Integrated Application of Enhanced Replacement Method and Ensemble Learning for the Prediction of BCRP/ABCG2 Substrates

Short running title: ERM and Ensemble Learning for Prediction of BCRP/ABCG2 Substrates

KEYWORDS: Breast Cancer Resistance Protein; ABC Transporters; ABCG2; Enhanced Replacement Method; Ensemble Learning; Linear Classifiers

Melisa E. Gantner[a], Lucas N. Alberca[a], Andrew G. Mercader[b], Luis E. Bruno-Blanch[a], Alan Talevi[a] *

a.- Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP) - La Plata (B1900AJI), Buenos Aires, Argentina

b.- Institute of Theoretical and Applied Physicochemical Research (INIFTA), National University of La Plata (UNLP) - CONICET, La Plata, Buenos Aires, Argentina

* Corresponding author: atalevi@biol.unlp.edu.ar – Phone: 542214235333 ext 41

ABSTRACT: Breast Cancer Resistance Protein (BCRP or ABCG2) is a polyspecific efflux-transporter which belongs to the ATP-binding Cassette superfamily. Up-regulation of BCRP is associated to multi-drug resistance in a number of conditions, e.g. cancer and epilepsy. Recent proteomic studies show that high-expression levels of BCRP are found in healthy human intestine and at the blood-brain barrier, limiting the absorption and brain distribution of its substrates. Here, we have jointly applied the Enhanced Replacement Method and ensemble learning approaches to obtain combinations of 2D linear classifiers capable of discriminating among substrates and non-substrates of the wild type human BCRP. The best model ensemble obtained outperforms previously reported 2D linear classifiers, showing the ability of the Enhanced Replacement Method and ensemble learning schemes to optimize the performance of individual models. This is the first report of the Enhanced Replacement Method to solve classification problems.

1. Introduction

Breast Cancer Resistance Protein (BCRP, also known as ABCG2) is an ATP-dependent efflux pump characterized by broad substrate specificity. It limits the absorption and biodistribution and promotes the elimination of structurally and functionally heterogeneous substrates, including anticancer, antiepileptic, antiviral and antihypertensive drugs, among others. Up-regulation of BCRP expression has been linked to multi-drug resistance issues in many disorders, such as cancer [1-2] and epilepsy [3-5]. Though most of the early research on ATP-binding cassette (ABC) transporters has focused on P-glycoprotein (Pgp, ABCB1), attention has recently been drawn to other members of the superfamily: recent reports demonstrate that BCRP expression levels at both the small intestine and the blood-brain barrier of healthy tissues are comparable (and even higher) than those of Pgp [6-7], suggesting a prominent role of BCRP in drug pharmacokinetics. Therefore, regulation of BCRP activity and/or early recognition of BCRP substrates are critical in the early phase of drug discovery to enhance drug bioavailability and design novel therapeutics aimed at diseases linked to BCRP-mediated multi-drug resistance issues.

To the moment, limited studies have been reported on the development of high-throughput *in silico* models for the early recognition of BCRP substrates [8-9]. The main limitations of these models are the inclusion of conformation-dependent (3D) molecular descriptors and that they have usually been derived from unbalanced training sets in which substrates are overrepresented in comparison to non-substrates (with the exception of the recent report from Erić and colleagues [10], where a balanced dataset was used). The use of 3D descriptors (which demand geometry pre-optimization of the

screened compounds) limits the application of such models in virtual screening campaigns, in which chemical repositories containing thousands to millions of small molecules are typically screened through *in silico* filters. On the other hand, inferring models from unbalanced training sets tends to produce predictions that are biased towards the overrepresented category of training examples [11-12].

We have recently reported an ensemble of linear classifiers obtained through Stepwise Forward Linear Discriminant Analysis. These classifiers are entirely based on conformation-independent (0-2D) molecular descriptors and they have been derived from a balanced and representative training set obtained through a clustering procedure [13].

Here, we have resorted to the Enhanced Replacement Method (ERM) [14-15] and ensemble learning in an attempt to improve the performance of the previously reported model-ensemble.

2. Materials and Methods

2.1. Dataset compilation and partition into representative training and test sets

The dataset was compiled from literature. For this purpose, we have considered a compound as a substrate only if it is efficiently transported by BCRP and as a non-substrate otherwise. Correspondingly, compounds that bind to BCRP but are not transported are not considered substrates. A 305-compound diverse dataset containing BCRP substrates and non-substrates was initially compiled from 150 articles. From this initial dataset, additional inclusion and exclusion criteria were considered to define the final dataset. First, since it has been reported that single-nucleotide substitutions modify

the substrate specificity of BCRP [16-24], and taking into account that the clinical importance of such variants are not clear to the moment [25], we have only kept substrates from wild type BCRP in the final dataset. Second, substrates of BCRP homologs from other species with no evidence of human BCRP-mediated transport were not included to avoid noise due to inter-species variability in substrate specificity. As a result, from the initial 305-compound dataset only 262 compounds were kept: 156 substrates and 106 non-substrates of human wild-type BCRP. The dataset was split into a balanced and representative 164-compound training set (consisting in 85 substrates and 79 non-substrates) and a 98-compound independent test set (71 substrates and 27 non-substrates). In order to obtain representative partitions of the dataset, the LibraryMCS v0.7 (ChemAxon, 2011) hierarchical clustering procedure has been jointly applied with the k-means optimization clustering algorithm implemented in Statistica 10 Cluster Analysis Module (Statsoft Inc., 2011). In the recent years, many novel clustering algorithms have been developed for different applications [26-30]. The LibraryMCS is a hierarchical clustering procedure that uses the maximum common substructure (MCS, the largest subgraph found in two chemical graphs) in combination with molecular fingerprints to group a set of small molecules. It has specifically been developed for the clustering of small molecules and it has been widely applied for such purpose [31-34]. Following Everitt et al. advice [35], hierarchical clustering was applied to decide on an initial partition of n molecules into k groups; this preliminary clustering was then optimized through the k-means approach, minimizing the Euclidean distance to the group centers. A number of molecular descriptors calculated with Dragon 4.0 (Milano Chemometrics, 2003) reflecting diverse features of the molecular structure (logP, number of H bonds donors and acceptors, molecular weight, sum of atomic van der

Waals volumes, polar surface area, sum of atomic Sanderson electronegativities, 2D Petitjean shape index and total information index of atomic composition) were normalized and used to compute such distance. After the clusters were independently recognized in the substrates and non-substrates categories, around 50% of each cluster from the substrate category and 25% of each cluster in the non-substrate category were randomly assigned to a test set for validation purposes. The remaining elements were kept as calibration set. This scheme provided a balanced training set where neither the substrates nor the non-substrates were overrepresented. Details on the training and test set compounds are available as Supplementary information so that the reader can appreciate the structural diversity of the dataset.

2.2. Descriptor computation and modeling procedure

Dragon software v. 6.0 (Milano Chemometrics, 2012) was applied for the computation of 2029 conformation-independent descriptors, distributed along 19 descriptor-blocks, among them topological indices, ring descriptors, 2D atom pairs, information indices and others. A very important issue in multivariate analysis is to remove data having low variance using pre-filtering processes [36-37]. Here we have applied the initial filtering steps provided by Dragon software to exclude molecular descriptors with low information content, including constant or near-constant values (identical values for all the training examples but one) and descriptors with standard deviation below 0.001.

A binary, dummy variable linked to the category of each compound was used as dependent variable (class = 1 for substrates and class = −1 for non-substrates). The ERM [14-15] was used to select, from the descriptor pool, linear combinations of descriptors capable of predicting whether a chemical compound is or is not a BCRP substrate. The original Replacement Method (RM) was developed to explore the descriptor space in an efficient manner, in search of a subset of molecular descriptors from a large set of descriptors [38-40]. It is a rapidly convergent iterative algorithm which produces linear models that perform quite close to the full search solution with much less computational cost. Briefly, an initial subset of descriptors *d* from the pool of *D* descriptors is randomly selected. One of these selected descriptors, X_i, is then replaced one by one with all the remaining (*D*-*d*) descriptors in the pool. The subset of descriptors with the smallest standard deviation (*S*) is kept. From this resulting subset, the descriptor with the largest standard error in its regression coefficient is substituted one by one with all the remaining (*D*-*d*) descriptors in the pool. If the replacement of the descriptor with the largest error by those in the pool does not decrease the value of *S*, then the descriptor is not changed. The procedure is repeated until the selected subset of descriptors remains unmodified. An improvement on the RM, called the Modified Replacement Method (MRM) [15] uses the same strategy except that in each step the descriptor with the largest error is substituted even when that replacement is not associated to a smaller value of *S*. The MRM converges to different solutions, commonly bounces from one point to another and simulates 'a higher noise' than the RM, though keeping the overall decreasing trend of the *S* function. This apparent thermal agitation makes the MRM less likely to get trapped in local minima at the cost of larger computer time. The ERM arises from the combinations of the RM and MRM following the RM-MRM-RM sequence; it

combines the good features of both methods and is the only algorithm that goes through a complete simulated annealing cycle [15]. ERM is less dependent on the starting subset of descriptors and shows less propensity to fall in local minima. ERM also has proven to provide better models than the more complex Genetic Algorithms [14].

Here, we have applied the ERM to obtain models including from 4 to 12 molecular descriptors. Models with a larger number of predictors were not considered to maintain the observations to predictor ratio above 15, reducing the chance of overfitting [41-42].

Only those models containing descriptors with significant coefficients at an alpha level of 0.05 were retained. Leave-one-out cross-validation and external validation (discussed in the next section) were used to assess each model's robustness and predictive ability.

2.3. External validation and model comparison

The 98-compound independent test set was used for external validation. Later, we have applied Receiving Operating Characteristic (ROC) curves analysis to assess and compare model performance [43]. The ROC curves are graphical plots of Sensitivity (Se) versus 1 minus Specificity (Sp), which provide a rational frame to balance type I and type II errors; they are also useful to optimize a model cutoff score value and to compare models statistically. Se represents the true positives (TP) rate, while Sp refers to the true negatives (TN) rate:

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

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

where FP denotes False Positives and FN stands for False Negatives. In our particular application, we are searching for drugs which are not recognized by BCRP (BCRP non-substrates) which will be considered our hits or positives, whereas we want to discard BCRP substrates, which will be then regarded as negatives. A perfect classifier presents an Area Under the Receiving Operating Characteristic curve (AUROCc) of 1, while random classification is associated to an AUROCc of 0.5. MedCalc (MedCalc software, 2011) was used to obtain and statistically compare ROC curves. For statistical comparison of two AUROCcs, the nonparametric method of DeLong et al. [44] was used for the calculation of the standard error of each AUROCc and then the z-statistic was calculated in order to obtain the correspondent p-value [45].

It has been observed that standard errors of enrichment metrics such as the AUROCc used here are higher for small datasets than for large datasets [46]. Moreover, using a limited test set with a relatively high proportion of positives leads to a saturation effect: once the hit compounds saturate the early part of the ranking the enrichment metric cannot get any higher. If a model is conceived to analyze large chemical repositories where very few hits might be dispersed among a large number of non-hits, a more challenging and informative test is to conduct a pilot screening on a large database in which relatively few hits are dispersed among a high proportion of non-hits. Therefore, we have built two pilot chemical libraries in order to estimate in a more realistic way the utility of our model in a real virtual screening setting.

On the one hand, we built a 577-compound pilot chemical library (called simulated library) in which our 98-compound test set was dispersed among 479 putative BCRP substrates acting as decoys. Since the human BCRP substrates reported in literature are limited, our putative substrates are substrates of non-human BCRP homologs from other species or highly similar compounds to known human BCRP substrates retrieved from ZINC and PubChem databases through molecular similarity searches (similarity score > 0.75 compared to known substrates). The resulting pilot library contains 27 known non-substrates among 550 known or putative substrates, leading to a hit ratio smaller than 0.05. The same pilot library has previously been applied to assess the performance of our already reported classificatory models (which were derived from the same training set through Stepwise Forward Linear Discriminant Analysis) [13].

As a final challenge to our models, we have used the Enhanced Directory of Useful Decoys resource (DUD-E) [47-48] to build a second, larger and more diverse chemical library (which will be called DUD-E library) containing 1346 compounds (1248 decoys plus the original 98-compound test set) where each decoy is physicochemically similar but topologically dissimilar to its corresponding non-substrate. For this purpose we used the automated decoy generation tool available online. Succinctly, the decoys are property-matched to known hits (in our case, the known non-substrates) using molecular weight, an estimated LogP (miLogP), hydrogen bond donor and acceptors count, rotatable bonds and net molecular charge. About 50 decoys for each compound are chosen from ZINC [49] using a dynamic protocol that adapts to the local chemical space by narrowing or widening windows around the 6 properties. The goal is to obtain from ZINC 3000 to 9000 potential decoys matching the compounds. In a final step, ECFP4 fingerprints are generated for the compounds and its potential decoys; each compound's

decoys are ranked by their maximum Tanimoto coefficient and the most dissimilar 25% are kept.

2.4. Ensemble learning

Ensemble learning uses multiple learning algorithms to obtain better predictive performance than the one that could be obtained from any of the individual constituent learning algorithms [50]. Here we have combined a) the scores of the 10 individual models that displayed the best AUROCc for the test set and; b) the scores of all the individual models that showed a global accuracy above 70% for the test set (in total, 97 models). We have used five combination schemes to obtain a combined score: MAX Operator; MIN Operator; Average Score; Average Ranking and; Average Voting. Voting was computed according to the equation previously used by Zhang and Muegge [51]: the vote obtained by the j^{th} compound in the i^{th} model is equal to max(0, int(11-rankij/0.02NDB)), where rankij is the ranking of the j^{th} compound according to the i^{th} model, and NDB is the number of compounds in the entire screened library. This procedure gives 10 votes to the first 2% ranked compounds, 9 votes for the next 2%, and so on. Compounds in the bottom 80% of the ranking list receive no votes. The five combination schemes were analyzed and compared through ROC curves.

Figure **1** shows a flowchart summarizing the modeling methodology.

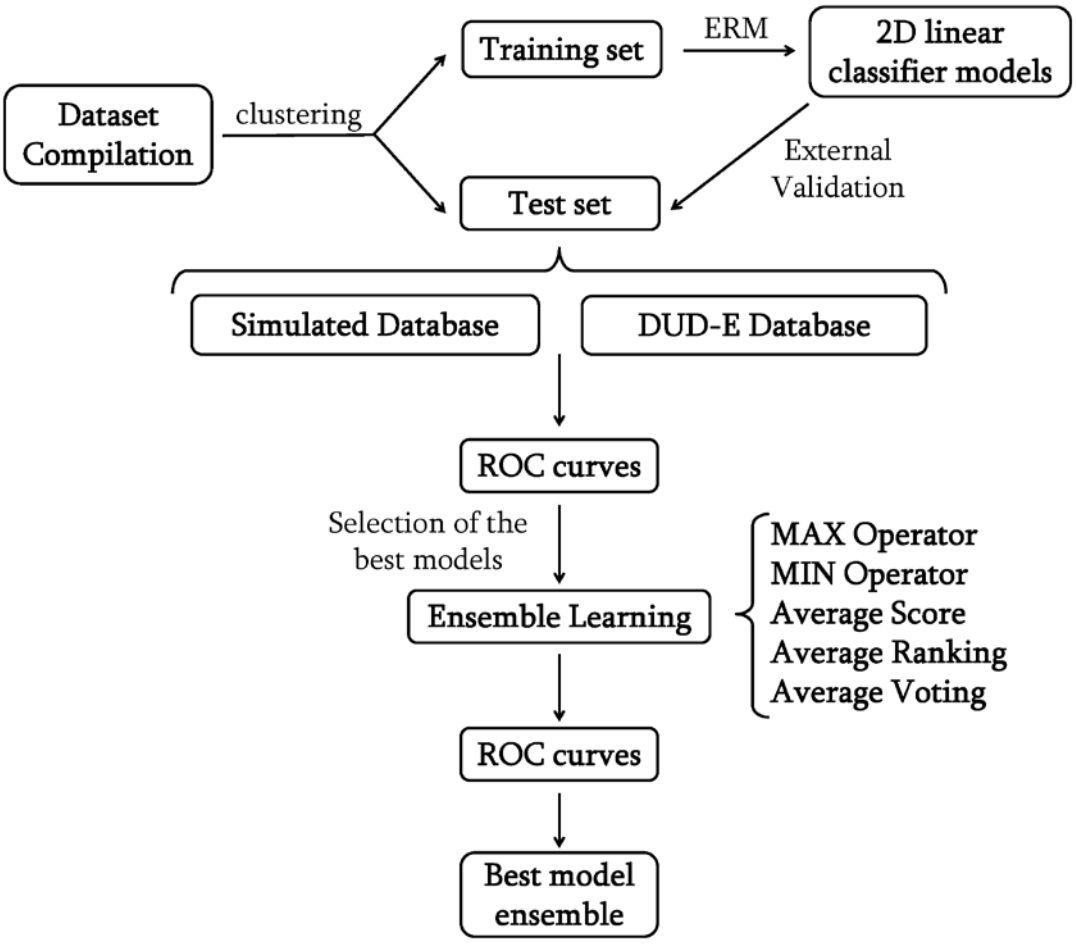

Figure **1**. Schematic representation of the modeling procedure.

3. Results and discussion

459 models incorporating between 4 and 12 molecular descriptors were obtained through the ERM. All of them displayed good results in the LOO cross-validation. 97 of them showed an overall percentage of good classification above 70% for the test set. The features of the 10 models showing the highest AUROCc for the test set are provided in Table **1**. Dragon's nomenclature for the descriptors has been kept. A more detailed

insight into the performance of each of these models on the training and test sets, the simulated library and the DUD-E library is presented in Table **2**. We have also included the results for the best individual model previously reported [13] (inferred through Stepwise Forward Linear Discriminant Analysis) and the two best 2-model ensembles obtained in that same study. All the individual models showed an AUROCc statistically different from random classification ($p < 0.0001$) for the training set, the test set and the 577-compound simulated library (Table **2**). All the 10 best models obtained through the ERM outperformed the best model obtained through Stepwise Forward Linear Discriminant Analysis (SF) for both the training and test sets, while all models display a similar behavior on the simulated library. Considering the AUROCc for the test set, the best ERM individual model was M438, which slightly outperforms the two best 2-model ensembles obtained through the Stepwise Forward approach (SF-E1 and SF-E2) in relation to the classification of the training and test set compounds; it also slightly outperforms SF-E1 regarding the simulated library classification while it shows similar performance to SF-E2 ($p = 0.6651$). The results suggest that the ERM provides better results than the Stepwise Forward approach. However, all the individual models experimented a significant drop on their performance when they were applied on the DUD-E library, evidently our most challenging test.

Table **1**. Summary of the features of the 10-ERM individual models which showed the best performance on the test set.

Model	Descriptors included	F	Wilk's λ	Overall accuracy Training Set	Overall accuracy Test Set
M36	**SM5_B(e):** spectral moment of order 5 from Burden matrix weighted by Sanderson electronegativity **P_VSA_i_2:** P_VSA-like on ionization potential, bin 2 **SpDiam_EA:** spectral diameter from edge adjacency mat. **nRCONHR:** number of secondary amides (aliphatic)	19.970	0.666	83.5	75.5
M61	**SpMax2_Bh(s):** largest eigenvalue n. 2 of Burden matrix weighted by I-state **SpMaxA_EA(ed):** normalized leading eigenvalue from edge adjacency mat. weighted by edge degree **nRCONHR** **CATS2D_07_NL:** CATS2D Negative-Lipophilic at lag 07 **F01[C-C]:** Frequency of C - C at topological distance 1	19.211	0.622	79	71
M107	**EE_B(e):** Estrada-like index (log function) from Burden matrix weighted by Sanderson electronegativity **P_VSA_i_2** **SpDiam_EA** **nRCONHR** **B10[C-C]:** Presence/absence of C - C at topological distance 10 **DLS_01:** modified drug-like score from Lipinski (4 rules)	18.373	0.587	83.5	71
M203	**IDE:** mean information content on the distance equality	18.426	0.512	83	70

	SpMaxA_Dz(e): normalized leading eigenvalue from Barysz matrix weighted by Sanderson electronegativity **SpPosA_B(v):** normalized spectral positive sum from Burden matrix weighted by van der Waals volume **nArOR:** number of ethers (aromatic) **S-108:** R=S **NaaaC:** Number of atoms of type aaaC **CATS2D_07_NL** **B10[C-C]**				
M289	**N%:** percentage of N atoms **SpMax2_Bh(s)** **SM11_EA:** spectral moment of order 11 from edge adjacency mat. **Eig05_EA**: eigenvalue n. 5 from edge adjacency mat **nCs:** number of total secondary C(sp3) **nRCONHR** **NaaaC** **CATS2D_07_NL** **Ui:** unsaturation index	17.838	0.490	83.5	74.5
M356	**N%** **nR04:** number of 4-membered rings **SpMax2_Bh(s)** **SM11_EA** **SM13_AEA(bo):** spectral moment of order 13 from augmented edge adjacency mat. weighted by bond order **nCs**	17.841	0.462	85	73.5

	nRCONHR **NaaaC** **CATS2D_07_NL** **Ui**				
M380	**IDE** **WiA_Dz(m):** average Wiener-like index from Barysz matrix weighted by mass **SpPosA_B(v)** **SpMax2_Bh(s)** **SaaaC:** Sum of aaaC E-states **SdS:** Sum of dS E-states **CATS2D_05_DA:** CATS2D Donor-Acceptor at lag 05 **CATS2D_01_NL:** CATS2D Negative-Lipophilic at lag 01 **B10[C-C]** **DLS_01**	17.415	0.468	85	73.5
M438	**nR07:** number of 7-membered rings **D/Dtr04:** distance/detour ring index of order 4 **piPC09:** molecular multiple path count of order 9 **ATSC2e:** Centred Broto-Moreau autocorrelation of lag 2 weighted by Sanderson electronegativity **GATS1e:** Geary autocorrelation of lag 1 weighted by Sanderson electronegativity **SpMax2_Bh(i):** largest eigenvalue n. 2 of Burden matrix weighted by ionization potential **SpMax2_Bh(s)**	17.588	0.440	86	75.5

	SM04_EA(ed): spectral moment of order 4 from edge adjacency mat. weighted by edge degree **Eig10_EA(dm):** eigenvalue n. 10 from edge adjacency mat. weighted by dipole moment **Uc**: unsaturation count **DLS_01**				
M446	**nR04** **SpMax_A:** leading eigenvalue from adjacency matrix (Lovasz-Pelikan index) **ATSC2e** **SpMax2_Bh(s)** **P_VSA_i_2** **Eig05_EA** **nRCONHR** **NaaaC** **B03[O-O]:** Presence/absence of O - O at topological distance 3 **F09[C-S]:** Frequency of C - S at topological distance 9 **cRo5:** Complementary Lipinski Alert index	17.419	0.442	85	71
M471	**SRW09:** self-returning walk count of order 9 **IDE** **WiA_Dz(e):** average Wiener-like index from Barysz matrix weighted by Sanderson electronegativity **SpMax2_Bh(s)** **Eta_betaS_A:** eta sigma average VEM coun	18.666	0.403	88	74.5

	C-034: R--CR..X **CATS2D_08_DP:** CATS2D Donor-Positive at lag 08 **CATS2D_07_NL** **B02[C-O]:** Presence/absence of C - O at topological distance 2 **B08[N-O]:** Presence/absence of N - O at topological distance 8 **B10[C-C]** **Ui**				

Table **2**. Se, Sp, Global accuracy and AUROCc values of the 10 best individual models obtained with the ERM for the training and test sets, the simulated library and the DUD-E library. Such values for the best previously reported individual model (SF) and the two best 2-model ensembles (SF-E1 and SF-E2) obtained through the Stepwise Forward technique have also been included for comparison purposes. A score of zero has been taken as the cutoff value to assess Sp and Se. The best AUROCc for each set of compounds has been indicated in bold.

	Training Set				**Test Set**				**Simulated Library**				**DUD-E Library**			
Model	***Sp***	***Se***	**Global**	**AUROCc**	***Sp***	***Se***	**Global**	**AUROCc**	***Sp***	***Se***	**Global**	**AUROCc**	***Sp***	***Se***	**Global**	**AUROCc**
M438	91	81	86	0.945**	85	52	75.5	**0.823****	76	52	74.5	0.755**	51.8	61.6	61.4	0.503
M380	86	85	85	0.939**	72	78	73.5	0.789**	68	78	67	0.746**	77.8	57.5	57.9	0.576
M107	82	85	83.5	0.894**	73	67	71	0.788**	61	67	61	0.719**	66.7	48.6	49	0.636*
M471	86	90	88	**0.957****	76	70	74.5	0.788**	75	70	74.5	0.765**	70.4	45.4	45.9	0.611*
M356	87	84	85	0.946**	79	59	73.5	0.783**	70	59	69.5	0.747**	59.3	51	51.2	0.591
M61	86	71	79	0.885**	75	63	71	0.774**	70	63	67	0.756**	63	46.7	47	**0.666***

M203	82	84	83	0.924**	70.4	70.4	70	0.774**	67	70.4	71	0.736**	70.4	57.4	57.6	0.590
M289	85	82	83.5	0.936**	59	80	74.5	0.774**	71.5	80	69.5	0.751**	59.3	50.1	50.3	0.580
M36	88	78	83.5	0.863**	74	76	75.5	0.773**	65	76	66	0.725**	74.1	47.4	47.9	0.634*
M446	86	85	85	0.948**	77.5	55.5	71	0.772**	66	55.5	65	0.732**	55.5	51.1	51.2	0.610†
SF	79	68	74	0.796**	63	74	66	0.748**	66	74	66	0.732**	46	74.1	46.6	0.622†
SF-E1	84	75	79	0.850**	70	74	71	0.785**	64	74	64	0.736**	53.6	74.1	54.1	0.660*
SF-E2	85	80	82	0.902**	76	70	74.5	0.804**	68	70	68	**0.771****	48	70.4	48.5	0.637†

AUROCc statistically different from a random classification (AUROCc = 0.5) ** $p < 0.0001$; * $p < 0.01$, † $p < 0.05$

The sharp drop in all the individual models performance when applied to the DUD-E library demonstrates the difficulty to find a single linear relationship capable of accurately classifying substrates and non-substrates when facing a real *in silico* screening application. In this sense, one should keep in mind that the broad substrate specificity of BCRP makes our modeling problem particularly challenging. It has been pointed out that the polyspecificity of ABC transporters due to multiple binding sites and high protein flexibility determine a complex phenomenon which can only be partially addressed by current methods in the computational drug design field [52-53]. This explains why many modeling efforts to identify ABC transporters substrates have resorted to ensemble learning or locally weighted methods [54-57]. Thus, we have applied ensemble learning methods combining the best ERM individual models through five different combination schemes, leading to remarkably improved results (Table **3**; the best ERM individual model –M438– has been included in the table to facilitate the comparison).

Table **3**. The table shows the AUROCcs values for the 10- and 97-ERM model ensembles obtained through the five combination schemes for the training and test sets, the simulated library and the DUD-E library; the AUROCcs values for the best individual model M438 have also been included. The best AUROCc for the 10-model ensemble has been highlighted in bold for each set of compounds

	AUROCc			
10-model ensemble	**Training Set**	**Test Set**	**Simulated Library**	**DUD-E Library**
MAX Operator	0.954**	**0.850****	**0.824****	**0.743****
MIN Operator	0.939**	0.776**	0.724**	0.526
Average Score	**0.963****	0.833**	0.791**	0.628*
Average Ranking	**0.963****	0.830**	0.789**	0.629†
Average Voting	0.871**	0.724**	0.711**	0.558
97-model ensemble	**Training Set**	**Test Set**	**Simulated Library**	**DUD-E Library**
MAX Operator	0.829**	0.647†	0.530	0.589
MIN Operator	0.934**	0.740**	0.573	0.512
Average Score	0.964**	0.778**	0.745**	0.598
Average Ranking	0.963**	0.781**	0.745**	0.591
Average Voting	0.906**	0.712**	0.651*	0.622†
M438	0.945**	0.823**	0.755**	0.503

AUROCc statistically different from a random classification (AUROCc = 0.5) ** $p < 0.0001$; * $p < 0.01$, † $p < 0.05$

Compared to the best ERM individual model, the combination of the 10 best individual ERM models using MAX Operator showed very similar performance on the training and the test sets ($p = 0.5626$ and $p = 0.4522$, respectively), but it markedly improved the performance on the simulated library ($p = 0.0353$) and the DUD-E library ($p < 0.0087$). No improvement compared to the best individual model (and, in fact, reduced accuracy) was observed when combining the 97 models that showed best performance on the test set. This suggests that, when resorting to ensemble learning,

selective combination of relatively few good models might be better than combining several weak models. This is in agreement with previous studies applying ensemble learning that suggest that the combination of a few and selected classifiers (selective ensemble) may provide better accuracy and generalization than combining all available learners [58-60]. Based on the previous results, we have also tested combinations of the 2, 3, 5, and 15 models (data not shown) with the best AUROCc on the test sets. None of these ensembles showed statistically significant differences on the training set, the test set and the simulated library AUROCcs. The 10-model ensemble displayed the best behavior on the DUD-E library, outperforming the 2-, 3- and 5-model ensembles ($p < 0.0001$) and showing very similar performance to the 15-model ensemble ($p = 0.6679$), constituting the best ensemble obtained.

It is also interesting that, for both the individual models and the model ensembles (no matter which combination scheme was used), the performance on the 98-compound test set is generally better than the performance on the 577-compound simulated library, falling sharply on the 1346-compound DUD-E library. In the one hand, this seems to confirm that assessing the model's performance on large chemical libraries is a more stringent test than assessing performance on small compound sets, and the ability of the Enhanced Directory of Useful Decoys to generate suitable decoys for validation purposes, on the other. The MAX Operator displayed the best results to combine individual learners, with an AUROCcs of 0.824 on the simulated library and of 0.743 on the DUD-E library, being the only ensemble scheme that showed a good performance on the larger and more diverse library tested, while all other individual models and model ensembles showed either poor or no classificatory power at all on this library. Figure **2**

shows the ROC curves of the 10-ERM model ensembles and the best ERM individual model for each set of compounds evaluated.

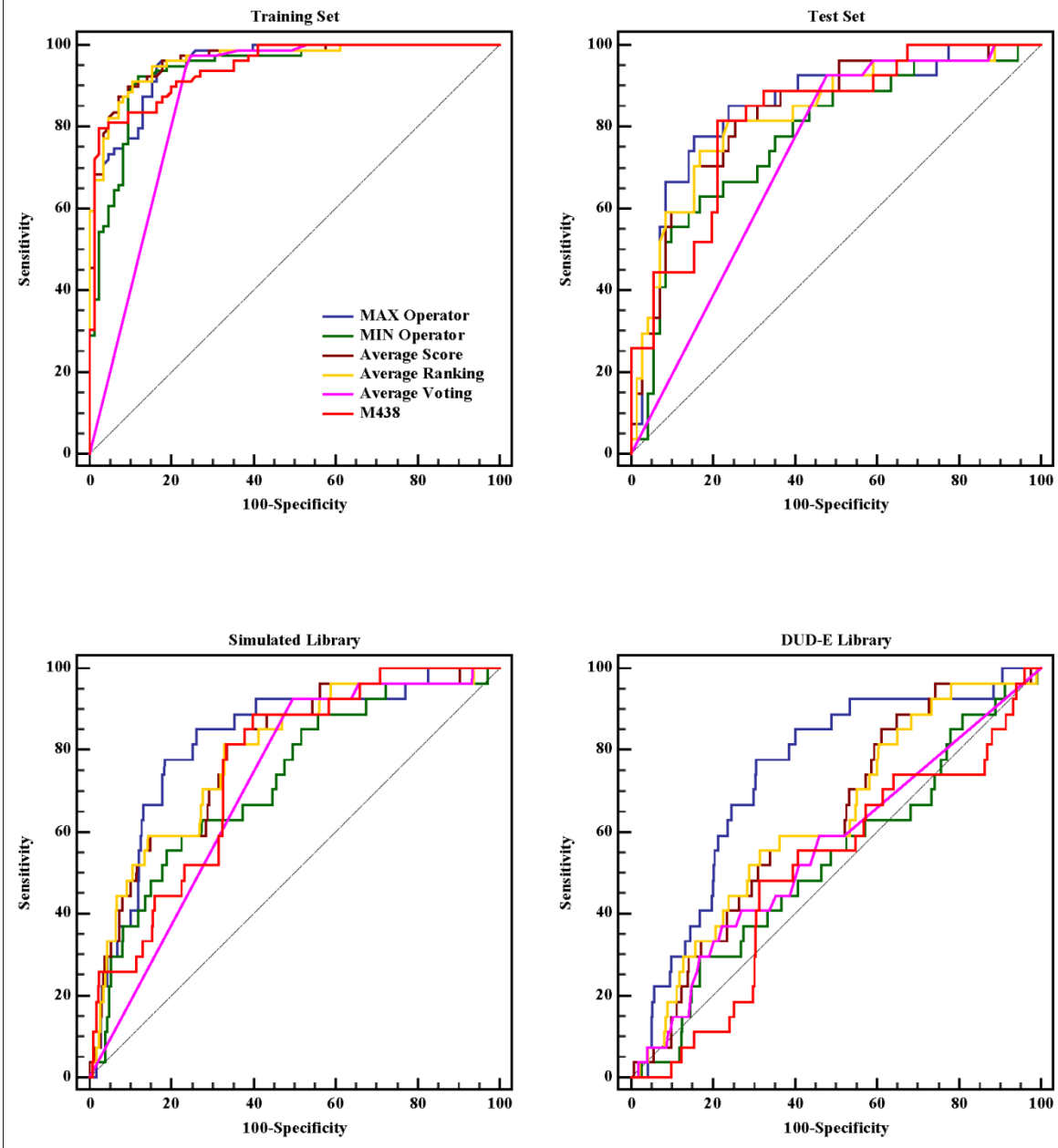

Figure **2**. Graphical comparison of the ROC curves from the best performing ensembles and the best ERM individual model (M438) for each set of compounds assessed.

Selective ensemble of 10 models through the MAX Operator is the only approach that consistently shows good performance on all the validations sets/libraries used.

4. Conclusions

Using the ERM to deal with a classification problem for the first time, we have developed linear model ensembles capable of identifying BCRP substrates and non-substrates; these models are entirely based on conformation-independent descriptors, thus being capable of screening large chemical repositories with high throughput. It should be highlighted that systematic comparisons of different virtual screening approaches with different levels of sophistication have surprisingly shown that the less sophisticated approaches might sometimes outperform more complex ones in terms of enrichment metrics: more complex methods are not unequivocally better; which method is more suitable seems to be highly target-dependent [51, 61]. Simpler approaches are, however, always more efficient.

The models reported here might be efficiently apply at an early stage in drug discovery projects to discard drug candidates predicted as BCRP substrates, which might present BCRP-associated bioavailability issues, drug-drug interactions and multi-drug resistance issues linked to BCRP. The models reported here clearly outperform previously reported linear classifiers by our group, which were obtained through the simpler Stepwise Forward Linear Discriminant Analysis, and confirm the utility of the ERM to establish useful QSAR relationships. The ensemble learning approach combining the 10-ERM best individual models obtained through MAX Operator

displayed the best capacity to discriminate between BCRP substrates and non-substrates across all the validation sets/libraries used here, proving to be effective to improve the predictive ability of the individual models. These results seem to confirm the strength of the selective ensemble approximation. The sharp drop in the classificatory power of the individual models and model ensembles when studying the DUD-E library underlines the ability of the Enhanced Directory of Useful Decoys to provide challenging collections of small molecules in order to assess computational models' predictive ability previous to proceed to real *in silico* screening applications. Furthermore, it was confirmed that assessing model performance on a large and diverse database is a more stringent test than assessing performance on smaller libraries, allowing estimating in a more realistic way the utility of a model in a real virtual screening setting.

Conflict of interest

The authors declare no conflict of interest related to the present article.

Acnowledgements

M. E. Gantner and L. N. Alberca are CONICET fellowship holders. A. Talevi and A. G. Mercader are members of the Scientific Research Career at CONICET. L. E. Bruno-Blanch is a researcher of the Faculty of Exact Sciences, National University of La Plata (UNLP). 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.

References

[1] Agarwal S, Hartz AM, Elmquist WF, Bauer B. Breast cancer resistance protein and P-glycoprotein in brain cancer: two gatekeepers team up. Curr Pharm Des. 2011;17(26):2793-802.
[2] Natarajan K, Xie Y, Baer MR, Ross DD. Role of breast cancer resistance protein (BCRP/ABCG2) in cancer drug resistance. Biochem Pharmacol. 2012;83(8):1084-103.
[3] Aronica E, Gorter JA, Redeker S, van Vliet EA, Ramkema M, Scheffer GL, et al. Localization of breast cancer resistance protein (BCRP) in microvessel endothelium of human control and epileptic brain. Epilepsia. 2005;46(6):849-57.
[4] 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):20-8.
[5] Sisodiya SM, Martinian L, Scheffer GL, van der Valk P, Scheper RJ, Harding BN, et al. Vascular colocalization of P-glycoprotein, multidrug-resistance associated protein 1, breast cancer resistance protein and major vault protein in human epileptogenic pathologies. Neuropathol Appl Neurobiol. 2006;32(1):51-63.
[6] Tucker TG, Milne AM, Fournel-Gigleux S, Fenner KS, Coughtrie MW. 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):279-85.
[7] Uchida Y, Ohtsuki S, Katsukura Y, Ikeda C, Suzuki T, Kamiie J, et al. Quantitative targeted absolute proteomics of human blood-brain barrier transporters and receptors. J Neurochem. 2011;117(2):333-45.
[8] Hazai E, Hazai I, Ragueneau-Majlessi I, Chung SP, Bikadi Z, Mao Q. Predicting substrates of the human breast cancer resistance protein using a support vector machine method. BMC Bioinformatics. 2013;14:130.
[9] Zhong L, Ma CY, Zhang H, Yang LJ, Wan HL, Xie QQ, et al. 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):1006-13.
[10] Eric S, Kalinic M, Ilic 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):939-66.
[11] Ghafourian T, Freitas AA, Newby D. The impact of training set data distributions for modelling of passive intestinal absorption. Int J Pharm. 2012;436(1-2):711-20.
[12] Van Hulse J, Khoshgoftaar T. Knowledge discovery from imbalanced and noisy data. Data & Knowledge Engineering. 2009;68:1513-42.

[13] Gantner ME, Di Ianni ME, Ruiz ME, Talevi A, Bruno-Blanch LE. Development of conformation independent computational models for the early recognition of breast cancer resistance protein substrates. Biomed Res Int. 2013;2013:863592.
[14] Mercader AG, Duchowicz PR, Fernandez FM, Castro EA. Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories. J Chem Inf Model. 2010;50(9):1542-8.
[15] Mercader AG, Duchowicz PR, Fernández FM, Castro EA. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. Chemometrics and Intelligent Laboratory Systems. 2008;92:138-44.
[16] Allen JD, Jackson SC, Schinkel AH. A mutation hot spot in the Bcrp1 (Abcg2) multidrug transporter in mouse cell lines selected for Doxorubicin resistance. Cancer Res. 2002;62(8):2294-9.
[17] Eddabra L, Wenner T, El Btaouri H, Baranek T, Madoulet C, Cornillet-Lefebvre P, et al. Arginine 482 to glycine mutation in ABCG2/BCRP increases etoposide transport and resistance to the drug in HEK-293 cells. Oncol Rep. 2012;27(1):232-7.
[18] Ejendal KF, Diop NK, Schweiger LC, Hrycyna CA. The nature of amino acid 482 of human ABCG2 affects substrate transport and ATP hydrolysis but not substrate binding. Protein Sci. 2006;15(7):1597-607.
[19] Janvilisri T, Shahi S, Venter H, Balakrishnan L, van Veen HW. Arginine-482 is not essential for transport of antibiotics, primary bile acids and unconjugated sterols by the human breast cancer resistance protein (ABCG2). Biochem J. 2005;385(Pt 2):419-26.
[20] Kathawala RJ, Gupta P, Ashby CR, Jr., Chen ZS. The modulation of ABC transporter-mediated multidrug resistance in cancer: a review of the past decade. Drug Resist Updat. 2015;18:1-17.
[21] Ozvegy-Laczka C, Koblos G, Sarkadi B, Varadi A. Single amino acid (482) variants of the ABCG2 multidrug transporter: major differences in transport capacity and substrate recognition. Biochim Biophys Acta. 2005;1668(1):53-63.
[22] Polgar O, Robey RW, Bates SE. ABCG2: structure, function and role in drug response. Expert Opin Drug Metab Toxicol. 2008;4(1):1-15.
[23] Pozza A, Perez-Victoria JM, Sardo A, Ahmed-Belkacem A, Di Pietro A. Purification of breast cancer resistance protein ABCG2 and role of arginine-482. Cell Mol Life Sci. 2006;63(16):1912-22.
[24] Robey RW, Honjo Y, Morisaki K, Nadjem TA, Runge S, Risbood M, et al. Mutations at amino-acid 482 in the ABCG2 gene affect substrate and antagonist specificity. Br J Cancer. 2003;89(10):1971-8.
[25] Cervenak J, Andrikovics H, Ozvegy-Laczka C, Tordai A, Nemet K, Varadi A, et al. The role of the human ABCG2 multidrug transporter and its variants in cancer therapy and toxicology. Cancer Lett. 2006;234(1):62-72.
[26] Saha S, Spandana R, Ekba A, Bandyopadhyay S. Simultaneous feature selection and symmetry based clustering using multiobjective framework. Applied Soft Computing. 2015;29:479-86.
[27] Saha S, Bandyopadhyay S. A symmetry based multiobjective clustering technique for automatic evolution of clusters. Pattern Recognition. 2010;43(3):738-51.
[28] Vijendra S, Laxman S. Symmetry Based Automatic Evolution of Clusters: A New Approach to Data Clustering. Comput Intell Neurosci. 2015;2015:796276.
[29] Abubaker A, Baharum A, Alrefaei M. Automatic Clustering Using Multi-objective Particle Swarm and Simulated Annealing. PLoS One. 2015;10(7):e0130995.

[30] Karaboga D, Ozturk C. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. Applied Soft Computing. 2011;11(1):652-7.
[31] Stahl M, Mauser H. Database clustering with a combination of fingerprint and maximum common substructure methods. J Chem Inf Model. 2005;45(3):542-8.
[32] Bocker A. Toward an improved clustering of large data sets using maximum common substructures and topological fingerprints. J Chem Inf Model. 2008;48(11):2097-107.
[33] Herhaus C. Introducing fuzziness into maximum common substructures for meaningful cluster characterisation. J Cheminform. 2014;6(Suppl 1):p17.
[34] Hariharan R, Janakiraman A, Nilakantan R, Singh B, Varghese S, Landrum G, et al. MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. J Chem Inf Model. 2011;51(4):788-806.
[35] Everitt B. Cluster analysis. 5th ed. Chichester, West Sussex, U.K.: Wiley; 2011.
[36] Bandyopadhyay S, Mallik S, Mukhopadhyay A. A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. IEEE/ACM Trans Comput Biol Bioinform. 2013.
[37] Mallik S, Mukhopadhyay A, Maulik U. RANWAR: rank-based weighted association rule mining from gene expression and methylation data. IEEE Trans Nanobioscience. 2015;14(1):59-66.
[38] Duchowicz PR, Castro EA, Fernández FM. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem. 2006;55(1):179-92.
[39] Mercader AG, Duchowicz PR, Fernandez FM, Castro EA. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. J Chem Inf Model. 2011;51(7):1575-81.
[40] Talevi A, Bellera CL, Di Ianni M, Duchowicz PR, Bruno-Blanch LE, Castro EA. An integrated drug development approach applying topological descriptors. Curr Comput Aided Drug Des. 2012;8(3):172-81.
[41] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503-10.
[42] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
[43] Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO. 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):2534-47.
[44] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.
[45] Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med. 2013;4(2):627-35.
[46] Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007;47(2):488-508.
[47] Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem. 2006;49(23):6789-801.

[48] Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem. 2012;55(14):6582-94.
[49] Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model. 2005;45(1):177-82.
[50] Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33(1-2):1-39.
[51] 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):1536-48.
[52] Demel MA, Kramer O, Ettmayer P, Haaksma EE, Ecker GF. Predicting ligand interactions with ABC transporters in ADME. Chem Biodivers. 2009;6(11):1960-9.
[53] Ecker GF. QSAR studies on ABC transporter—How to deal with polyspecificity. In: Ecker G, Chiba P, editors. Transporters as Drug Carriers: Structure, Function, Substrates. Weinheim, Germany: Wiley-VCH; 2009.
[54] Cao D-S, Huang J-H, Yan J, Zhang L-X, Hu Q-N, Xu Q-S. Kernel k-nearest neighbor algorithm as a flexible SAR modeling tool. Chemometrics and Intelligent Laboratory Systems. 2012;114:19-23.
[55] Li WX, Li L, Eksterowicz J, Ling XB, Cardozo M. Significance analysis and multiple pharmacophore models for differentiating P-glycoprotein substrates. J Chem Inf Model. 2007;47(6):2429-38.
[56] Penzotti JE, Lamb ML, Evensen E, Grootenhuis PD. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. J Med Chem. 2002;45(9):1737-40.
[57] Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q. Boosting: an ensemble learning tool for compound classification and QSAR modeling. J Chem Inf Model. 2005;45(3):786-99.
[58] Li l, Hu Q, Wu X, Yu D. Exploration of classification confidence in ensemble learning. Pattern Recogn. 2014;47:3120-31.
[59] Madhavi Sastry G, Andeep Inakollu VSS, Sherman W. Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimendional fingerprints, shape, and doking. J Chem Inf Model. 2013;53:1531-42.
[60] Zhou Z-H, Wu J, Tang W. Ensembling neural networks: Many could be better than all. Artif Intell Rev. 2002;137(1-2):239-63.
[61] Bender A, Glen RC. A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. J Chem Inf Model. 2005;45(5):1369-75.