensemble obtained was applied on a small library of 21 compounds (Figure 2) synthesized or acquired by our group (LIDeB - UNLP), with the following characteristics:

1) Proven anticonvulsant activity[71−80]

2) Adequate UV absorptivity for subsequent analytical determination by HPLC-DAD.

We selected 5 compounds among those classified as nonsubstrates by our model ensemble; these 5 compounds fell within the applicability domain of the models that constitute the combination. While the model ensemble may provide some independence from the determination of the applicability domain of each individual model,[81] we have adopted a conservative approach calculating the applicability domain estimated through the leverage approach.[82] It consists in computing the leverage $h_i = x_i^T(X^TX)^{-1}x_i$ for each compound of the database, where x_i is the descriptor vector of the considered compound *i*, and *X* is the model matrix derived from the training set descriptor values. The threshold value is defined as $3k/n$, with *k* being the number of model parameters and *n* being the number of training set compounds.

Animals. Male Wistar rats (280−320 g body weight) were maintained under a 12:12-h light:dark cycle, at controlled room temperature with food and water *ad libitum*. Experiments were conducted in accordance to the Guide for the Care and Use of Laboratory Animals of the National Research Council (USA, 1996) and also in accordance with the guidelines of the 6344/96 regulation of the Argentinean National Drug, Food and Medical Technology Administration (ANMAT).

Experimental Model. The *ex vivo* everted rat intestinal sac model was used for studying BCRP mediated transport for each of the 5 compounds previously selected as BCRP nonsubstrates.[83,84] The main advantage of this technique with respect to *in vitro* assays is that the results obtained using *ex vivo* techniques often match those obtained from *in vivo* studies.[85−89] The BCRP is expressed in all segments of the rat small intestine from the duodenum and proximal jejunum to the distal jejunum and the ileocecal valve, including the ileum. The highest levels of BCRP expression are observed in the segment including the distal jejunum and ileum, and therefore this portion of the intestine was used to study the possible interaction of anticonvulsant drugs with the BCRP.[84,90]

Evaluation of Drug Transport Across the Everted Rat Intestinal Sacs. Briefly, in each trial the rats were anesthetized with urethane (1.2 g/kg, i.p. injection), the abdomen was opened, and the distal extreme of the small intestine was removed and placed in a chamber with Krebs buffer (in mM: NaCl 118; KCl 4.7; $MgCl_2$ 1.2; NaH_2PO_4 1.0; $CaCl_2$ 2.6; $NaHCO_3$ 25.0; glucose 11.1; sodium ethylenediamine tetraacetic acid (Na_2 EDTA) 0.004; final pH 7.4), under bubbling with 95% O_2/5% CO_2. The intestine was gently everted on a glass rod, and six sacs (5−7 cm) were prepared. To generate the sacs, one end of each segment of intestine was firmly ligated with thread, while the other end was also tightly tied with thread to a short cannula where the drug solutions were administered. A thin wire steel inserted into the cannula cap acts as a hook that allows hanging the intestinal sac in the container where the test was conducted. The six sacs were randomly immersed into the containers with 5 mL of Krebs buffer prewarmed at 37 °C under bubbling with 95% O_2/5% CO_2. After 15 min stabilization, the trial was initiated by introducing the 0.7 mL of drug solution within each sac with a syringe. This time point was considered time 0. To evaluate transporter inhibition, the inhibitor was added to the medium containing the corresponding sac 30 min before the addition of the drug solution. The transport of drugs across the intestine from the serosal to mucosal surface was evaluated by sampling the medium (150 μL with replacement) every 5 min during 30

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

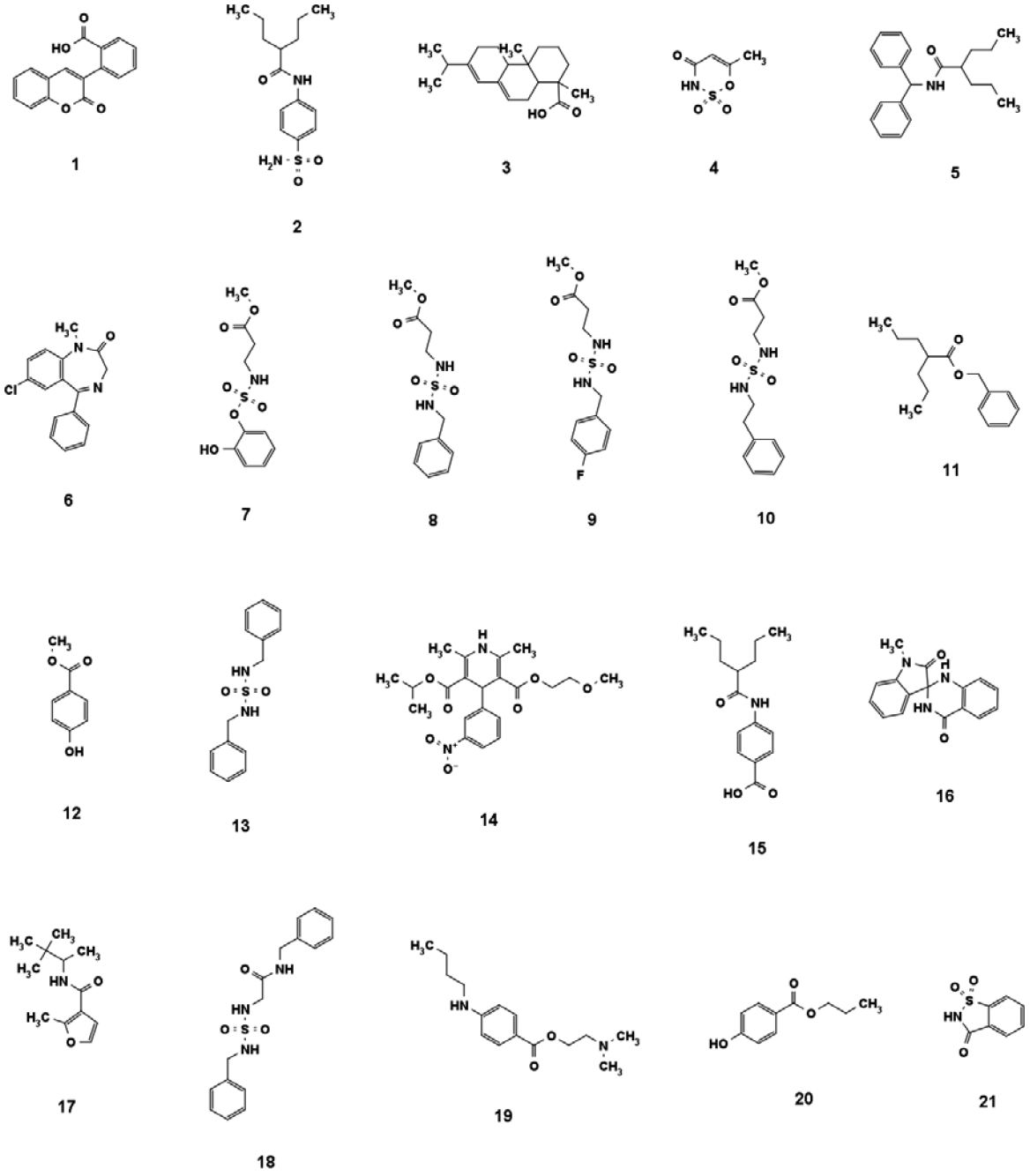


Figure 2. Chemical structures of the 21-compound library on which we applied the best model ensemble.

min. We also performed a control at 4 °C in order to evaluate a possible BCRP-independent transport of the drugs across the rat intestine. Figure 3 shows a diagram of the procedure carried out during the everted rat intestinal sac experiment, while Figure 4 shows the location of the BCRP in the intestine before and after the eversion.

To check the active transport across the rat intestine mediated by BCRP, we used the couple substrate-inhibitor of BCRP nitrofurantoin (NF)[91]-Ko143.[92] Three concentrations of NF (1, 10, and 100 μM) were evaluated to establish the optimal working concentration at which active transport is analytically observed to be used as positive control. Once established the optimal NF concentration, two inhibitor concentrations (10 and 50 μM) were evaluated in order to establish the optimal inhibitor concentration at which the transport inhibition is observed.

The evaluation of the selected anticonvulsant drugs was performed through two protocols:

Protocol 1: Evaluation of the potential effects of the drugs on the transporter by measuring the transport of NF in the presence of each drug. Each drug was added to the buffer 30 min before starting the test. The sampling of the medium was performed every 5 min during 30 min. The already established optimal NF concentration was used, and two concentrations of each drug were evaluated corresponding to the optimal concentration of the inhibitor and NF, respectively.

Protocol 2: Transport assessment of the drugs in the presence and absence of the specific BCRP inhibitor Ko143. It was only possible to evaluate in this second protocol those

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

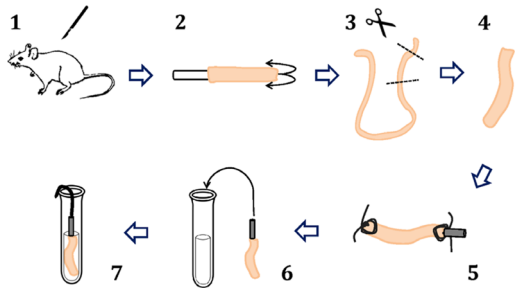


Figure 3. Schematic representation of the steps followed during the everted rat intestinal sacs experiment. Step 1: After anesthetizing the rat, the abdomen is opened, and the distal end of the small intestine is removed. Step 2: The intestine is gently everted onto a glass rod. Steps 3 and 4: The intestine is divided into segments of 5−7 cm each. Step 5: To produce the sacs, one end of each segment of intestine is firmly tied with string, while the other end is tied with thread to a short cannula where the drugs solutions are administered. Steps 6 and 7: A thin wire steel inserted into the cannula cap acts as a hook to hang the everted intestinal sac in the container where the trial is performed.

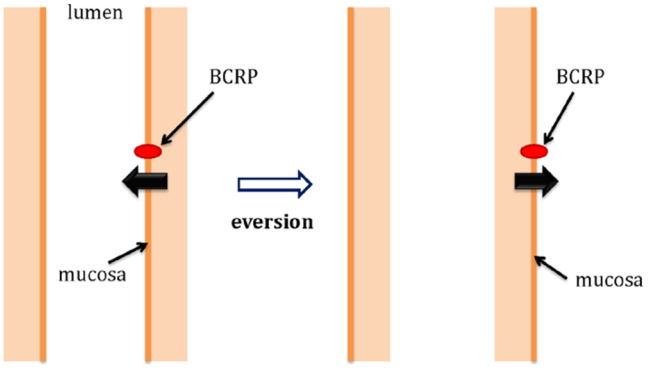


Figure 4. BCRP is localized in the intestinal lumen. In the everted intestinal sac assay the mucosa is exposed to the outside, and for this reason it is expected that the BCRP substrates introduced inside the sacs (serosa) are transported to the outside, i.e. from the serosa to the mucosa.

drugs that were within the detection limit of the analytical quantification method (HPLC-DAD). Each drug was individually evaluated following the protocol established for NF; that is, the drug solution was injected into the sac, and the same sampling scheme was followed. The assayed concentration for each drug and for the inhibitor corresponded to the optimum concentration established for NF and the inhibitor, respectively.

Drugs. NF was cordially donated by Laboratorios Bagó S.A., 2-(2-oxo-2H-chromen-3-yl)benzoic acid (Compound **1**, Figure 2) was purchased from InterBioScreen Ltd., 1-methyl-1,2,3′,4′-tetrahydro-1′H-spiro[indole-3,2′-quinazoline]-2,4′-dione (Compound **16**, Figure 2) was purchased from Princeton Biomolecular Research, *N*-(3,3-dimethylbutan-2-yl)-2-methyl-furan-3-carboxamide (Compound **17**, Figure 2) was acquired from UkrOrgSyntez (UORSY) Ltd., *N*,*N*′-dibenzylsulfamide (Compound **13**, Figure 2) was synthesized in our laboratory,[80] and methylparaben (Compound **12**, Figure 2) and Ko143 were purchased from Sigma-Aldrich Argentina.

Quantitative Analysis by HPLC. An analytical HPLC-DAD method was developed and validated for the quantitation of the evaluated drugs in the matrices from the *ex vivo* studies. Protein precipitation was performed by adding one volume of acetonitrile on each sample followed by centrifugation. An appropriate internal standard was used in each case. All samples were quantified using an UHPLC Dionex Ultimate 3000 (Thermo Scientific, Dionex, Sunnyvale, California, USA) equipped with a diode array detector. The stationary phase was a Luna RP18 column (5 μm, 150 × 4.6 mm) (Phenomenex, Torrance, CA, USA), operated at room temperature. In all cases the mobile phase was composed of a mixture of 20 mM KH_2PO_4 buffer (adjusted to pH 2.5 with H_3PO_4) and methanol in different proportions as required.

For quantification, the linearity, specificity, precision, and accuracy in the working range concentrations of each drug were demonstrated.

Statistical Analysis of Experimental Results. We verified the assumptions required to perform parametric tests: the randomness of residuals, the normality of the data, and the homogeneity of variance. The logarithmic transformation of the data was applied in those cases where it was necessary. Finally, we applied the Student's t test or the factorial analysis of variance (one-way ANOVA).

Table 1. Overall Accuracy, Sensitivity (Se), and Specificity (Sp) for the Training and the Test Sets and the Results of the Internal Validation for the 12 Best Models Obtained[a]

model	overall accuracy training set	Se training set	Sp training set	overall accuracy test set	Se test set	Sp test set	LGO[b]	Y-randomization[c]
1	73.78	65.82	81.18	73.47	62.96	77.46	63.93 (13.08)	58.98 (2.33)
2	95.73	98.73	92.94	73.47	77.78	71.83	68.55 (16.13)	56.75 (5.73)
3	93.29	91.14	95.29	74.49	81.48	71.83	62.81 (14.56)	56.16 (4.81)
4	95.12	96.20	94.12	69.39	70.37	69.01	68.37 (14.34)	57.16 (4.51)
5	96.95	97.47	96.47	68.37	70.37	67.61	69.95 (13.95)	58.16 (4.66)
6	96.34	97.47	95.29	67.34	70.37	66.20	70.60 (14.14)	58.53 (5.06)
7	92.07	92.40	91.77	76.53	74.07	77.46	68.22 (12.33)	57.69 (5.76)
8	83.54	72.15	94.12	78.57	62.96	84.51	68.57 (12.31)	57.38 (5.67)
9	78.05	72.15	83.53	75.51	85.19	71.83	65.93 (12.97)	54.12 (4.61)
10	91.46	89.87	92.94	75.51	74.07	76.06	64.19 (13.98)	54.44 (4.59)
11	82.32	68.35	95.29	77.55	59.26	84.51	66.57 (13.87)	56.64 (4.96)
12	91.46	91.14	91.77	75.51	74.07	76.06	67.92 (14.28)	56.71 (4.90)

[a]A score cutoff value of 0.5 was considered here to differentiate substrates and nonsubstrates. [b]The results are presented as the average result of 150 replications and the standard deviation between parentheses. [c]The results are presented as the average performance of the 50 randomized models and the standard deviation between parentheses.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

Table 2. AUC ROC Values Accompanied by Their 95% Confidence Interval between Parentheses for the Training and Test Sets, the Simulated Library, and the DUD-E Library for the 12 Best Models Obtained

model	AUC ROC training set	AUC ROC test set	AUC ROC simulated library	AUC ROC DUD-E library
1	0.785 (0.714−0.845)	0.737 (0.638−0.821)	0.696 (0.657−0.734)	0.615 (0.588−0.641)
2	0.992 (0.963−0.999)	0.791 (0.697−0.866)	0.735 (0.697−0.771)	0.675 (0.649−0.700)
3	0.968 (0.928−0.989)	0.829 (0.740−0.898)	0.734 (0.696−0.770)	0.701 (0.676−0.726)
4	0.993 (0.964−1.000)	0.788 (0.694−0.864)	0.718 (0.679−0.754)	0.598 (0.571−0.625)
5	0.996 (0.971−1.000)	0.785 (0.691−0.862)	0.716 (0.678−0.753)	0.590 (0.563−0.617)
6	0.995 (0.968−1.000)	0.788 (0.694−0.864)	0.718 (0.680−0.755)	0.598 (0.571−0.624)
7	0.969 (0.930−0.990)	0.856 (0.771−0.919)	0.780 (0.744−0.813)	0.809 (0.787−0.830)
8	0.867 (0.806−0.915)	0.744 (0.646−0.827)	0.725 (0.686−0.761)	0.742 (0.718−0.766)
9	0.815 (0.747−0.871)	0.880 (0.799−0.937)	0.820 (0.786−0.851)	0.651 (0.625−0.677)
10	0.965 (0.924−0.987)	0.850 (0.764−0.914)	0.758 (0.721−0.792)	0.757 (0.733−0.779)
11	0.843 (0.778−0.895)	0.728 (0.628−0.813)	0.707 (0.668−0.744)	0.681 (0.655−0.705)
12	0.962 (0.921−0.986)	0.839 (0.751−0.905)	0.795 (0.760−0.827)	0.756 (0.732−0.779)

RESULTS AND DISCUSSION

***In Silico* Modeling.** From a comprehensive analysis through the variation of the J48 algorithm run parameters (as detailed in the previous section) for the 114 subsets of molecular descriptors preselected by GA, we chose the 12 models which showed the best performance on the training set considering the overall accuracy, the sensitivity (true positive rate, Se), the specificity (true negative rate, Sp), the AUC ROC, and the internal validation (LGO and Y-randomization) results. We also verified that these 12 models demonstrate a good performance on the test set. The diagrams of the 12 decision trees are presented in the Supporting Information. The reader can also observe the descriptors of each model and the cutoff values of each node, together with the RF variable importance measures (mean decrease in accuracy and mean decrease in node impurity) for each descriptor of the model ensemble. Table 1 shows the overall accuracy, the Se and Sp for the training and the test sets, and the results of the internal validation; whereas Table 2 shows the results of the AUC ROC for the training and the test sets, the simulated library, and the DUD-E library for the 12 selected models.

Table 1 shows that the overall accuracy is lower for the test set than for the training set, and the same behavior can be observed for the AUC ROC in Table 2, suggesting some degree of overfitting which is frequently observed in decision trees and, in general, in flexible methods. According to Table 2, there is a general trend to loss predictivity when we move from the test set to the DUD-E library. While all the models displayed acceptable results in the internal cross-validation, it can be observed that when the number of descriptors incorporated into the model increases, the difference between the percentage of overall good classifications for the training set and for the instances removed during the LGO validation process also increases following a linear trend (see Table 3 and Figure 5). The J48 algorithm has a strong tendency to overfitting which is accentuated with the number of independent variables in the model (which is a general problem of highly flexible methods[93,94]); due to this behavior we put great emphasis on the pruning tools available to control this trend.

In addition, the J48 algorithm has some degree of instability, which means that a small variation in the training set may lead, in certain cases, to a very different decision tree.[94,95] This emerges directly from the divisive hierarchical process by which the decision tree is constructed. At this point, it is important to note the great structural variability of the compounds of the database which in part emerges from the known broad substrate specificity of BCRP. Data fusion techniques appear as a possible solution to the instability and the tendency to overfitting of the J48 algorithm.[94,95] Accordingly, we generated the combinations of the 12 best models (selective ensemble) using the five data fusion schemes described in the Materials and Methods section. The results are shown in Table 4.

Table 3. Difference between the Percentages (%) of Overall Good Classifications for the Training Set and the Removed Instances on the LGO Internal Validation Depending on the Number of Descriptors Incorporated into the Model

model	no. of descriptors	no. of instances training set/no. of model descriptors	difference between % overall good classifications for training set and LGO validation
1	5	32.8	9.85
9	6	27.33	12.12
11	6	27.33	15.74
8	7	23.42	14.96
7	9	18.22	23.85
10	9	18.22	27.27
12	9	18.22	23.54
4	10	16.40	26.75
6	10	16.40	25.73
5	11	14.91	26.99
2	12	13.66	27.18
3	13	12.61	30.48

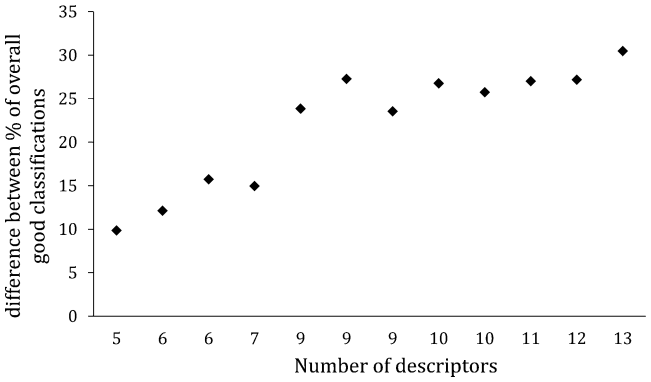


Figure 5. Apparent linear evolution of the difference between the percentages of overall good classifications for the training set and the instances removed during LGO validation (y axis) vs the number of descriptors incorporated into the model (x axis). A clear trend to overfitting is observed with the increasing the number of descriptors added to the model.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

Table 4. AUC ROC Values Accompanied by Their 95% Confidence Interval between Parentheses for the Training Set, the Test Set, the Simulated Library, and the DUD-E Library for the Five Data Fusion Schemes Applied to the 12 Best Models Obtained

ensemble	AUC ROC training set	AUC ROC test set	AUC ROC simulated library	AUC ROC DUD-E library
MAX operator	0.978 (0.942−0.994)	0.845 (0.757−0.910)	0.741 (0.704−0.777)	0.668 (0.642−0.693)
MIN operator	0.986 (0.954−0.998)	0.825 (0.735−0.894)	0.797 (0.762−0.829)	0.786 (0.763−0.808)
average score	0.994 (0.967−1.000)	0.890 (0.811−0.944)	0.826 (0.792−0.856)	0.779 (0.756−0.801)
average ranking	0.997 (0.973−1.000)	0.945 (0.880−0.981)	0.840 (0.808−0.869)	0.801 (0.779−0.822)
average voting	0.952 (0.907−0.979)	0.956 (0.894−0.987)	0.834 (0.801−0.864)	0.818 (0.796−0.838)

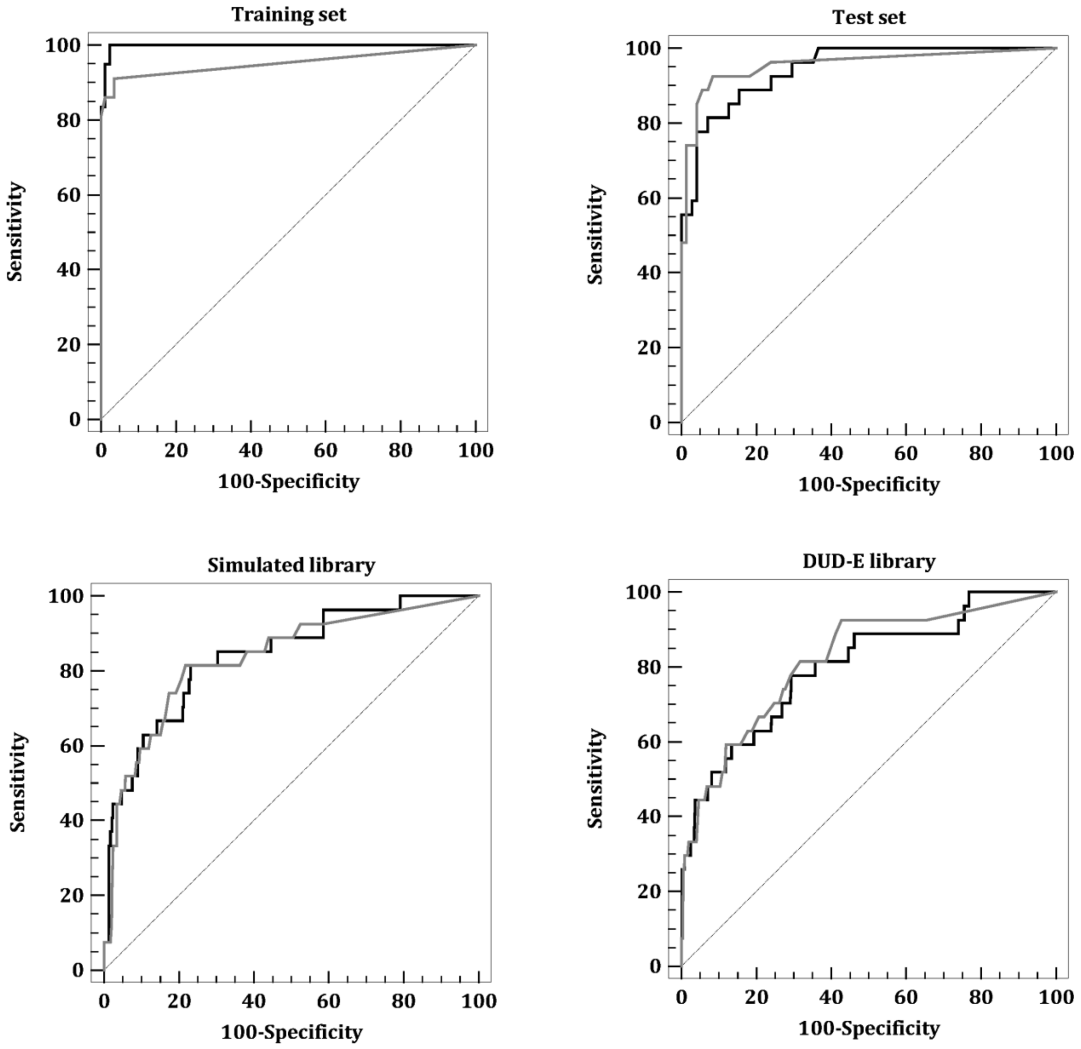


Figure 6. ROC curves for the training set, the test set, the simulated library, and the DUD-E library for both the average ranking (black) and the average voting (gray) of the 12 best nonlinear models obtained. The similar performance of both data fusion schemes is evident. The average ranking was chosen for experimental validation because it presented a better performance for the training set and it is a much simpler data fusion scheme.

According to Table 4, the DUD-E library is the more challenging external validation step and presents the worst performance among all the sets of compounds used for that purpose. Nonetheless, a good performance for all data fusion schemes along all libraries used in the external validation is observed (with the exception of the MAX operator, which leads to an AUC ROC below 0.7 for the DUD-E library). Regarding the training set, the average ranking presents the greater AUC ROC outperforming the average voting (p = 0.0055), while there are no statistically significant differences between the average ranking and the remaining data fusion schemes. For the test set, the average ranking and the average voting have the higher AUC ROC with no statistically significant differences between them ($p > 0.05$), while both are slightly different from the remaining three schemes ($p < 0.03$). The maximum AUC ROC for the simulated library corresponds to the average ranking, showing no differences with the remaining combinations ($p > 0.05$). For the DUD-E library, the highest AUC ROC value corresponds to the average voting, displaying no significant differences with the other combinations except the MAX operator ($p < 0.0001$).

The average ranking and the average voting consistently show the best performances across all libraries examined during the external validation, and there is no statistically significant difference between them with the exception of the training set, where the average ranking was higher. According to this, we

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

decided to use the average ranking to move to the experimental validation stage also considering that is a simpler data fusion scheme.

Figure 6 shows the ROC curves for the training set, the test set, the simulated library, and the DUD-E library for both the average ranking and the average voting, evidencing a very similar performance.

Experimental Validation. *Drugs Selected for Evaluation.* The 5 anticonvulsant compounds classified as nonsubstrates by our model ensemble and selected for experimental evaluation were methylparaben (Compound **12**, Figure 2),[79] 2-(2-oxo-2H-chromen-3-yl)benzoic acid (Compound **1**, Figure 2), 1-methyl-1,2,3′,4′-tetrahydro-1′H-spiro[indole-3,2′-quinazoline]-2,4′-dione (Compound **16**, Figure 2), *N*-(3,3-dimethylbutan-2-yl)-2-methylfuran-3-carboxamide (Compound **17**, Figure 2),[71,72] and *N*,*N*′-dibenzylsulfamide (Compound **13**, Figure 2).[80]

Experimental Model Validation. We verified the active transport of NF through the tissue and its inhibition in the presence of Ko143. The results are shown in Figure 7. Across

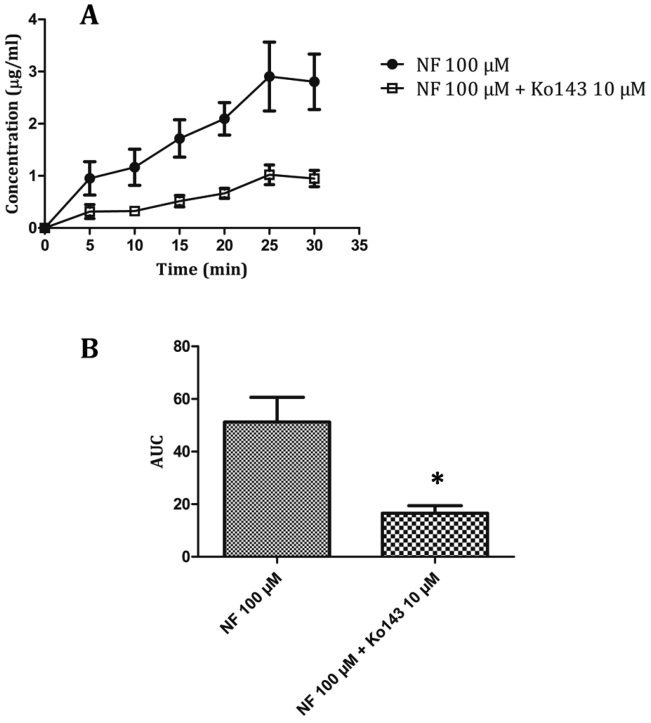


Figure 7. Results obtained for NF 100 μM evaluated at 37 °C with and without Ko143 10 μM (n = 9). **A.** Concentration versus time curves. **B.** Area under the concentration versus time curves (AUC). Vertical segments indicate standard error of the mean (SEM).

all the concentrations evaluated, 100 μM for NF and 10 μM for Ko143 were the optimal concentrations at which both, active transport and inhibition were best observed analytically. We found statistically significant differences (p = 0.000343) between the area under the concentration versus time curve (AUC) of NF at 100 μM with and without Ko143 10 μM. Inhibition of the transport by Ko143 confirms that BCRP is involved in the active transport of NF in our experimental conditions.

As shown in Figure 7, the greater data dispersion was observed in the curve corresponding to NF 100 μM, while the dispersion was much lower in the presence of the inhibitor. This may be due to intraindividual (different portion of intestine) as well as interindividual variability in the BCRP expression levels observed *in vivo*, affecting the active transport of NF, which is abolished in the presence of the specific inhibitor.

Evaluation of the Selected Drugs According to Protocol 1. Active transport of NF was evaluated in the presence of the 5 drugs predicted as nonsubstrates by the model ensemble, which were dissolved in the external media at 10 and 100 μM. Such concentrations were established according to the optimal concentrations found during the model validation for the inhibitor (10 μM) and NF (100 μM). The results obtained are shown in Figure 8. No statistically significant differences were found between the AUC of NF in the presence and absence of the 5 drugs at both 10 and 100 μM ($p > 0.05$ in all cases). The results indicate that none of the evaluated drugs significantly interfere, at the concentrations tested, with the rat BCRP-mediated transport of NF. This finding adds evidence to the hypothesis that none of the 5 drugs is a substrate of BCRP, at least not for the same binding site or with similar affinity for BRCP than NF. For confirmation, we tested 3 of the 5 drugs (Compounds **1**, **12**, and **16**) following protocol 2.

Evaluation of the Selected Drugs According to Protocol 2. The transport across the intestine for Compounds **1**, **12**, and **16** was evaluated at 100 μM (same concentration than the optimal concentration found for NF) at 37 and 4 °C in order to check a possible active transport through the tissue and at 37 °C in the presence and absence of Ko143 10 μM to assess the possible transport of the drugs mediated by the BCRP. The results are shown in Figure 9. No statistically significant differences between the AUC for the 3 tested drugs were found in any case ($p > 0.05$ in all cases).

We can conclude that none of the 3 drugs is actively transported across the rat intestine in the assayed conditions; furthermore, no evidence of BCRP mediated transport was observed.

The evidence of both protocols agrees with the prediction made by the average ranking of the 12 best decision trees, according to which none of the 5 drugs would be a transportable BCRP-substrate.

CONCLUSIONS

We have developed a computational nonlinear model ensemble based on conformational independent molecular descriptors for the early identification of substrates and nonsubstrates of BCRP efflux transporter, a protein linked to MDR-phenomena in diseases such as epilepsy and cancer. The model ensemble is easy and quick to use because no previous conformational analysis of the chemical structures to be evaluated is required, which is particularly suitable for virtual screening campaigns on large chemical libraries. All generated models were derived from a relatively large and highly structurally diverse data set, which was divided into representative training and test sets through a rational clustering procedure leading to an adequate balance between the number of substrates and nonsubstrates on the training set. According to the results obtained during the *in silico* modeling step, the importance of using pilot libraries of greater size and structural diversity to assess model behavior can be highlighted. The broad substrate specificity of BCRP increases the difficulty of finding a single model capable of achieving good prediction rates for both, substrates and nonsubstrates, a fact that justified the application of more complex strategies such as nonlinear modeling, data fusion, and, particularly, selective ensemble.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

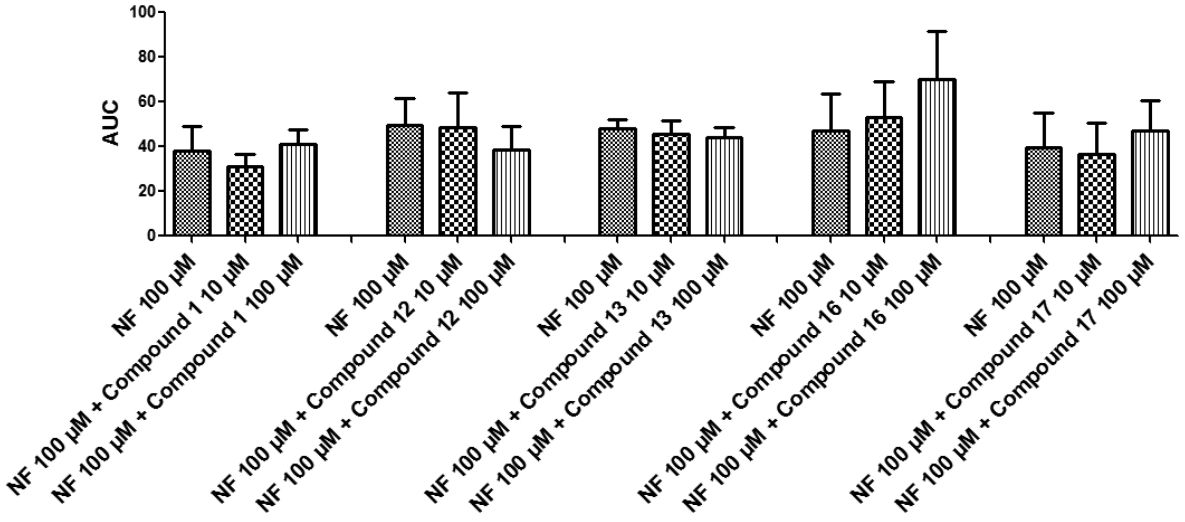


Figure 8. AUC for NF 100 μM evaluated at 37 °C in the absence and presence of the 5 tested drugs at 10 and 100 μM concentrations (n = 6). Vertical segments indicate SEM.

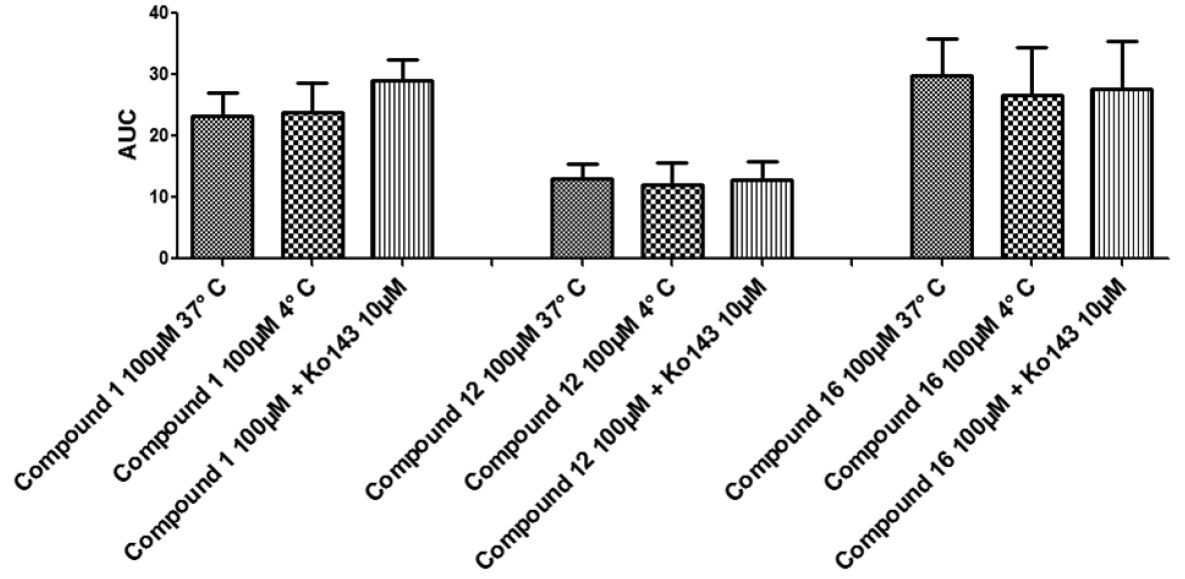


Figure 9. AUC for Compounds **1**, **12**, and **16** 100 μM evaluated at 4 °C and at 37 °C with and without Ko143 10 μM (n = 9). Vertical segments indicate SEM.

The experimental validation of the ensemble predictions was performed using the *ex vivo* everted rat intestinal sac model. The evidence of the first protocol for the 5 drugs evaluated (Compounds **1**, **12**, **13**, **16**, and **17**) together with the evidence of the second protocol for Compounds **1**, **12**, and **16** seems to support the predictions made by the average ranking of the 12 best decision trees, according to which none of the 5 assayed drugs would be a transportable substrate of BCRP. These results demonstrate the predictive ability of the computational model ensemble reported, suggesting that it is a potentially valuable tool to be used as an *in silico* ADME filter in computer-aided drug discovery oriented to overcome BCRP-mediated MDR problems, e.g. for the treatment of refractory epilepsy. As an additional advantage, the model ensemble allows the prediction of potential drug−drug interactions associated with simultaneous administration of two or more drugs that are BCRP substrates, inhibitors, or inducers.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00016.

Diagrams of the 12 best decision trees, where the descriptors of each model and the cutoff values of each node are detailed; nomenclature of Dragon 4.0 for molecular descriptors was retained; values of the RF variable importance measures (mean decrease in accuracy and mean decrease in node impurity, i.e. mean decrease Gini) for each descriptor of the model ensemble calculated using the R package randomForest (Table S1) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +542214235333ext 41. E-mail: melisagantner@gmail.com; mgantner@biol.unlp.edu.ar.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

ORCID
Melisa E. Gantner: 0000-0001-7491-4268

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

M.E.G. and M.L.V. are postdoctoral fellowship holders of CONICET. R.N.P., M.E.R., and A.T. are members of the Scientific Research Career at CONICET. J.F.M. is a CONICET fellowship holder. The authors would like to thank UNLP (Incentivos X-597), CONICET (PIP 11220090100603), and ANPCyT (PICTs 2010-2531 and 2010-1774, PPL 2011-0003) for providing funds to develop our research.

■ ABBREVIATIONS

BCRP, Breast Cancer Resistance Protein; MDR, multidrug resistance; ABC, ATP-binding cassette; Pgp, P-glycoprotein; MRPs, Multidrug Resistance-Associated Proteins; AEDs, antiepileptic drugs; BBB, blood-brain barrier; CETA, concentration equilibrium transport assay; SVM, support vector machines; GA-CG-SVM, genetic algorithm-conjugate gradient-support vector machines; RF, Random Forest; ANN, artificial neural network; GA, genetic algorithms; ROC, Receiver Operating Characteristic; AUC ROC, area under the Receiver Operating Characteristic curve; LGO, leave-group-out; DUD-E, Enhanced Directory of Useful Decoys; NF, nitrofurantoin; Se, sensitivity; Sp, specificity; AUC, area under the concentration versus time curve; SEM, standard error of the mean

■ REFERENCES

(1) Sheps, J.; Ling, V. Introduction: What Is Multidrug Resistance? In *ABC Transporters and Multidrug Resistance*; Boumendjel, A., Boutonnat, J., Robert, J., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009; pp 1−459, DOI: 10.1002/9780470495131.ch.
(2) Rees, D. C.; Johnson, E.; Lewinson, O. ABC Transporters: The Power to Change. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (3), 218−227.
(3) Voloshyna, I.; Reiss, A. B. The ABC Transporters in Lipid Flux and Atherosclerosis. *Prog. Lipid Res.* **2011**, *50* (3), 213−224.
(4) Vasiliou, V.; Vasiliou, K.; Nebert, D. W. Human ATP-Binding Cassette (ABC) Transporter Family. *Hum. Genomics* **2009**, *3* (3), 281−290.
(5) Thurman, D. J.; Beghi, E.; Begley, C. E.; Berg, A. T.; Buchhalter, J. R.; Ding, D.; Hesdorffer, D. C.; Hauser, W. A.; Kazis, L.; Kobau, R.; et al. Standards for Epidemiologic Studies and Surveillance of Epilepsy. *Epilepsia* **2011**, *52* (SUPPL. 7), 2−26.
(6) World Health Organisation. WHO | Epilepsy.http://www.who.int/mediacentre/factsheets/fs999/en/ (accessed July 25, 2017).
(7) Brodie, M. J.; Barry, S. J. E.; Bamagous, G. A.; Norrie, J. D.; Kwan, P. Patterns of Treatment Response in Newly Diagnosed Epilepsy. *Neurology* **2012**, *78* (20), 1548−1554.
(8) Kwan, P.; Arzimanoglou, A.; Berg, A. T.; Brodie, M. J.; Hauser, W. A.; Mathern, G.; Moshé, S. L.; Perucca, E.; Wiebe, S.; French, J. Definition of Drug Resistant Epilepsy: Consensus Proposal by the Ad Hoc Task Force of the ILAE Commission on Therapeutic Strategies. *Epilepsia* **2010**, *51* (6), 1069−1077.
(9) Sinha, S.; Siddiqui, K. A. Definition of Intractable Epilepsy. *Neurosciences* **2011**, *16* (1), 3−9.
(10) Aronica, E.; Sisodiya, S. M.; Gorter, J. A. Cerebral Expression of Drug Transporters in Epilepsy. *Adv. Drug Delivery Rev.* **2012**, *64* (10), 919−929.
(11) Loscher, W.; Potschka, H. Drug Resistance in Brain Diseases and the Role of Drug Efflux Transporters. *Nat. Rev. Neurosci.* **2005**, *6* (8), 591−602.
(12) Loscher, W.; Potschka, H. Role of Drug Efflux Transporters in the Brain for Drug Disposition and Treatment of Brain Diseases. *Prog. Neurobiol.* **2005**, *76* (1), 22−76.
(13) Luna-Tortós, C.; Fedrowitz, M.; Löscher, W. Several Major Antiepileptic Drugs Are Substrates for Human P-Glycoprotein. *Neuropharmacology* **2008**, *55* (8), 1364−1375.
(14) Löscher, W.; Luna-Tortós, C.; Römermann, K.; Fedrowitz, M. Do ATP-Binding Cassette Transporters Cause Pharmacoresistance in Epilepsy? Problems and Approaches in Determining Which Antiepileptic Drugs Are Affected. *Curr. Pharm. Des.* **2011**, *17* (26), 2808−2828.
(15) Zhang, C.; Kwan, P.; Zuo, Z.; Baum, L. The Transport of Antiepileptic Drugs by P-Glycoprotein. *Adv. Drug Delivery Rev.* **2012**, *64* (10), 930−942.
(16) Nakanishi, H.; Yonezawa, A.; Matsubara, K.; Yano, I. Impact of P-Glycoprotein and Breast Cancer Resistance Protein on the Brain Distribution of Antiepileptic Drugs in Knockout Mouse Models. *Eur. J. Pharmacol.* **2013**, *710* (1−3), 20−28.
(17) Römermann, K.; Helmer, R.; Löscher, W. The Antiepileptic Drug Lamotrigine Is a Substrate of Mouse and Human Breast Cancer Resistance Protein (ABCG2). *Neuropharmacology* **2015**, *93*, 7−14.
(18) Tucker, T. G.; Milne, A. M.; Fournel-Gigleux, S.; Fenner, K. S.; Coughtrie, M. W. Absolute Immunoquantification of the Expression of ABC Transporters P-Glycoprotein, Breast Cancer Resistance Protein and Multidrug Resistance-Associated Protein 2 in Human Liver and Duodenum. *Biochem. Pharmacol.* **2012**, *83* (2), 279−285.
(19) Dauchy, S.; Dutheil, F.; Weaver, R. J.; Chassoux, F.; Daumas-Duport, C.; Couraud, P. O.; Scherrmann, J. M.; De Waziers, I.; Declèves, X. ABC Transporters, Cytochromes P450 and Their Main Transcription Factors: Expression at the Human Blood-Brain Barrier. *J. Neurochem.* **2008**, *107* (6), 1518−1528.
(20) Uchida, Y.; Ohtsuki, S.; Katsukura, Y.; Ikeda, C.; Suzuki, T.; Kamiie, J.; Terasaki, T. Quantitative Targeted Absolute Proteomics of Human Blood-Brain Barrier Transporters and Receptors. *J. Neurochem.* **2011**, *117* (2), 333−345.
(21) Cherigo, L.; Lopez, D.; Martinez-Luis, S. Marine Natural Products as Breast Cancer Resistance Protein Inhibitors. *Mar. Drugs* **2015**, *13* (4), 2010−2029.
(22) FDA. Guidance for Industry. Drug Interaction Studies: Study Design, Data Analysis, Implications for Dosing, and Labeling Recommendations. *Guid. Doc.* 2012, No. February, 79.
(23) EMA. Guideline on the Investigation of Drug Interactions. *Guid. Doc.* 2012, No. June, 59.
(24) Ding, Y. L.; Shih, Y. H.; Tsai, F. Y.; Leong, M. K. In Silico Prediction of Inhibition of Promiscuous Breast Cancer Resistance Protein (BCRP/ABCG2). *PLoS One* **2014**, *9* (3), e90689.
(25) Shityakov, S.; Förster, C. In Silico Structure-Based Screening of Versatile P-Glycoprotein Inhibitors Using Polynomial Empirical Scoring Functions. *Adv. Appl. Bioinforma. Chem.* **2014**, *7* (1), 1−9.
(26) Montanari, F.; Ecker, G. F. BCRP Inhibition: From Data Collection to Ligand-Based Modeling. *Mol. Inf.* **2014**, *33* (5), 322−331.
(27) Mudududdla, R.; Guru, S. K.; Wani, A.; Sharma, S.; Joshi, P.; Vishwakarma, R. A.; Kumar, A.; Bhushan, S.; Bharate, S. B. 3-(Benzo[d][1,3]dioxol-5-Ylamino)-N-(4-Fluorophenyl)thiophene-2-Carboxamide Overcomes Cancer Chemoresistance via Inhibition of Angiogenesis and P-Glycoprotein Efflux Pump Activity. *Org. Biomol. Chem.* **2015**, *13* (14), 4296−4309.
(28) Belekar, V.; Lingineni, K.; Garg, P. Classification of Breast Cancer Resistant Protein (BCRP) Inhibitors and Non-Inhibitors Using Machine Learning Approaches. *Comb. Chem. High Throughput Screening* **2015**, *18* (5), 476−485.
(29) Thai, K.-M.; Huynh, N.-T.; Ngo, T.-D.; Mai, T.-T.; Nguyen, T.-H.; Tran, T.-D. Three- and Four-Class Classification Models for P-Glycoprotein Inhibitors Using Counter-Propagation Neural Networks. *SAR QSAR Env. Res.* **2015**, *26* (2), 139−163.
(30) Deeken, J. F.; Löscher, W. The Blood-Brain Barrier and Cancer: Transporters, Treatment, and Trojan Horses. *Clin. Cancer Res.* **2007**, *13* (6), 1663−1674.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

(31) Lhommé, C.; Joly, F.; Walker, J.; Lissoni, A.; Nicoletto, M.; Manikhas, G.; Baekelandt, M.; Gordon, A.; Fracasso, P.; Mietlowski, W.; et al. Phase III Study of Valspodar (PSC 833) Combined with Paclitaxel and Carboplatin Compared with Paclitaxel and Carboplatin Alone in Patients with Stage IV or Suboptimally Debulked Stage III Epithelial Ovarian Cancer or Primary Peritoneal Cancer. *J. Clin. Oncol.* **2008**, *26* (16), 2674−2682.
(32) Teodori, E.; Dei, S.; Martelli, C.; Scapecchi, S.; Gualtieri, F. The Functions and Structure of ABC Transporters: Implications for the Design of New Inhibitors of Pgp and MRP1 to Control Multidrug Resistance (MDR). *Curr. Drug Targets* **2006**, *7* (7), 893−909.
(33) Hazai, E.; Hazai, I.; Ragueneau-Majlessi, I.; Chung, S. P.; Bikadi, Z.; Mao, Q. Predicting Substrates of the Human Breast Cancer Resistance Protein Using a Support Vector Machine Method. *BMC Bioinf.* **2013**, *14*, 130.
(34) Zhong, L.; Ma, C. Y.; Zhang, H.; Yang, L. J.; Wan, H. L.; Xie, Q. Q.; Li, L. L.; Yang, S. Y. A Prediction Model of Substrates and Non-Substrates of Breast Cancer Resistance Protein (BCRP) Developed by GA-CG-SVM Method. *Comput. Biol. Med.* **2011**, *41* (11), 1006−1013.
(35) Sedykh, A.; Fourches, D.; Duan, J.; Hucke, O.; Garneau, M.; Zhu, H.; Bonneau, P.; Tropsha, A. Human Intestinal Transporter Database: QSAR Modeling and Virtual Profiling of Drug Uptake, Efflux and Interactions. *Pharm. Res.* **2013**, *30* (4), 996−1007.
(36) Erić, S.; Kalinić, M.; Ilić, K.; Zloh, M. Computational Classification Models for Predicting the Interaction of Drugs with P-Glycoprotein and Breast Cancer Resistance Protein. *SAR QSAR Environ. Res.* **2014**, *25* (12), 939−966.
(37) Garg, P.; Dhakne, R.; Belekar, V. Role of Breast Cancer Resistance Protein (BCRP) as Active Efflux Transporter on Blood-Brain Barrier (BBB) Permeability. *Mol. Diversity* **2015**, *19* (1), 163−172.
(38) Lee, Y.; Jana, S.; Acharya, G.; Lee, C. H. Computational Analysis and Predictive Modeling of Polymorph Descriptors. *Chem. Cent. J.* **2013**, *7* (1), 23.
(39) Ose, A.; Toshimoto, K.; Ikeda, K.; Maeda, K.; Yoshida, S.; Yamashita, F.; Hashida, M.; Ishida, T.; Akiyama, Y.; Sugiyama, Y. Development of a Support Vector Machine-Based System to Predict Whether a Compound Is a Substrate of a Given Drug Transporter Using Its Chemical Structure. *J. Pharm. Sci.* **2016**, *105* (7), 2222−2230.
(40) Pinto, M.; Digles, D.; Ecker, G. F. Computational Models for Predicting the Interaction with ABC Transporters. *Drug Discovery Today: Technol.* **2014**, *12*, e69−e77.
(41) Chen, L.; Li, Y.; Yu, H.; Zhang, L.; Hou, T. Computational Models for Predicting Substrates or Inhibitors of P-Glycoprotein. *Drug Discovery Today* **2012**, *17* (7−8), 343−351.
(42) Gantner, M. E.; Di Ianni, M. E.; Ruiz, M. E.; Talevi, A.; Bruno-Blanch, L. E. Development of Conformation Independent Computational Models for the Early Recognition of Breast Cancer Resistance Protein Substrates. *BioMed Res. Int.* **2013**, *2013*, 863592.
(43) Gantner, M.; Alberca, L.; Mercader, A.; Bruno-Blanch, L.; Talevi, A. Integrated Application of Enhanced Replacement Method and Ensemble Learning for the Prediction of BCRP/ABCG2 Substrates. *Curr. Bioinf.* **2017**, *12* (3), 239−248.
(44) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45* (3), 542−548.
(45) Böcker, A. Toward an Improved Clustering of Large Data Sets Using Maximum Common Substructures and Topological Fingerprints. *J. Chem. Inf. Model.* **2008**, *48* (11), 2097−2107.
(46) Hartigan, J. *Clustering Algorithms*; John Wiley & Sons, Inc: New York, NY, USA, 1975.
(47) Hartigan, J. A.; Wong, M. A. A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Applied Stat.* **1979**, *28* (1), 100−108.
(48) Everitt, B. S.; Landau, S.; Leese, M.; Stahl, D. Optimization Clustering Techniques. In *Cluster Analysis*; Wiley: 2011; DOI: 10.1002/9780470977811.ch5.
(49) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29* (6−7), 476−488.
(50) Quinlan, J. R. *C4.5: Programs for Machine Learning*; 1993; Vol. *1*.
(51) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11* (1), 10−18.
(52) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. Virtual Screening Workflow Development Guided by The "receiver Operating Characteristic" curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, *48* (7), 2534−2547.
(53) DeLong, E. R.; DeLong, D. M.; Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44* (3), 837−845.
(54) Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Int. Med.* **2013**, *4* (2), 627−635.
(55) Kunal, R.; Kar, S. How to Judge Predictive Quality of Classification and Regression Based QSAR Models. In *Frontiers in Computational Chemistry*; Betham Science: 2014; pp 71−120.
(56) Truchon, J. F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for The "early Recognition" problem. *J. Chem. Inf. Model.* **2007**, *47* (2), 488−508.
(57) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177−182.
(58) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202−D1213.
(59) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789−6801.
(60) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582−6594.
(61) Giri, N.; Agarwal, S.; Shaik, N.; Pan, G.; Chen, Y.; Elmquist, W. F. Substrate-Dependent Breast Cancer Resistance Protein (Bcrp1/Abcg2)-Mediated Interactions: Consideration of Multiple Binding Sites in in Vitro Assay Design. *Drug Metab. Dispos.* **2009**, *37* (3), 560−570.
(62) Nakagawa, R.; Hara, Y.; Arakawa, H.; Nishimura, S.; Komatani, H. ABCG2 Confers Resistance to Indolocarbazole Compounds by ATP-Dependent Transport. *Biochem. Biophys. Res. Commun.* **2002**, *299* (4), 669−675.
(63) Takenaka, K.; Morgan, J. A.; Scheffer, G. L.; Adachi, M.; Stewart, C. F.; Sun, D.; Leggas, M.; Ejendal, K. F. K.; Hrycyna, C. A.; Schuetz, J. D. Substrate Overlap between Mrp4 and Abcg2/Bcrp Affects Purine Analogue Drug Cytotoxicity and Tissue Distribution. *Cancer Res.* **2007**, *67* (14), 6965−6972.
(64) Mo, W.; Qi, J.; Zhang, J. T. Different Roles of TM5, TM6, and ECL3 in the Oligomerization and Function of Human ABCG2. *Biochemistry* **2012**, *51* (17), 3634−3641.
(65) Hazai, E.; Bikádi, Z. Homology Modeling of Breast Cancer Resistance Protein (ABCG2). *J. Struct. Biol.* **2008**, *162* (1), 63−74.
(66) Clark, R.; Kerr, I. D.; Callaghan, R. Multiple Drug Binding Sites on the R482G Isoform of the ABCG2 Transporter. *Br. J. Pharmacol.* **2006**, *149* (5), 506−515.
(67) Li, L.; Hu, Q.; Wu, X.; Yu, D. Exploration of Classification Confidence in Ensemble Learning. *Pattern Recognit.* **2014**, *47* (9), 3120−3131.
(68) Zhou, Z. H.; Wu, J.; Tang, W. Ensembling Neural Networks: Many Could Be Better than All. *Artif. Intell.* **2002**, *137* (1−2), 239−263.
(69) Sastry, G. M.; Inakollu, V. S. S.; Sherman, W. Boosting Virtual Screening Enrichments with Data Fusion: Coalescing Hits from Two-Dimensional Fingerprints, Shape, and Docking. *J. Chem. Inf. Model.* **2013**, *53* (7), 1531−1542.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880

(70) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49* (5), 1536−1548.
(71) Di Ianni, M. E.; Enrique, A. V.; Palestro, P. H.; Gavernet, L.; Talevi, A.; Bruno-Blanch, L. E. Several New Diverse Anticonvulsant Agents Discovered in a Virtual Screening Campaign Aimed at Novel Antiepileptic Drugs to Treat Refractory Epilepsy. *J. Chem. Inf. Model.* **2012**, *52* (12), 3325−3330.
(72) Couyoupetrou, M.; Gantner, M.; Di Ianni, M.; Palestro, P.; Enrique, A.; Gavernet, L.; Ruiz, M.; Pesce, G.; Bruno-Blanch, L.; Talevi, A. Computer-Aided Recognition of ABC Transporters Substrates and Its Application to the Development of New Drugs for Refractory Epilepsy. *Mini-Reviews Med. Chem.* **2017**, *17* (3), 205−215.
(73) Tasso, S. M.; Moon, S. C.; Bruno-Blanch, L. E.; Estiú, G. L. Characterization of the Anticonvulsant Profile of Valpromide Derivatives. *Bioorg. Med. Chem.* **2004**, *12* (14), 3857−3869.
(74) Talevi, A.; Cravero, M. S.; Castro, E. a; Bruno-Blanch, L. E. Discovery of Anticonvulsant Activity of Abietic Acid through Application of Linear Discriminant Analysis. *Bioorg. Med. Chem. Lett.* **2007**, *17* (6), 1684−1690.
(75) Talevi, A.; Enrique, A. V.; Bruno-Blanch, L. E. Anticonvulsant Activity of Artificial Sweeteners: A Structural Link between Sweet-Taste Receptor T1R3 and Brain Glutamate Receptors. *Bioorg. Med. Chem. Lett.* **2012**, *22* (12), 4072−4074.
(76) Gavernet, L.; Elvira, J. E.; Samaja, G. A.; Pastore, V.; Cravero, M. S.; Enrique, A.; Estiu, G.; Bruno-Blanch, L. E. Synthesis and Anticonvulsant Activity of Amino Acid-Derived Sulfamides. *J. Med. Chem.* **2009**, *52* (6), 1592−1601.
(77) Villalba, M. L.; Enrique, A. V.; Higgs, J.; Castaño, R. A.; Goicoechea, S.; Taborda, F. D.; Gavernet, L.; Lick, I. D.; Marder, M.; Bruno Blanch, L. E. Novel Sulfamides and Sulfamates Derived from Amino Esters: Synthetic Studies and Anticonvulsant Activity. *Eur. J. Pharmacol.* **2016**, *774*, 55−63.
(78) Villalba, M. L.; Palestro, P.; Ceruso, M.; Gonzalez Funes, J. L.; Talevi, A.; Bruno Blanch, L.; Supuran, C. T.; Gavernet, L. Sulfamide Derivatives with Selective Carbonic Anhydrase VII Inhibitory Action. *Bioorg. Med. Chem.* **2016**, *24* (4), 894−901.
(79) Talevi, A.; Bellera, C. L.; Castro, E. A.; Bruno-Blanch, L. E. A Successful Virtual Screening Application: Prediction of Anticonvulsant Activity in MES Test of Widely Used Pharmaceutical and Food Preservatives Methylparaben and Propylparaben. *J. Comput.-Aided Mol. Des.* **2007**, *21* (9), 527−538.
(80) Gavernet, L.; Barrios, I. A.; Cravero, M. S.; Bruno-Blanch, L. E. Design, Synthesis, and Anticonvulsant Activity of Some Sulfamides. *Bioorg. Med. Chem.* **2007**, *15* (16), 5604−5614.
(81) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena Pyriformis. *J. Chem. Inf. Model.* **2008**, *48* (4), 766−784.
(82) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22* (1), 69−77.
(83) Rong, Z.; Xu, Y.; Zhang, C.; Xiang, D.; Li, X.; Liu, D. Evaluation of Intestinal Absorption of Amtolmetin Guacyl in Rats: Breast Cancer Resistant Protein as a Primary Barrier of Oral Bioavailability. *Life Sci.* **2013**, *92* (3), 245−251.
(84) Peroni, R. N.; Di Gennaro, S. S.; Hocht, C.; Chiappetta, D. A.; Rubio, M. C.; Sosnik, A.; Bramuglia, G. F. Efavirenz Is a Substrate and in Turn Modulates the Expression of the Efflux Transporter ABCG2/BCRP in the Gastrointestinal Tract of the Rat. *Biochem. Pharmacol.* **2011**, *82* (9), 1227−1233.
(85) Yumoto, R.; Hamada, S.; Okada, K.; Kato, Y.; Ikehata, M.; Nagai, J.; Takano, M. Effect of Ursodeoxycholic Acid Treatment on the Expression and Function of Multidrug Resistance-Associated Protein 2 in Rat Intestine. *J. Pharm. Sci.* **2009**, *98* (8), 2822−2831.
(86) Patel, J. P.; Korashy, H. M.; El-Kadi, A. O. S.; Brocks, D. R. Effect of Bile and Lipids on the Stereoselective Metabolism of Halofantrine by Rat Everted-Intestinal Sacs. *Chirality* **2010**, *22* (2), 275−283.
(87) Ni, L.; Yu, X.; Yu, Q.; Chen, X.; Jia, L. Effects of Cyclosporine A and Itraconazole on Permeability, Biliary Excretion and Pharmacokinetics of Amlodipine. *Drug Metab. Lett.* **2008**, *2* (3), 163−168.
(88) Machavaram, K. K.; Gundu, J.; Yamsani, M. R. Effect of Ketoconazole and Rifampicin on the Pharmacokinetics of Ranitidine in Healthy Human Volunteers: A Possible Role of P-Glycoprotein. *Drug Metab. Drug Interact.* **2006**, *22* (1), 47−65.
(89) Kashimura, J.; Nagai, Y. Inhibitory Effect of Palatinose on Glucose Absorption in Everted Rat Gut. *J. Nutr. Sci. Vitaminol.* **2007**, *53* (1), 87−89.
(90) MacLean, C.; Moenning, U.; Reichel, A.; Fricker, G. Closing the Gaps: A Full Scan of the Intestinal Expression of P-Glycoprotein, Breast Cancer Resistance Protein, and Multidrug Resistance-Associated Protein 2 in Male and Female Rats. *Drug Metab. Dispos.* **2008**, *36* (7), 1249−1254.
(91) Merino, G.; Jonker, J. W.; Wagenaar, E.; van Herwaarden, A. E.; Schinkel, A. H. The Breast Cancer Resistance Protein (BCRP/ABCG2) Affects Pharmacokinetics, Hepatobiliary Excretion, and Milk Secretion of the Antibiotic Nitrofurantoin. *Mol. Pharmacol.* **2005**, *67* (5), 1758−1764.
(92) Allen, J. D.; van Loevezijn, A.; Lakhai, J. M.; van der Valk, M.; van Tellingen, O.; Reid, G.; Schellens, J. H. M.; Koomen, G.-J.; Schinkel, A. H. Potent and Specific Inhibition of the Breast Cancer Resistance Protein Multidrug Transporter in Vitro and in Mouse Intestine by a Novel Analogue of Fumitremorgin C. *Mol. Cancer Ther.* **2002**, *1* (6), 417−425.
(93) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall: Belmont, CA, 1984; Vol. *19*.
(94) Yang, H.; Zhang, J.; Yu, B.; Zhao, W. *Statistical Methods for Immunogenicity Assessment*; Chapman & Hall/CRC, Ed.; CRC Press Taylor & Francis group: 2015.
(95) Luo, X.; Yu, J.; Li, Z. *Advanced Data Mining and Applications. 10th International Conference, ADMA 2014, Guilin, China, December 19−21, Proceedings*; Springer International Publishing: 2014.

DOI: 10.1021/acs.jcim.7b00016
J. Chem. Inf. Model. 2017, 57, 1868−1880