JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Classifier Ensemble Based on Feature Selection and Diversity Measures for Predicting the Affinity of A_{2B} Adenosine Receptor Antagonists

Isis Bonet,[†,‡] Pedro Franco-Montero,[‡] Virginia Rivero,[§] Marta Teijeira,[§] Fernanda Borges,[∥] Eugenio Uriarte,[⊥] and Aliuska Morales Helguera*[,∥,#,▲]

[†]Escuela de Ingeniería de Antioquia, Envigado, 055428 Antioquia, Colombia

[‡]Centro de Estudios de Informática, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba

[§]Departamento de Química Orgánica, Universidade de Vigo, Vigo, 36320 Pontevedra, España

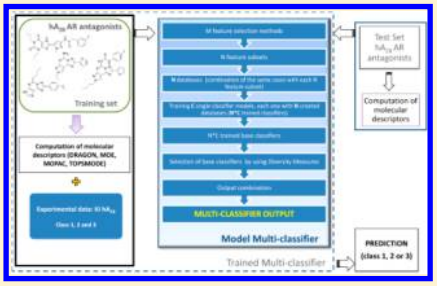
[∥]CIQ, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto, 4169-007 Porto, Portugal

[⊥]Departamento de Química Orgánica. Facultad de Farmacia, Universidad de Santiago de Compostela, Santiago de Compostela, 15782 La Coruña, España

[#]Centro de Bioactivos Químicos/[▲]Departamento de Química, Universidad Central "Marta Abreu" de Las Villas, Santa Clara, 54830 Villa Clara, Cuba

S *Supporting Information*

ABSTRACT: A_{2B} adenosine receptor antagonists may be beneficial in treating diseases like asthma, diabetes, diabetic retinopathy, and certain cancers. This has stimulated research for the development of potent ligands for this subtype, based on quantitative structure-affinity relationships. In this work, a new ensemble machine learning algorithm is proposed for classification and prediction of the ligand-binding affinity of A_{2B} adenosine receptor antagonists. This algorithm is based on the training of different classifier models with multiple training sets (composed of the same compounds but represented by diverse features). The *k*-nearest neighbor, decision trees, neural networks, and support vector machines were used as single classifiers. To select the base classifiers for combining into the ensemble, several diversity measures were employed. The final multiclassifier prediction results were computed from the output obtained by using a combination of selected base classifiers output, by utilizing different mathematical functions including the following: majority vote, maximum and average probability. In this work, 10-fold cross- and external validation were used. The strategy led to the following results: i) the single classifiers, together with previous features selections, resulted in good overall accuracy, ii) a comparison between single classifiers, and their combinations in the multiclassifier model, showed that using our ensemble gave a better performance than the single classifier model, and iii) our multiclassifier model performed better than the most widely used multiclassifier models in the literature. The results and statistical analysis demonstrated the supremacy of our multiclassifier approach for predicting the affinity of A_{2B} adenosine receptor antagonists, and it can be used to develop other QSAR models.



INTRODUCTION

Adenosine receptors (ARs) are a family of G-protein coupled receptors (GPCRs) of great interest as targets for therapeutic intervention due to their ubiquitous distribution throughout the body and their important modulatory effects on cell function.[1] Four distinct receptor subtypes of ARs have been characterized and cloned: A_1, A_{2A}, A_{2B}, and A_3.[2] From all subtypes, the A_{2B} adenosine receptor (A_{2B}AR) requires the highest concentrations of adenosine for activation.[3] Thus, unlike other AR subtypes, the A_{2B}AR is not stimulated by physiological levels of adenosine but may therefore play an important role in pathophysiological conditions associated with massive adenosine release. Such conditions occur in ischemia or in tumors where hypoxia is commonly observed.[4] Although potent and selective tools are scarce for the A_{2B}AR subtype, it has become increasingly clear in recent years that this AR subtype regulates a number of functions (e.g., vascular tone, cytokine release, and angiogenesis[5]). The A_{2B}ARs may also play a role in cancer therapy.[6] There is an increasing amount of data confirming that A_{2B}ARs play an important role in mediating the effects of adenosine on tumor growth and progression. The effects which are most interesting for a potential anticancer treatment based on A_{2B}ARs as a target are inhibition of angiogenesis and inhibition of ERK 1/2 phosphorylation.[6] In addition, a number of studies have demonstrated that activation of the A_{2B}AR may play an important role in asthma.[6]

The above evidence for the therapeutic applications of A_{2B}AR antagonists has stimulated the search for potent and selective ligands for this receptor subtype. Thus, a great number

Received: October 25, 2012
Published: December 1, 2013

dx.doi.org/10.1021/ci300516w | *J. Chem. Inf. Model.* 2013, 53, 3140−3155

of antagonists of the A_{2B}AR have been studied, including xanthine, deazaxanthine, adenine, 2-aminopyridine, bipyrimidine, pyrimidine, imidazopyridine, pyrazine, and pyrazolotriazolo-pyrimidine scaffolds. Of these, the xanthine derivatives represent most of the high affinity receptor antagonists.[7−9] Many potential ligands have been synthetized. However, there are no rules available for the rational design of new potent antagonists of the A_{2B}AR.

Rational drug design involves a combination of advanced experimental and computational methods. A broad variety of medicinal chemistry approaches can be used for the identification of hits, generation of leads, and optimization of leads into drug candidates. Quantitative structure−activity relationships (QSAR methods) are a potentially useful strategy that[10−12] could be used for designing potent antagonists of the A_{2B}AR.

Field-based 3D-QSAR, like Comparative Molecular Field Analysis (CoMFA), pharmacophore modeling, and docking studies, are computational methods that have been used to predict A_{2B}AR affinity.[13−15] A critical step in CoMFA studies is the positioning and subsequent alignment of the molecules, a requirement that limits their usage to homogeneous series of compounds. On the other hand, docking approaches require time-consuming optimization of molecular geometry and are not always based on real receptor modeling. Moreover, almost all the QSAR models reported so far in the prediction of A_{2B}AR affinity are limited to congeneric series of compounds, e.g. xanthines[15] deazaxanthines,[16] and they assume a linear relationship between molecular descriptors (MDs) and K_i (the affinity of the ligand for the receptor), by using multiple linear regression-type modeling.

Advances in combinatorial chemistry and high throughput screening technology have resulted in a huge growth of structural and biological data, making the development of QSAR models more challenging. This has prompted the development of fast nonlinear classification QSAR methods that can capture structure−activity relationships for large and complex data.[17,18] Nonlinear classification methods of machine learning, such as *k* Nearest Neighbors (*k*NN), Decision Tree, Support Vector Machine (SVM), and Artificial Neural Networks (ANN), have become routine tools in QSAR studies.[19,20] While several examples of the use of these methods in medicinal chemistry have been reported, only SVM has been used to predict the affinity of A_{2B}AR antagonist.[7]

Recently, there has been interest in the use of multiple classifiers ensemble systems, otherwise known as consensus or combinatorial models.[21] The use of such approaches has been demonstrated to be more accurate than single classifiers in a wide range of application areas.[22] An ensemble behaves like an expert committee in predicting the class to which a sample belongs. In most cases the class prediction is determined by a majority vote of the members of the committee or by an average of the "opinions" of committee members. Two of the most often-used methods are Bagging[23] Boosting.[24] Both methods employ the same classification model (e.g., decision tree, ANN). The efficiency of such methods comes primarily from the diversity caused by resampling the training set. Svetnik et al.[25] compared Boosting and Random Forest with other single learning machines for handling QSAR problems and found that were not always superior to other methods of classification.

Novotarskyi et al. used selection methods to generate features subsets in order to build classifiers with several machine-learning algorithms (e.g., *k*NN, decision tree, SVM) in combination with the Bagging approach.[26] Another method, Signal, was proposed in ref 27. Signal creates a classifier ensemble of significant descriptors chosen from a much larger property space, which showed better performance than other methods. Dutta et al. used different machine learning methods to make an ensemble to build QSAR models, and feature selection is used to produce different subsets for different machine learning methods.[28]

The Combi-QSAR approach, developed by Tropha et al.,[17,29,30] explores all possible combinations of various descriptor types and different learning machines. This method selects models characterized by high accuracy in predicting both training and test sets to make consensus predictions. This combinatorial methodology uses multiple training and test sets of chemicals for model development and validation. Recently Pérez-Castillo et al.[31] proposed the GA(M)E-QSAR algorithm which combines the search and optimization capabilities of Genetic Algorithms with the simplicity of the Adaboost (a variant of Boosting) ensemble-based classification algorithm to solve binary classification problems. This ensemble guarantees the diversity of the classifiers by using a distance criterion proposed by Todeschini that takes into account the correlation of original variables within and between models and allows the discovery of clusters of similar models.[32]

While the above combinatorial methods for QSAR modeling have proved useful, many are not based on the use of explicit diversity measures. Diversity is the degree to which classifiers make different decisions relating to the same problem. Classifiers with high accuracy and low diversity are more likely to make the same correct decision on the same cases (see Kuncheva et al.[22,33] for discussion), which may not improve the global performance of the ensemble.

On the other hand, in QSAR methods there are many molecular features and a few training cases. This means the data set contains redundant information which, if used to train a model, would lead to overfitted models that lack generalization capabilities. A common practice, therefore, is to employ feature selection techniques to select relevant descriptors to the property to describe.[26,28,34]

This paper presents data on the development and validation of a novel and fully automatic algorithm based on a classifier ensemble for predicting the affinity antagonists for binding to the A_{2B}AR. It uses feature selection techniques to produce different training sets (the same cases with different subsets of features) for training multiple classification models. All training sets were trained with classifiers such as *k*NN, decision tree, ANN, and SVM. To select the base classifiers (members of an ensemble), several diversity measures were used. The final multiclassifier prediction results were computed from the output obtained by using a combination of selected base classifiers, by utilizing different mathematical functions.

COMPUTATIONAL METHODS

Data Set. The data set was retrieved from the literature[16,35−48] (Table S1 of the Supporting Information). The A_{2B}AR antagonist data set contains 381 compounds. The binding data were drawn from different sources, but the protocols for competitive radioligand binding were very similar. Among all of the binding data for specific targets, only measurements using human A_{2B} adenosine receptor subtype cloned in HEK-293 cells and [^{3}H]DPCPX, as the radiolabeled ligands, were considered. The cloning, sequencing, expression,

and membrane preparation were carried out by standard techniques. The binding data were analyzed by nonlinear regression analysis, and IC_{50} values were converted to K_i values by using the Cheng-Prusoff equation.[4] Expression of the data in this way ensured that the biological end points from different sources are comparable and can be used in a combined data set. The range of binding affinities varied from 1.00 nM to >10000 nM (K_i) for the A_{2B}AR data set.

The compounds were first divided into three classes according to their K_i values. The first class, designated as potent antagonists, included all chemicals with a K_i < 100 nM (pK_i > 7). The second class, named as moderately potent antagonists, included those compounds with a K_i of 100 nM ≤ K_i < 1000 nM (pK_i = ⌈7; 6⌉). Finally the third class, named as weak antagonists, was formed by compounds with a K_i of ≥1000 nM (pK_i ≤ 6). As a result of this categorization, 204 compounds are potent antagonists, 95 chemicals were moderately potent antagonists, and 82 were weak antagonists (Table 1).

Table 1. Data Distribution Indicating the Number of A_{2B}AR Antagonists by Classes (Potent, Moderately Potent, and Weak Antagonists) in the Training Data and Test Sets

classes	training	test	total per row
potent antagonists	161	43	**204** (53.54%)
moderately potent antagonists	79	16	**95** (24.93%)
weak antagonists	66	16	**82** (21.52%)
total per column	**306** (80.31%)	**75** (19.69%)	**381**

To obtain validated QSAR models, the data were divided into training and test sets by using cluster analysis (Joining Tree Clustering and *k*-Means Cluster Analysis). The training data set was used for model building; while the test set was only employed for external model validation. Thus, the test set was considered as an external set.

Cluster Analysis. Choosing the suitable variables to be used in the cluster strategy can be a problem due to the availability of a large number of MDs. However, we overcame this problem by using a linear dimensionality reduction technique so that as much useful structural information as possible can be retained to be used in the clustering technique without rejecting any descriptor in advance. This was achieved by using Principal Component Analysis (PCA) implemented in the DRAGON software (version 5.4).[49] The PCA was performed separately for 16 different subsets of descriptors, specifically for the first 16 descriptor blocks implemented in the DRAGON software (version 5.4).[49] Thus, 134 principal components produced by PCA were standardized. After that, they formed the input for the cluster analysis. Joining Tree Clustering (JTC) was then applied to the whole data set of compounds to set the appropriate number of clusters, using, as a criterion of linking, the complete linkage, implemented in the STATISTICA software version 8.0.[50] The distances between clusters were determined by the greatest distance between any two objects in the different clusters ("furthest neighbors"), while the distances between cases were computed by the Euclidean distance.

According to this, Figure 1 shows that the appropriate number of clusters is 5. The optimal number of clusters has been highlighted for a better visualization with a red line cutting the branches.

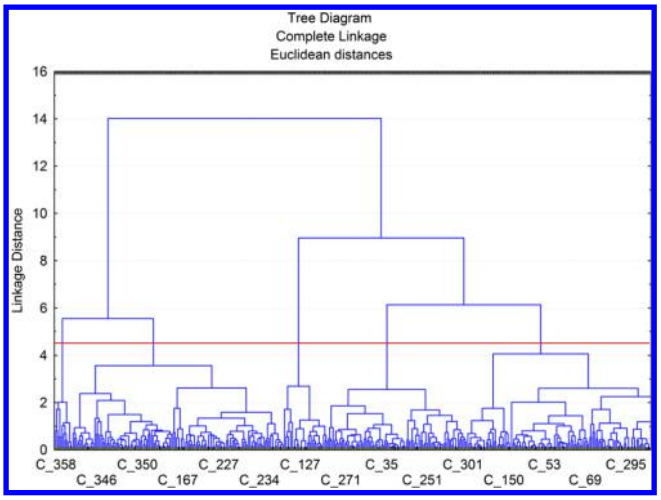


Figure 1. Graph of Joining Tree Clustering (JTC) Technique applied to the whole data set of compounds to set the appropriate number of clusters.

Then *k*-Means Cluster Analysis (*k*-MCA) was performed to design the training data and external set. By using this method it was possible to obtain exactly 5 different clusters of greatest possible distinction which could be assessed by comparing the means of the resulting 5-clusters. Another indicator of the good performance of this technique is the magnitude of the Fisher ratio values from the analysis of variance developed on each dimension (Table S2 of the Supporting Information), calculated with STATISTICA, version 8.0.[50]

Judging from the magnitude (and significance *p*-levels) of the F values derived from analysis of variance, the first principal components of constitutional descriptors (PC01-01), topological descriptors (PC02-01), connectivity indices (PC04-01), information indices (PC05-01), Burden eigenvalues (PC08-01), eigenvalue-based indices (PC10-01), and geometrical descriptors (PC12-01) are the major criteria for assigning cases to clusters (Table S2 of the Supporting Information). Setting up these conditions, the *k*-MCA and JTC techniques yielded 5 different clusters, with 117, 107, 68, 57, and 32 members, respectively.

After the partition of the whole data into different statistically representative clusters of compounds, the external set was selected by leaving out 20% of cases (specifically one compound for every five compounds) in each cluster after the members of each cluster were sorted by the Euclidean distance criterion. Table 1 reports the data distribution indicating the number of A_{2B}AR antagonists by classes (potent, moderately potent, and weak antagonists) in the training and external sets.

Molecular Descriptors (Features). *Molecular Structures.* The structures of all antagonists were optimized before the calculation of the 3D MDs. The optimized conformations were generated from the SMILES strings using the 3D conversion function and energy minimization by means of the quantum-mechanics semiempirical Parametric Method Number 6 (PM6) as implemented in the MOPAC 7.0 program.[51] After this, a frequency calculation was made on each optimized structure. None of the final output files produced imaginary frequencies, indicating that all the structures of our database are, at least, in a local minimum. This minimum-energy structure selected by the MOPAC program was considered as prototype of the "bioactive conformation". We decided to adopt this conforma-

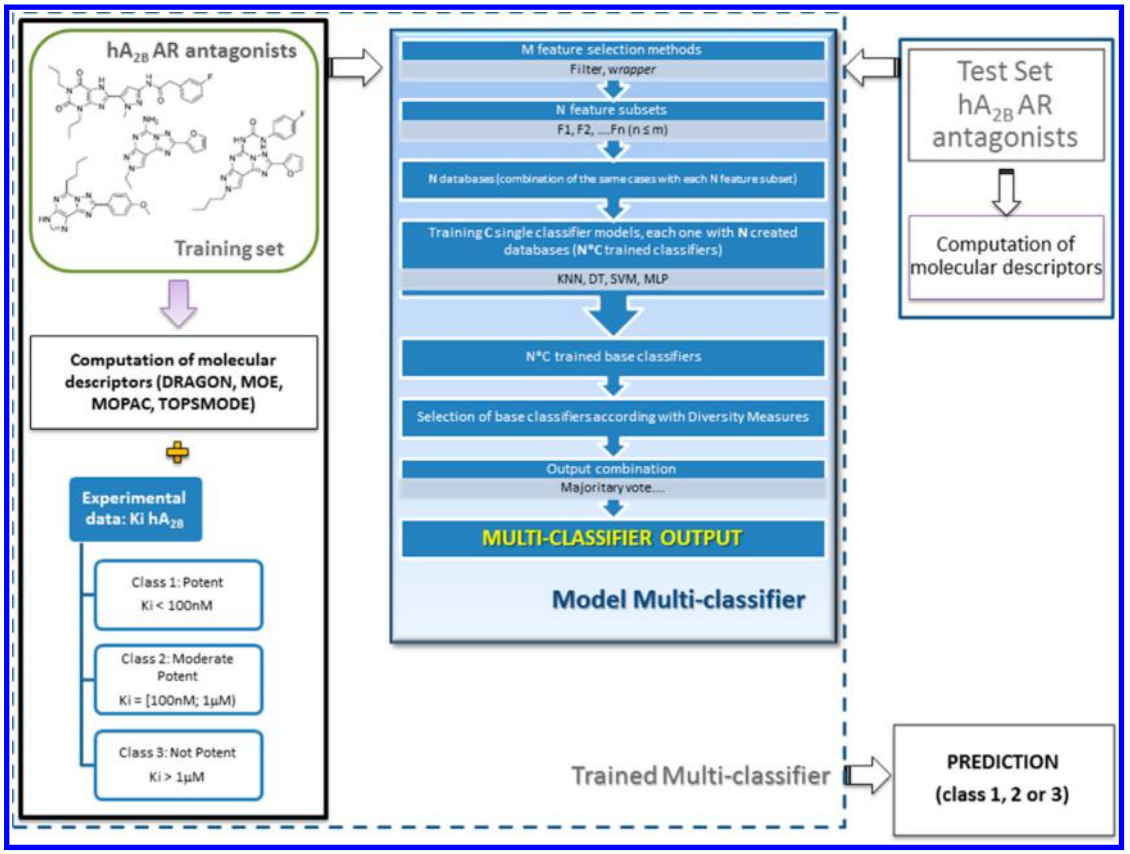


Figure 2. The description of our multiclassifier and how it is included in our QSAR methodology.

tional selection criterion for two reasons; a) the identification of the global minimum-energy conformation using a conformational analysis is computationally expensive and it may not necessarily correspond to a biologically active 3D structure and b) the information about the possible binding mode of antagonists to the human A2B receptor is very limited. Therefore, regardless of the limitations of this selection criterion, it may be considered as a reasonable approach to standardize the conformational selection.

MOE Descriptors. MOE descriptors include both 2D and 3D MDs. 2D descriptors include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices; adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors.[10] The 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors. 184 2D and 68 3D MDs were calculated (252 total) using MOE 2007.09 software.[52]

DRAGON Descriptors. DRAGON descriptors were classified into 0D, 1D, 2D, and 3D descriptors. The version 5.4 of the DRAGON software generated a total of 1455 descriptors, covering a wide variety of types. For example, its 0D descriptors contained constitutional descriptors; 1D descriptors included functional group counts and atom-centered fragments; 2D descriptors included topological descriptors, connectivity indices, information indices, and eigenvalue based indices.[10] The 3D-DRAGON descriptors were Randić molecular profiles, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors,[10] GETAWAY descriptors,[53,54] and geometrical descriptors.[10] All descriptors were cleaned up by eliminating the constant variables and near-constant variables by using the built-in function of the DRAGON software (version 5.4).[49] The pairwise correlations for all descriptors were examined, and, for each comparison, one of the two descriptors with the correlation coefficient R of 0.95, or higher, was excluded from further analysis.

MOPAC Descriptors. A total of 14 quantum-chemical descriptors, including energy of highest occupied, molecular orbital (EHOMO) and lowest unoccupied molecular orbital (ELUMO), dipole, ionization potential, hardness, and heat of formation, were generated by using MOPAC 7.0[9] after the PM6-optimization procedure.

TOPS-MODE Descriptors. MDs were computed following the TOPological Substructural MOlecular Design (TOPS-MODE) theoretical approach,[55−57] implemented in the MODESLAB software (version 1.5) available at the following Internet site: http://www.modeslab.com.

Two hundred and twenty-six (226) MDs were computed. More specifically, the first 15 spectral moments (μ_1 −μ_{15}) for each bond weight, and the number of bonds in the molecules (μ_0), excluding the hydrogen atoms, were computed.

In summary, DRAGON, MOE, MODESLAB, and MOPAC software packages were used for calculation of MDs. Each computer program generated data with a different number and type of features for the same chemical. In such cases, the data from each program were called databases (e.g. DRAGON, MOE, MODESLAB, and MOPAC databases).

New Variant of Classifier Ensemble. In this article we propose a new multiclassifier approach for predicting the affinity of A_{2B} adenosine receptor antagonists. This multiclassifier is based on two different ideas: (1) use of different classifier models and (2) use of different training sets to develop the classifier models. The training sets contain the

same cases but different descriptor subsets, which were selected by using multiple feature selection techniques. Different classifiers were trained with these training sets. To select the base classifiers, several diversity measures were used.

Figure 2 describes our multiclassifier and how it is included in our methodology. We adopted the following four-step procedure for establishing our ensemble.

Step 1: Feature Selection To Generate Different Training Sets. This proceeds as follows:

1.1. Select and apply M feature selection methods.

1.2. Return the N feature subsets (i.e. F1, F2..., Fn).

1.3. Obtain N database. The task is to obtain N different training sets that combine the same cases with each N feature subset selected in 1.2.

Step 2: Base Classifiers Selection. This proceeds as follows:

2.1. Train base classifiers. The base classifiers are trained by using C classifier models and each database obtained in 1.3, resulting in N*C trained base classifiers.

2.2. Use multiple diversity measures to select base classifiers. First, groups resulting in combinations of K classifiers are created, and the diversity measures of each group are calculated. The values of K chosen were 3 and 5, in order to use as few classifiers as possible to minimize the complexity of the model. Finally, M groups with the highest diversity measures are selected by taking into account the consensus of all diversity measures because each of them can give different results. It should be highlighted that the base classifiers diversity metrics are computed on the training set. Please see Diversity Measure below for more details.

Step 3: Output of Multiclassifier. For the combination of base classifier outputs, different mathematical functions can be used, such as majority vote, average, product, min, and max.

Step 4: Validation Procedure. Two kinds of diagnostic statistical tools are suggested for evaluating the predictive ability of the multiclassifier model internal and external validation (see the Performance Measures and Evaluation section).

In summary, we propose the use of a new multiclassifier model that transforms the original training data into different training sets with the same chemicals but with different descriptors (subsets of features). To obtain the descriptor subsets some feature selection techniques were applied (please see Feature Selection below). The reduction feature process permits simpler training of the classifiers. With these subsets, several base classifiers were trained. Some diversity measures were used to evaluate all possible combinations of individual classifiers in groups of size K. In order to use different kinds of diversity measures, we averaged all diversity measures applied to groups of K classifiers. Then, the subsets of bases classifiers with more average diversity were selected. The final multiclassifier predictions were obtained from the combination of base classifier outputs using different methods such as majority vote, average, product, min, and max.

The following sections describe the machine learning techniques and software used to build the ensemble.

Machine Learning Techniques. *Weka.* Weka (version 3.6; Waikato Environment for Knowledge Analysis), a software developed at the Waikato University, New Zealand, and available at http://www.cs.waikato.ac.nz/ml/weka/index.html[58] was used. Weka provides implementations of machine learning algorithms and some tools for transforming data sets. It also allows a straightforward way to experiment and test different learning algorithms and is also a good mode to add new methods, in view of its open source code. The algorithms we used are included in Weka, and our proposed multiclassifier was implemented by using additions to this source code.

Feature Selection. A lot of feature (descriptors) selection methods exist in the literature.[59] These techniques can be divided in three categories: filter, wrapper, and embedded techniques. In this study we used those of them, which are implemented in Weka software, namely the following: CfsSubsetEval and WrapperSubsetEval (described below). In both cases, the search methods employed were as follows: best first and genetic algorithm (named Genetic Search in Weka), resulting in 4 selection techniques (CfsSubsetEval + best first, CfsSubsetEval + genetic search, WrapperSubsetEval + best first, and WrapperSubsetEval + genetic search).[58] Feature selection methods were only executed on the training data, as opposed to the entire data.

CfsSubsetEval is a based-filter technique. It assesses the predictive ability of each attribute individually and the degree of redundancy among all of the attributes, preferring sets of attributes that are highly correlated with the class but which have low intercorrelation. *WrapperSubsetEval* is a based-wrapper technique that uses a classifier to evaluate features sets. However it employs a cross-validation scheme to estimate the accuracy of the classifier for a set of features. The classifiers used are described below.

Single Classifiers. Different simple classifiers were generated by using Weka software.

k Nearest Neighbors (kNN). This is a lazy learning method in which objects are classified from closest training examples in the feature space, based on Euclidean distance.[60]

Decision Tree (J48). J48 is an implementation of the C4.5 algorithm in Weka, which builds decision trees from a training set based on the criterion of normalized information gain.[61]

Support Vector Machine (SVM). SVM is a supervised machine learning technique proposed by Vapnik, which originates from statistical learning theory.[62] It can be divided into lineal and nonlinear SVM: the latter being obtained by the introduction of kernel, the most widely used being polynomial and Radial Basis Function (RBF).

$$\text{Polynomial: } k(x, x') = \langle x \cdot x' \rangle^d \tag{1}$$

$$\text{Radial Basis Function: } k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{2}$$

Multilayer Perceptron (MLP). The MLP is a feed-forward artificial neural network model, which uses a Back-propagation algorithm for training.[63]

Multiclassifiers. Combinations of multiple classifiers have been found to be consistently more accurate than a single classifier. For that reason several multiclassifiers have also been developed in this work, i.e. Vote, Bagging, and Boosting.

Vote. When combining the results of several methods, an intuitive approach is to make a vote or average of the opinions or results. This represents a straightforward way to combine the results in multiclassifiers and is implemented in Weka by using 'Vote'. The Vote algorithm uses several classification models with the same training set. The output combination is based on the vector of distribution probability obtained for every classifier: averaging them or using majority vote or selecting minimum, maximum, or mode.[58]

Bagging.[23] Bagging was one of the first multiclassifiers to be developed. It is also known as bootstrap aggregating and comprises classifiers built from different training sets. The

training sets are generated randomly, and this is accompanied by replacement from the original training set. It guarantees the diversity necessary to make the ensemble work. In addition, the base classifiers should be unstable, that is, small changes in the training set should lead to large changes in the classifier output.[58] Bagging employs the same classification model (e.g., J48 or MLP), and the classifier outputs are combined by the majority vote.

Boosting. Like Bagging, Boosting aggregates multiple classifiers generated by using the same learning algorithm and different training sets, created by sampling with replacement from the training instances. This multiclassifier also uses a vote output combination. However, in the boosting algorithm, a weight is assigned to each compound in the training set, after the first base classifier is trained. The weights are increased in each iteration for the cases misclassified by previous classifiers, thereby increasing the likelihood of those cases being members of the training set. Several implementations of Boosting have been developed, although, the most widely used is Adaboost.[24]

Diversity Measures. It is well-known that an ensemble would only be more successful than individual classifiers if the individual classifier outputs disagree with each other. Such independence among classifiers is known as the diversity. The diversity of the individual classifier outputs is therefore a vital requirement for the success of the ensemble.[33] There are many diversity measures, which can be categorized into two types: pairwise and nonpairwise. The first is defined between two classifiers, while the second is defined by the whole set of classifiers.[33] In this paper some diversity measures are used in order to select the appropriate base classifiers. They are related as described below.

Pairwise Measure. The Q statistic, correlation coefficient (ρ), disagreement measure (D), double fault measure (DF), and a combination of D and DF (R) were used as pairwise diversity measures.[33]

The pairwise measures consider a pair of classifiers (Ci, Cj) at a time. An ensemble of L classifiers produced $L(L-1)/2$ pairwise measures, which were averaged to get an overall diversity value for the ensemble (see Table 2). The entries in Table 2 are 1 or 0 if a case is recognized correctly or incorrectly, respectively.

Table 2. 2 × 2 Table of the Relationship between a Pair of Classifiers for One Case[a]

	Cj correct (1)	Cj incorrect (0)
Ci correct (1)	a	b
Ci incorrect (0)	c	d

[a]Total: $a + b + c + d = 1$.

Considering the pair of classifiers, by adding a, b, c, and d values for all cases, it is possible to obtain results (Table 3) from which the pairwise measures were calculated.

Table 3. 2 × 2 Table of the Relationship between a Pair of Classifiers for All Cases[a]

	Cj correct (1)	Cj incorrect (0)
Ci correct (1)	A	B
Ci incorrect (0)	C	D

[a]Total: $A + B + C + D = N$, where N is the total of cases.

Eqs 3, 4, 5, 6, and 7 describe the calculations of Q, ρ, D, DF, and R, respectively

$$Q_{Ci,Cj} = \frac{A\cdot D - B\cdot C}{A\cdot D + B\cdot C},\ -1 \le Q \le 1 \tag{3}$$

$$\rho_{Ci,Cj} = \frac{A\cdot D - B\cdot C}{\sqrt{(A+B)\cdot(C+D)\cdot(A+C)\cdot(B+D)}},\quad -1 \le \rho \le 1 \tag{4}$$

$$D_{Ci,Cj} = \frac{B+C}{N} \tag{5}$$

$$DF_{Ci,Cj} = \frac{D}{N} \tag{6}$$

$$R_{Ci,Cj} = \frac{D_{Ci,Cj}}{DF_{Ci,Cj}} \tag{7}$$

Q and ρ vary between −1 and 1. Classifiers that tend to recognize the same objects correctly will have positive values of Q and ρ, and those which commit errors on different objects (more diverse) will render Q and ρ negative. Maximum diversity is obtained for $Q = 0$ and $\rho = 0$. D is the probability that the two classifiers will disagree on their decisions. Therefore, a high probability (high D value) implies a high diversity. DF is defined as the proportion of the cases that have been misclassified by both classifiers. Therefore, it is necessary to minimize this measure to increase the diversity. Finally a combination of D and DF, denoted as R, was also used as pairwise measure (eq 7). A high value of R is indicative of high value of diversity between both classifiers.

Nonpairwise Measure. The entropy (E), Kohavi-Wolpert variance (KW), and measure of inter-rater agreement (κ) were used as nonpairwise measures.[33] These measures consider all the classifiers together and calculate directly one diversity value for the ensemble.

Equations 8, 9, and 10 describe the calculations of E, KW, and κ, respectively

$$E = \frac{1}{N}\cdot\frac{2}{L-1}\sum_{j=1}^{N}\min\{(\sum_{i=1}^{L} y_{j,i}),\ (L - \sum_{i=1}^{L} y_{j,i})\},\quad y_{j,i} \in \{0,1\},\quad 0 \le E \le 1 \tag{8}$$

$$KW = \frac{1}{NL^2}\sum_{j=1}^{N} Y(z_j)(L - Y(z_j)),\quad \text{where } Y(z_j) = \sum_{i=1}^{L} y_{i,j} \tag{9}$$

$$k = 1 - \frac{\frac{1}{L}\sum_{j=1}^{N} Y(Z_j)(L - Y(Z_j))}{N(L-1)p(1-p)} \tag{10}$$

where

$$p = \frac{1}{N\cdot L}\cdot\sum_{j=1}^{N}\sum_{i=1}^{L} y_{j,i} \tag{11}$$

In eqs 8, 9, and 10, L is the number of classifiers. The values of $y_{j,i}$ represent the results of classifier i in each case j. A value of 1 for $y_{j,i}$ indicates that the classifier i has classified the case j

dx.doi.org/10.1021/ci300516w | J. Chem. Inf. Model. 2013, 53, 3140−3155

Table 4. Feature Selection Methods Together with Search Methods Applied on DRAGON, MOE, MODESLAB, and MOPAC Training Data

				training sets			
no.	feature selection method	learning algorithm	search method	DRAGON	MOE	MODESLAB	MOPAC
1	CfsSubsetEval	-[a]	Best First	39	17	21	4
2	CfsSubsetEval	-[a]	Genetic Search	595	104	103	4
3	WrapperSubsetEval	*k*NN (k = 1)	Best First	30	12	4	7
4	WrapperSubsetEval	*k*NN (k = 3)	Best First	18	8	8	7
5	WrapperSubsetEval	*k*NN (k = 5)	Best First	9	8	9	7
6	WrapperSubsetEval	J48	Best First	11	8	13	6
7	WrapperSubsetEval	SVM(poly)	Best First	16	16	12	**1**
8	WrapperSubsetEval	SVM(RFB)	Best First	5	**1**	**1**	**1**
9	WrapperSubsetEval	*k*NN (k = 1)	Genetic Search	474	123	121	6
10	WrapperSubsetEval	*k*NN (k = 3)	Genetic Search	498	57	87	5
11	WrapperSubsetEval	*k*NN (k = 5)	Genetic Search	698	135	104	8
12	WrapperSubsetEval	J48	Genetic Search	412	133	57	7
13	WrapperSubsetEval	SVM (poly)	Genetic Search	657	136	113	**2**
14	WrapperSubsetEval	SVM (RBF)	Genetic Search	772	11	26	**2**
total of descriptor in original databases				**1454**	**251**	**226**	**13**

[a]The *CfsSubsetEval* does not use classifiers. In bold are shown the discarded bases.

correctly, and a value 0 indicates the case is misclassified. The E measure is based on the idea that an ensemble is most diverse when half of the votes are correct (or incorrect), and the other half of the votes are incorrect (or correct). E varies between 0 and 1, where 0 indicates no difference, and 1 indicates the highest possible diversity. Eq 9 shows KW measure of diversity, proposed by Kohavi and Wolpert,[33] where $Y(z_j)$ denotes the number of correct votes for the case j. A high value of KW measure indicates high diversity. Also, κ, a statistic developed as a measure of inter-rater reliability, is used when different raters (classifiers in this study) assess subjects (z_j in this paper) to measure the level of agreement while correcting for chance. In eq 10, the term denoted by p is the average individual classification accuracy. The diversity decreases when κ increases.

These eight diversity measures were implemented using the Java platform and computed over different ensembles following a 10-fold cross-validation scheme, applied on the training set. All possible combinations or ensembles of K classifiers (K = 3 or 5) were generated. The diversity measures were computed for each ensemble as the average over the 10-folds. Finally, the most diverse ensembles are selected by taking into account the consensus of all diversity measures. It was done by considering not only their frequency of appearance in the top-20 ensembles (decreasingly sorted by diversity measures) but also their importance (the ranking order) according to the respective diversity measures.

Therefore, for each database (DRAGON, MOE, MODESLAB, and MOPAC) were selected 20 ensembles based solely on diversity of its base classifiers. After that, they were compared taking into account the performance measures, described in the below section, which were used to select the top performing ensemble for each database.

Performance Measures and Evaluation. There are some measures in the literature to evaluate the performance of classifiers, which are principally focused on two-class problems. Some of them can also be applied to a multiclass problem, such as the situation in this study, where there are three classes.

Our database is imbalanced with respect to the number of cases in each class (Table 1). Measures such as accuracy do not represent the reality of how many cases were correctly classified in each class.[64−67] To fix this problem, the recall for each class as a performance measure was calculated. The recall provides the percentage of correctly classified cases into each class. Equation 12 represents the recall for class i.

$$\text{recall}_i = \frac{\text{number of correctly classified cases for class } i}{\text{total number of cases of class } i} \tag{12}$$

In order to give more general results, the accuracy (eq 13) is also reported.

$$\text{accuracy} = \frac{\text{number of correctly classified cases}}{\text{total number of cases}} \tag{13}$$

For internal validation, k-fold cross-validation (k = 10) using the training set was employed. For each data set, an input−output model was developed, based on the utilized modeling technique. The model was evaluated by measuring its recall and accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). The advantage of the 10-fold cross-validation is that all examples of the database are used for both, training and testing. The disadvantage is that, due to large volumes of data, the process can be delayed[68] although our databases for the present study are of medium size while presenting a large number of features. Notice that 10-fold cross-validation was also employed in a) the feature selection using training data and b) in the base classifier selection using diversity measures.

In the external validation, a sufficiently large and representative external set of compounds was used to evaluate the performance of the models. It was not employed in the process of model development. The external set was generated by cluster analysis, as explained in the Cluster Analysis section.

RESULTS AND DISCUSSION

The main goal of this work was to find the most reliable affinity classification model for A_{2B}AR antagonist. To this end, many models were generated with several machine learning algorithms on different descriptor sets, and these were trained. In addition, new ways to combine these algorithms and to

compare the results with other classifiers and ensembles are described.

The parameters that were used as points of reference in this study are as follows:

- Database: DRAGON, MOE, MODESLAB, MOPAC
- Feature selection methods: *CfsSubsetEval* and *WrapperSubsetEval*
- Diversity measures: Pairwise measure, Nonpairwise measure
- Machine learning algorithms: *k*NN, J48, SVM, MLP
- Multiclassifiers: Bagging, Boosting, Vote, VoteFSD

The sizes of training and external sets were 306 and 75 compounds, respectively. However, the number of MDs for each compound was much higher (1944). For this reason, they were grouped according to the software used for their calculation. In this way, four databases were conformed, DRAGON, MOE, MODESLAB, and MOPAC, containing 1454, 251, 226, and 13 MDs, respectively. There are overlaps between the DRAGON database and the rest of the sets (e.g., MOE and MODESLAB), but each of them includes unique types of descriptors as well.

In the second step, *CfsSubsetEval* and *WrapperSubsetEval* were used as feature selection methods. The first of these is called the filter method, because the attribute set is filtered to produce the most promising subset before the learning process starts. The second is the wrapper method, because the learning algorithm is wrapped into the selection procedure. Thus, in this work, *WrapperSubsetEval* employs, as a learning algorithm, the same classifiers as those used to train the base classifiers. These were J48, *k*NN (k = 1, 3 and 5), and SVM.

Table 4 shows the descriptor subsets for each database after applying the feature selection methods together with the search methods (best first and genetic search). There are 14 feature selection methods and thus 14 descriptor subsets for each database, giving a total of 56 different training sets (or bases). However some of them only have 1 or 2 MDs or have identical features, and so they were discarded. Consequently, 49 out of 56 training sets (DRAGON: 14; MOE: 13; MODESLAB: 13; and MOPAC: 9) were used as input for the machine learning algorithms.

Comparison of Individual Classifiers. Multiple individual classifiers derived from machine learning techniques were obtained for the A_{2B} ligand data set. These were *k*NN (with k = 1, 3, and 5 and Euclidean distance as metric distance), J48, SVM (polynomial and RBF), and MLP (considering 3 topologies with a hidden layer each and 4, 8, and 12 neurons, respectively). As a result, 9 classifier models were generated for each descriptor subset (Table 4), and their performances were assayed by the accuracy and recall results of predictions in 10-fold cross-validation for the training and external validation set. An example of accuracy and recall results on the external set, by using MOE descriptors, is shown in Figure 3 and Table S3 of the Supporting Information.

In general, the classifiers generated after feature subset selection produced better statistical models than the classifiers obtained with all the MDs. Specifically, the use of wrapper techniques, as feature selection methods, generated better classifiers than those produced by the filter techniques (Figure 3 and Figures S1 and S2 of the Supporting Information). The best models with MOE descriptors were obtained with the *k*NN method and with k = 1 (Figure 3). The best top-single MOE-model showed an external accuracy of 82.14% and was able to correctly classify 86.70%, 68.40%, and 85.00% of the

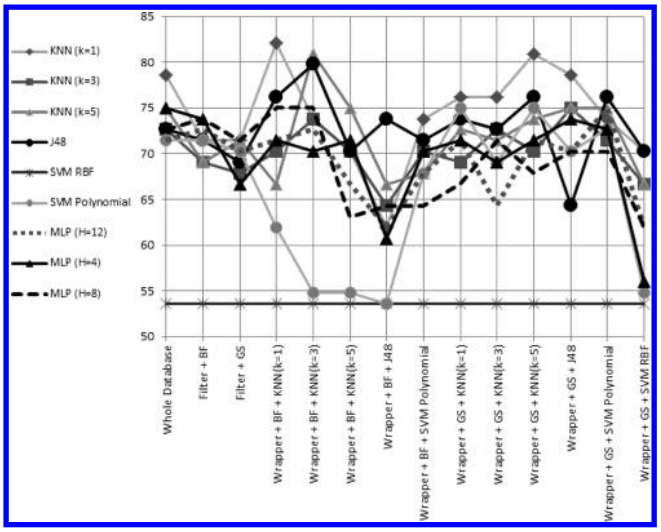


Figure 3. External accuracies of MOE-based classifiers obtained with the whole of MDs and with different descriptor subsets. The first column represents the whole database (all descriptors), and the rest are the different subsets obtained using the feature selection methods: CfsSubsetEval (Filter) and WrapperSubsetEval (Wrapper) with the respectively search method: best first (BF) and genetic search (GS).

antagonists belonging to the classes: "potent", "moderately potent", and "weak", respectively. The least predictive classifiers were developed with SVM and kernel Radial Basis Function (the highest accuracy was 53.57%). Similar behaviors were obtained with DRAGON and MODESLAB MDs (Figures S1 and S2 of the Supporting Information).

However none of the models based on MOPAC MDs had an accuracy higher than 75%. The highest accuracy values were obtained with *k*NN for k = 1 (Figure S3 of the Supporting Information). These results confirmed that *k*NN models revealed the strongest correlations between MOE, MODESLAB, and DRAGON descriptors and compound binding affinity.

Statistical tests (Friedman and Wilcoxon) were applied in order to compare the accuracy of classifiers. These tests confirmed that *k*NN (k = 1) outperforms others (p = 0). Kruskal–Wallis and Mann–Whitney tests were used to compare *k*NN in each database. From these results it is concluded that the accuracy levels of individual classifiers, derived from DRAGON, MOE, and MODESLAB database, are not statistically different (p > 0.45).

In spite of the good predictive performance of our individual models, their usage in combinations might improve their predictivities. For that reason, multiclassifiers implemented in Weka software (Bagging, Boosting, and Vote) were trained by using the same feature sets as were involved in the individual classifiers.

In the case of Bagging and Boosting, the base classifier models with the highest accuracy results were combined. As a result, 16 base classifier sets of size 10 were selected for each multiclassifier in order to decrease the ensemble computational cost. MLP, SVM, *k*NN, and J48 were used as machine learning algorithms (Table 5). Figure 4 shows the accuracy results obtained for both Bagging and Boosting and their base classifiers, built on cross-validation sets (Figure 4a) and an external set (Figure 4b). These plots show that the accuracies of the base classifiers and the multiclassifiers are very similar. Statistical tests, such as Friedman and Wilcoxon tests (p >

dx.doi.org/10.1021/ci300516w | J. Chem. Inf. Model. 2013, 53, 3140–3155

Table 5. Best 16 Ensembles Derived from Bagging and Boosting, Specifically Four Ensembles for Each Database

		feature selection method			base classifier	
ID	database	method	search method	learning algorithm	model	parameters
1	MOPAC	WrapperSubsetEval	Best First	*k*NN (k = 3)	J48	
2	MOPAC	WrapperSubsetEval	Best First	*k*NN (k = 5)	*k*NN	k = 5
3	MOPAC	WrapperSubsetEval	Best First	*k*NN (k = 5)	SVM	Polynomial Kernel
4	MOPAC	WrapperSubsetEval	Genetic Search	J48	MLP	N = 12
5	MOE	WrapperSubsetEval	Best First	*k*NN (k = 1)	*k*NN	k = 1
6	MOE	WrapperSubsetEval	Best First	*k*NN (k = 1)	MLP	N = 8
7	MOE	WrapperSubsetEval	Genetic Search	*k*NN (k = 5)	J48	
8	MOE	WrapperSubsetEval	Genetic Search	*k*NN (k = 5)	SVM	Polynomial Kernel
9	MODESLAB	WrapperSubsetEval	Best First	J48	*k*NN	k = 1
10	MODESLAB	WrapperSubsetEval	Best First	J48	J48	
11	MODESLAB	WrapperSubsetEval	Genetic Search	*k*NN (k = 5)	SVM	Polynomial Kernel
12	MODESLAB	WrapperSubsetEval	Genetic Search	*k*NN (k = 5)	MLP	N = 12
13	DRAGON	CfsSubsetEval	Genetic Search	MLP (N = 8)	MLP	N = 8
14	DRAGON	WrapperSubsetEval	Best First	*k*NN (k = 3)	*k*NN	k = 3
15	DRAGON	WrapperSubsetEval	Best First	J48	J48	
16	DRAGON	WrapperSubsetEval	Genetic Search	SVM (Polynomial Kernel)	SVM	Polynomial Kernel

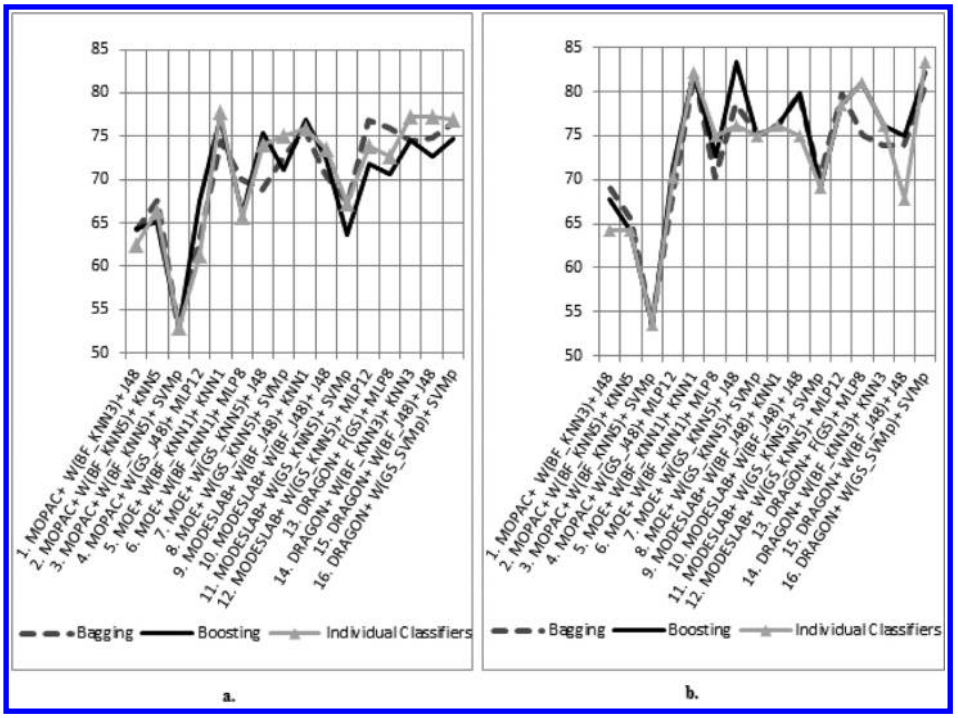


Figure 4. Accuracy results obtained for both Bagging and Boosting (using 16 ensembles) and their respective base classifiers. a. training sets from cross-validation and b. an external test set. The ensemble names in the figure are composed of the following: the name of database, the feature selection technique (search technique and learning algorithm), and the model of the base classifier. In feature selection techniques, abbreviations mean the following: W - WrapperSubsetEval; F - CfsSubsetEval; BF - Best First; GS - Genetic Search.

0.40), confirm this conclusion. This result could be attributed to the fact that Bagging and Boosting are usually most effective on unstable classifiers, such as decision trees and neural networks.[22] Bagging and Boosting were trained with stable (e.g., *k*NN, SVM) and unstable classifiers (e.g., J48, MLP), but none of them improved the accuracies (Figure 4). We believe that the main cause of this contradictory result could be related to low structural diversity of the original training set. The utility of the unstable classifier based ensemble is that small changes in the training set cause great variation in the classification.[22] However the great chemical similarity of our database and consequently of molecular descriptor values implies that no major changes occur in the predictions when light variations in the training set happen. The scaffolds in our A_{2B}AR training set fall into the following seven chemical classes: pyrazolo[4,3-e][1,2,4]triazolo[1,5-c]pyrimidines; pyrrolo[3,2-d]pyrimidine-2,4-diones; 8-(pyrazol-3-yl)-purine-2,6-diones; [1,2,4]triazolo-[1,5-c]quinazolines; purine-2,6-diones; [1,2,4]triazolo[5,1-i]-purines; and pyrimidin-2-amines (Figure 5). However there is a three-dimensional similarity between the conformers of the training set as is indicated by the range of the combo Tanimoto coefficient (0.6−1.7 for 86.3% of the training set). The combo Tanimoto coefficient represents a combination of shape matching (shape Tanimoto) and functional group matching (color Tanimoto) in the space.[69] The range of combo

dx.doi.org/10.1021/ci300516w | J. Chem. Inf. Model. 2013, 53, 3140−3155

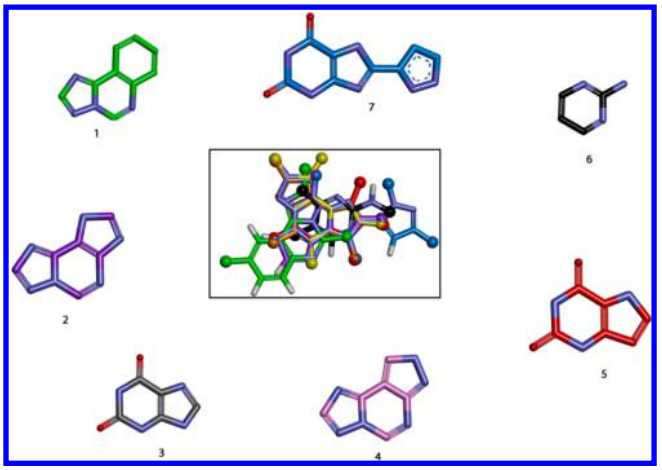


Figure 5. Scaffolds in our A_{2B}AR training set and the superposition of them. **1**) [1,2,4]triazolo[1,5-c]quinazolines; **2**) [1,2,4]triazolo[5,1-i]purines; **3**) purine-2,6-diones; **4**) pyrazolo[4,3-e][1,2,4]triazolo[1,5-c]pyrimidines; **5**) pyrrolo[3,2-d]pyrimidine-2,4-diones; **6**) pyrimidin-2-amines; **7**) 8-(pyrazol-3-yl)-purine-2,6-diones.

Tanimoto is 0−2, where 0 means "maximally dissimilar" and 2 means "maximally similar".

On the other hand, Vote used diversity measures in order to select base classifiers. It is important to note that Vote employs the same training set but several learning algorithms. For that reason, combinations of 9 single classifier models (e.g., SVM and J48) in groups of sizes 3 and 5 give 84 and 126 different base classifier combinations, respectively. Taking into account the number of descriptor subsets for each database (DRAGON: 14; MOE: 13; MODESLAB: 13; and MOPAC: 9) and considering groups of size 3 (or 5; as shown by the following numbers in parentheses), a total of 1176 (1764), 1092 (1638), 1092 (1638), and 756 (1134) multiclassifier topologies were generated by the DRAGON, MOE, MODESLAB, and MOPAC databases, respectively. In order to minimize the complexity of the Vote ensembles base classifier sets of sizes 3 and 5 were selected according to the eight diversity measures (described above). However an accuracy analysis of Vote models containing 3 and 5 base classifiers showed that they were very similar. Therefore, the results of Vote ensembles containing 3 base classifiers are only considered. Table 6 depicts the performance of the best-performing 3 Vote ensembles for each database using the cross-validation and external set. The accuracy results, considering three output combination functions, are also shown.

Table 6 shows that the Vote ensembles improve the performance of the top-single classifiers, with the exception of the Vote models (cross-validation) based on MOE and MODESLAB information. The accuracy results derived from three output combinations were not significantly different.

Our Ensemble Named VoteFSD. The name: VoteFSD is introduced here as an acronym for *Vote* based on *F*eature *S*election and *D*iversity measures. This proposed new multi-classifier also uses the individual classifiers (described earlier) as base classifiers, but in contrast with Bagging and Boosting, it combines several learning algorithms as Vote (Table 7). In this work, Vote and VoteFSD use diversity measures in order to decrease the number of base classifier combinations (Table 7). The principal difference between them is that our ensemble uses larger combinations of base classifiers than Vote (Table 7). For instance, VoteFSD makes a combination of 126 individual classifiers for the DRAGON database, in groups of size 3 and 5, while Vote combines 9 single classifier models for each training set, in groups of size 3 and 5. For the output combination, Vote and VoteFSD use majority vote, maximum, and average probability.

Figure 6, on the left, shows a comparison of the accuracy between VoteFSD and the best results (MAX) of individual classifiers used as base classifiers, specifically for MOE database using cross-validation, while on the right it displays the validation results using the external set. Different functions such as a) probabilities average (T_AVG), b) majority vote (T_MAJ), c) the max of probabilities (T_MAX), and d) the best result of the previous functions (Mult_MAX) were considered as output combinations. At least 100 VoteFSD ensembles of size 5 were analyzed for each database (Table 7). According to Figure 6, in many cases the best performances were achieved with VoteFSD ensembles. Probability average and majority vote were the output combination functions with best results, and statistical tests, like Friedman and Wilcoxon, confirm the earlier visual finding. Using the MOE database and the probability average, the 80.6% (83/103) of VoteFSD ensembles gave a better prediction for the external set than that provided by the top-single classifiers. However, comparing the best results of the different output combination showed that VoteFSD ensembles always produced the best predictions

Table 6. Performance of the Best Vote Ensemble for Each Database Using the Cross-Validation and External Set[b]

			cross-validation accuracy				external accuracy			
			BC	Vote			BC	Vote		
database	models of base classifiers	feature selection technique[a]	Max_BC	AVG	MAJ	MAX	Max_BC	AVG	MAJ	MAX
DRAGON	*k*NN (k = 1), J48, SVM (RBF Kernel)	WrapperSubsetEval (Genetic Search_*k*NN (k = 5))	70.9	**80.8**	**77.5**	**80.8**	72.6	**76.2**	72.6	**82.1**
MOE	*k*NN (k = 1), J48, SVM (RBF Kernel)	WrapperSubsetEval (Genetic Search_*k*NN (k = 1))	**77.2**	76.0	72.1	73.6	76.2	**81.0**	**82.1**	**82.1**
MODESLAB	*k*NN (k = 1), MLP (N = 8), SVM (Polynomial Kernel)	WrapperSubsetEval (Best First _J48)	**76.0**	74.5	68.5	75.4	76.2	**79.8**	75.0	75.0
MOPAC	*k*NN (k = 1), MLP (N = 8), SVM (Polynomial Kernel)	WrapperSubsetEval (Best First_*k*NN (k = 3))	64.0	**68.2**	**67.0**	**66.1**	70.2	**77.4**	65.5	**73.8**

[a]Feature selection techniques used to obtain the molecular descriptor subsets (search method and learning algorithm); BC: Base Classifiers; Max_BC: the best result obtained by the individual classifiers used as bases classifiers; AVG: output combination by AVeraGe of probabilities; MAJ: output combination by MAjority Vote; MAX: output combination by MAXimum of probabilities; N: Neuron of hidden layer; **k**: number of k in *k*NN; RBF: Radial Basis Function. In bold is represented the best accuracies in each row. [b]The accuracy results considering three output combination functions are also shown.

dx.doi.org/10.1021/ci300516w | *J. Chem. Inf. Model.* 2013, 53, 3140−3155

Table 7. Description of the Ensembles Used in This Paper

ensemble	base classifiers	training set	diversity measures	possible models[a]	trained models	final models
Bagging	same model	different training set, each of them modified by random selection	-	DRAGON: 126	all	DRAGON: 126
				MOE: 117		MOE: 117
				MODESLAB: 117		MODESLAB: 117
				MOPAC: 81		MOPAC: 81
Boosting	same model	different training set, obtained by selection with weighted cases	-	DRAGON: 126	all	DRAGON: 126
				MOE: 117		MOE: 117
				MODESLAB: 117		MODESLAB: 117
				MOPAC: 81		MOPAC: 81
Vote	different models	same training set	diversity combination technique (describe here)	[b]	20 better ensembles sorted by diversity measures	
				DRAGON: 1176 (1764)		DRAGON: 196 (209)
				MOE: 1092 (1638)		MOE: 143 (297)
				MODESLAB: 1092 (1638)		MODESLAB: 169 (311)
				MOPAC: 756 (1134)		MOPAC: 81 (113)
VoteFSD	different models	different training set with different features	diversity combination technique (describe here)	[c]	20 better ensembles sorted by diversity measures	
				DRAGON: 325500 (244 222 650)		DRAGON: 101 (107)
				MOE: 260130 (167549733)		MOE: 64 (103)
				MODESLAB: 260130 (167549733)		MODESLAB: 56 (101)
				MOPAC: 85320 (25621596)		MOPAC: 76 (100)

[a]Take into account the 9 machine learning algorithms and the number of training sets for each database: DRAGON (14), MOE (13), MODESLAB (13), and MOPAC (9). [b]Numbers of models derived from combinations of 3 base classifiers (and 5 base classifiers). The base classifiers result of applying 9 machine learning algorithms to the same training set of each database. [c]Numbers of models that result from combinations of 9 machine learning algorithms with all training sets of each database.

(Figure 6d). In Figures S4–S6 of the Supporting Information, similar behavior for the rest of the database (DRAGON, MODESLAB, and MOPAC) can be seen.

Table 8 depicts the external and internal accuracy of the best ensemble for each database and their top-single classifiers. The ensembles derived from DRAGON and MOE descriptors have the highest performances for internal accuracy (cross-validation). Accuracies higher than 80% were achieved if the probability average or majority votes are considered as an output combination. These results suggest that a combination of the descriptor sets of different structural information, e.g. 0D, 1D, 2D, and 3D (Table 8 and Table S4 of the Supporting Information), contributes relevant new information to the model and increases its performance. Both, DRAGON and MOE databases included 3D information. Therefore, the high performance values for models including DRAGON/MOE descriptors may demonstrate the importance of 3D information for modeling affinity of A_{2B} adenosine receptor antagonists although the optimization of 3D structures can be a limiting step and can significantly increase the computational cost for application of models using these sets of descriptors. Table 8 results also confirm that the VoteFSD ensembles perform better than the top-single classifiers for both internal and external sets, and the probability average and majority vote are the best output combinations.

The best VoteFSD models were based on MOE and DRAGON databases (Table 8). Both ensembles show good values for internal and external accuracy. However the VoteFSD-DRAGON uses a bigger number of MDs than VoteFSD-MOE, which significantly increases the training time. For that reason, the MOE ensemble is proposed as the top-best VoteFSD model. The best internal and external accuracies are achieved by this ensemble when majority vote is used as output combination (83.7% and 86.9%, respectively). This ensemble combines two *k*NN (k = 1 and k = 5), one decision tree (J48), and two artificial neural networks with 8 neurons in a hidden layer (Table 8).

Table 9 shows the recall results (calculated for the three classes) of the VoteFSD-MOE and its base classifiers. Analysis of the data in Table 9 shows the following: a) the ensemble achieves better recall results than the top-single classifiers for both: internal and external sets; b) in general, the majority vote is the best output combination; and c) for the ensemble, the classes "potent antagonists" and "weak antagonists" yielded the best percentage of correct classifications (>80%), while the class "moderately potent antagonists" yielded the worst of all (>67%). The latter result means that the "moderately potent antagonists" class is the most variable classification, probably because chemicals belonging to this class can be on the borderline between "potent antagonists" and "weak antagonists".

The VoteFSD ensembles perform better compared with the popular multiclassifiers. Given that Bagging and Boosting do not improve the performance of single classifiers whereas Vote does enhance predictivity, only Vote was used in the comparison with our multiclassifier. In contrast with Vote, VoteFSD performed better if groups of base classifier of size 5 were considered. For that reason, in this subsection our results

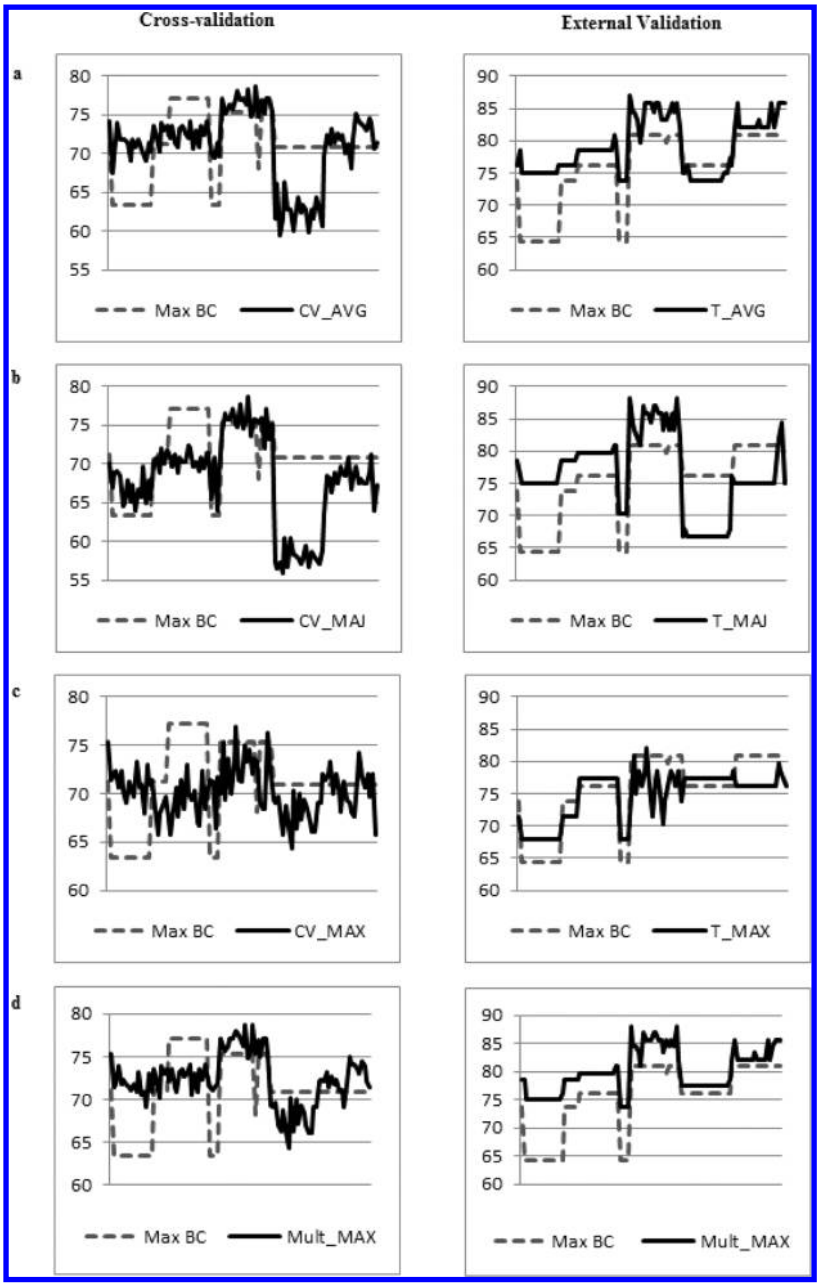


Figure 6. Comparison of accuracy results from the database MOE using the best results of the individual classifiers (Max) vs the VoteFSD using several output combinations: a. probability average (T_AVG), b. majority vote (T_MAJ), c. maximum of probabilities (T_MAX), d. maximum of three previous outputs (Mult_MAX).

are based on the best multiclassifiers, without considering the ensemble size. The internal and external accuracy of the top-performing models obtained from the Vote and the VoteFSD, using the database DRAGON and MOE, are shown in Figure 7. The probability average (AVG), majority vote (MAJ), and maximum of probabilities (MAX) were employed as output combinations.

It can be seen that, in general, our new multiclassifier performed better than the Vote. For example, the best VoteFSD ensemble exhibited 83.7% and 86.9% internal and external accuracy, respectively. While the top-Vote model managed 80.02% and 82.1% on the internal and external sets, respectively.

In summary, VoteFSD proposes a new methodology to combine single classifiers. The practical application of this new ensemble is when high dimensional databases are analyzed. Frequently, in QSAR studies, one is confronted with many more MDs that compounds in data sets. To solve this problem, feature selection methods can be used to identify relevant chemical information for the specific biological end point. In this context, VoteFSD reduces the data dimensionality by using different feature selection techniques. The selected descriptor sets are used to train several classifiers and then combine them in a logical form.

■ CONCLUSIONS AND FUTURE RESEARCH

In this paper a methodology of ensemble classifiers based on feature selection and diversity measures has been proposed for predicting A_{2B} adenosine receptor affinity. We have compared the results of using some individual classifiers (i.e., *k*NN, J48,

dx.doi.org/10.1021/ci300516w | *J. Chem. Inf. Model.* 2013, 53, 3140−3155

Table 8. External and Internal Accuracy of the Best VoteFSD Ensemble and the Performance of the Top-Single Classifiers[c]

				cross-validation accuracy				test accuracy			
				BC	VoteFSD			BC	VoteFSD		
database	models of base classifiers	feature selection technique[a]	molecular descriptor information (no. of features)[b]	Max_BC	AVG	MAJ	MAX	Max_BC	AVG	MAJ	MAX
DRAGON	*k*NN (*k* = 1)	WrapperSubsetEval (BF_*k*NN (*k* = 5))	0D, 1D, 2D, and 3D (9 + 698 + 595 + 5 + 657)								
	MLP (*N* = 4)	WrapperSubsetEval (GA_*k*NN (*k* = 5))									
	MLP (*N* = 8)	CfsSubsetEval (GA)		76.9	**83.3**	**84.4**	**83.1**	83.3	**86.3**	**86.9**	84.5
	MLP (*N* = 8)	WrapperSubsetEval (BF_SVM (RBF Kernel))									
	SVM (Polynomial Kernel)	WrapperSubsetEval (GA_SVM (Polynomial Kernel))									
MOE	*k*NN (*k* = 1)	WrapperSubsetEval (GA_*k*NN (*k* = 5))	2D and 3D (135 + 8+12 + 135 + 12)								
	*k*NN (*k* = 5)	WrapperSubsetEval (BF_*k*NN (*k* = 3))									
	J48	WrapperSubsetEval (GA_*k*NN (*k* = 5))		75.4	**82.4**	**83.7**	77	81.0	**86**	**86.9**	84
	MLP (*N* = 8)	WrapperSubsetEval (BF_*k*NN (*k* = 1))									
	MLP (*N* = 8)	WrapperSubsetEval (BF_*k*NN (*k* = 3))									
MODESLAB	*k*NN (*k* = 1)	WrapperSubsetEval (BF_*k*NN (*k* = 1))	2D (4 + 13 + 26 + 57 + 12)								
	*k*NN (*k* = 3)	WrapperSubsetEval (BF_J48)									
	*k*NN (*k* = 5)	WrapperSubsetEval (GA_SVM (RBF Kernel))		73.8	**78.4**	**75.7**	**75**	79.8	**81**	**83.4**	79.8
	J48	WrapperSubsetEval (GA_J48)									
	MLP (*N* = 8)	WrapperSubsetEva (BF_SVM (Polynomial Kernel))									
MOPAC	*k*NN (*k* = 1)	WrapperSubsetEval (GA_*k*NN (*k* = 1))	Quantum Mechanic (6 + 5 + 6 + 7 + 7)								
	*k*NN (*k* = 1)	WrapperSubsetEval (GA_*k*NN (*k* = 3))									
	J48	WrapperSubsetEval (BF_J48)		65.2	**70.7**	**70.7**	66	75	**76.9**	77	77
	MLP (*N* = 12)	WrapperSubsetEval (BF_*k*NN (*k* = 5))									
	MLP (*N* = 4)	WrapperSubsetEval (GA_J48)									

[a]Feature selection techniques used to obtain the molecular descriptor subsets (search method and learning algorithm). [b]Number of features obtained by each feature selection technique, that means the number of features to represent the different databases used as a training set by each base classifier. [c]BC: Base Classifiers; VoteFSD: the new multiclassifier proposes here for first time; Max_BC: the best result obtained by the individual classifiers used as bases classifiers; AVG: output combination by AVeraGe of probabilities; MAJ: output combination by MAjority Vote; MAX: output combination by MAXimum of probabilities; N: Neuron of hidden layer; *k*: number of *k* in *k*NN; RBF: Radial Basis Function. In bold are represented the best accuracies.

Table 9. External and Internal Recall Results for Three Classes Using VoteFSD-MOE and Its Top-Base Classifiers[b]

topology of VoteFSD-MOE			cross-validation/recall				test/recall			
			BC	VoteFSD-MOE			BC	VoteFSD		
models of BC	feature selection technique[a]	class	Max_BC	AVG	MAJ	MAX	Max_BC	AVG	MAJ	MAX
*k*NN (*k* = 1)	WrapperSubsetEval (GA_*k*NN (*k* = 5))									
*k*NN (*k* = 5)	WrapperSubsetEval (BF_*k*NN (*k* = 3))	potent antagonist	73.9	**88.2**	88.8	80.1	77.8	90.7	**90.9**	81.4
J48	WrapperSubsetEval (GA_*k*NN (*k* = 5))	moderately potent antagonist	64.7	**73.4**	74.7	67.1	68.4	75.0	73.7	**81.3**
MLP (*N* = 8)	WrapperSubsetEval (BF_*k*NN (*k* = 1))	weak antagonist	76.4	**80.3**	86.3	83.3	90.0	87.5	88.9	**93.8**
MLP (*N* = 8)	WrapperSubsetEval (BF_*k*NN (*k* = 3))									

[a]Feature selection techniques used to obtain the molecular descriptor subsets (search method and learning algorithm). [b]BC: Base Classifiers; VoteFSD-MOE: the new multiclassifier proposes here using MOE descriptors; Max_BC: the best result obtained by the individual classifiers used as bases classifiers; AVG: output combination by AVeraGe of probabilities; MAJ: output combination by MAjority Vote; MAX: output combination by MAXimum of probabilities; N: Neuron of hidden layer; *k*: number of *k* in *k*NN. In bold are represented the best recalls.

SVM, MLP) and multiclassifiers (Vote, Bagging, and Boosting). Although significant improvements were achieved by using simple classifiers and the Vote multiclassifier, the best results were obtained with the new method, which we have called

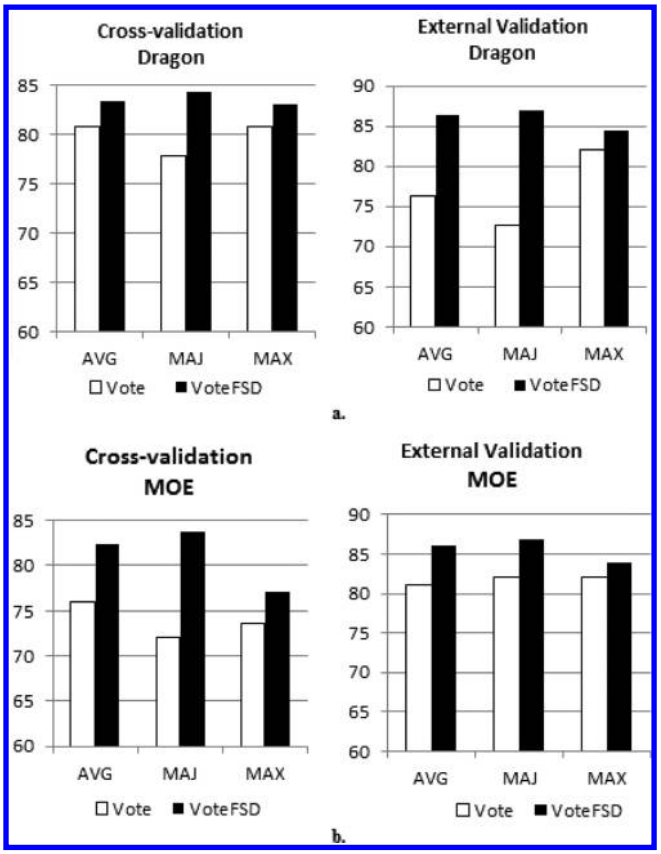


Figure 7. Internal and external accuracy of the top-performing models obtained from the Vote and the VoteFSD. a. These results are based on the DRAGON Database and are referred to as VoteFSD vsVote. b. These results are based on the MOE Database and are referred to as VoteFSD vsVote.

VoteFSD. This multiclassifier trains different classifier models with training sets. The training sets contained the same cases but different descriptors, which were selected by feature selection techniques and base classifiers were selected taking into account diversity measures.

The VoteFSD, which was implemented in Java using Weka version 3.6, uses two the CfsSubsetEval and WrapperSubsetEval feature selection algorithms and *k*NN, J48, SVM (called SMO), and MLP (MultilayerPerceptron) classifier models.

The data showed the following:

- The single classifiers, generated after feature subset selection, produced more-predictive statistical models than the classifiers obtained with the whole MDs.
- The best single models were obtained with *k*NN ($k = 1$) as the machine learning algorithm and using the Wrapper method for feature selection.
- The results obtained with Vote gave superior of single models, while Bagging and Bossting did not improve the performance of individual classifiers.
- The best predictive results were attained with VoteFSD by using the MOE database and majority vote as the output combination. The internal and external accuracies were 83.7% and 86.9%, respectively. This ensemble combines two *k*NN ($k = 1$ and $k = 5$), one decision tree (J48), and two artificial neural networks (MLP) with 8 neurons in hidden layer. The feature selection techniques used the Wrapper method.

Further issues to be explored are as follows:

1. To combine all MDs obtained by different software (e.g., DRAGON, MOE, MODESLAB, MOPAC, etc.) in the same multiclassifier.
2. To evaluate the new multiclassifier method in other medicinal chemistry problems.

■ ASSOCIATED CONTENT

Supporting Information

Training set and external set (Table S1), analysis of variance in the cluster analysis (Table S2), external Recall of MOE-based classifiers obtained with all the MDs and with different descriptors subsets (Table S3), lists of the most frequent descriptors (Table S4), accuracies of single classifiers based on DRAGON (Figure S1), MODESLAB (Figure S2), and MOPAC (Figure S3) obtained with all the MDs and with different descriptor subsets. Comparison between the individual classifiers (Max) and VoteFSD using DRAGON, MODESLAB, and MOPAC database and several output combinations (Figures S4−S6). Internal and external accuracy of the top-performing models obtained from the Vote and the VoteFSD and using MOE, MODESLAB, and MOPAC (Figure 7S). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +351 220402560. Fax: +351 220402659. E-mail: aliuskamhelguera@yahoo.es.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Foundation for Science and Technology (FCT), Portugal, Universidade de Vigo and Xunta de Galicia (CN2012/184) and AECID (projects PTDC/QUI-QUI/113687/2009, 09VIA05, 11VIA18 and D/024153/09) are acknowledged for financial support. The authors acknowledge to the FCT grant (SFRH/BPD/63946/2009) and the Dr. Maykel Cruz Monteagudo, Dr. Yunierki Pérez Castillo, and Dr. Reinaldo Molina Ruiz for valuable technical contribution to this work. We also extend our most heartfelt thanks and appreciation to Dr. Robert D. Combes and Dr. Stephen J. Barigye for their editorial assistance with the manuscript. Finally, the authors thank the anonymous reviewers of the original version of the manuscript for highly relevant comments and suggestions that have helped us significantly to improve and enrich the manuscript.

■ REFERENCES

(1) Kaiser, S. M.; Quinn, R. J. Adenosine receptors as potential therapeutic targets. *Drug Discovery Today* **1999**, *4*, 542−551.

(2) Gao, Z. G.; Jacobson, K. A. Emerging adenosine receptor agonists. *Expert Opin. Emerging Drugs* **2007**, *12*, 479−492.

(3) Fredholm, B. B.; IJzerman, A. P.; Jacobson, K. A.; Klotz, K. N.; Linden, J. Nomenclature and classification of adenosine receptors. *Pharmacol. Rev.* **2001**, *53*, 527−552.

(4) Brunton, L. L. *Goodman & Gilman's The Pharmacological Basis of Therapeutics*, 11th ed.; McGraw-Hill: New York, NY, 2007.

(5) Volpini, R.; Costanzi, S.; Vittori, S.; Cristalli, G.; Klotz, K. N. Medicinal chemistry and pharmacology of A2B adenosine receptors. *Curr. Top. Med. Chem.* **2003**, *3*, 427−443.

(6) Wilson, C. N.; Mustafa, S. J. Adenosine receptors in health and disease. In *Handbook of Experimental Pharmacology*; München, F. B. H., Ed.; Springer: Berlin, 2009; Vol. *193*, pp 410−414.

(7) Michielan, L.; Stephanie, F.; Terfloth, L.; Hristozov, D.; Cacciari, B.; Klotz, K. N.; Spalluto, G.; Gasteiger, J.; Moro, S. Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. *J. Chem. Inf. Model.* **2009**, *49*, 2820−2836.
(8) Baraldi, P. G.; Tabrizi, M. A.; Gessi, S.; Borea, P. A. Adenosine receptor antagonists: translating medicinal chemistry and pharmacology into clinical utility. *Chem. Rev.* **2008**, *108*, 238−263.
(9) Moro, S.; Gao, Z. G.; Jacobson, K. A.; Spalluto, G. Progress in the pursuit of therapeutic adenosine receptor antagonists. *Med. Res. Rev.* **2006**, *26*, 131−159.
(10) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
(11) Riera-Fernández, P.; Martín-Romalde, R.; Prado-Prado, F.; Escobar, M.; Munteanu, C.; Concu, R.; Duardo-Sanchez, A.; González-Díaz, H. From QSAR models of drugs to complex networks: state-of-art review and introduction of new Markov-spectral moments indices. *Curr. Top. Med. Chem.* **2012**, *12*, 927−960.
(12) González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. Medicinal chemistry and bioinformatics−current trends in drugs discovery with networks topological indices. *Curr. Top. Med. Chem.* **2007**, *7*, 1015−1029.
(13) Gonzalez, M. P.; Teran, C.; Teijeira, M. Search for new antagonist ligands for adenosine receptors from QSAR point of view. How close are we? *Med. Res. Rev.* **2008**, *28*, 329−371.
(14) Baraldi, P. G.; Baraldi, S.; Saponaro, G.; Preti, D.; Romagnoli, R.; Piccagli, L.; Cavalli, A.; Recanatini, M.; Moorman, A. R.; Zaid, A. N.; Varani, K.; Borea, P. A.; Tabrizi, M. A. Novel 1,3-dipropyl-8-(3-benzimidazol-2-yl-methoxy-1-methylpyrazol-5-yl)xanthines as potent and selective A(2)B adenosine receptor antagonists. *J. Med. Chem.* **2012**, *55*, 797−811.
(15) Song, Y.; Coupar, I. M.; Iskander, M. N. Structural predictions of adenosine 2B antagonist affinity using molecular field analysis. *Quant. Struct.-Act. Relat.* **2001**, *20*, 23−30.
(16) Carotti, A.; Stefanachi, A.; Raviña, E.; Sotelo, E.; Loza, M. I.; Cadavid, M. I.; Centeno, N. B.; Nicolotti, O. 8-Substituted-9-deazaxanthines as adenosine receptor ligands: design, synthesis and structure-affinity relationships at A_{2B}. *Eur. J. Med. Chem.* **2004**, *39*, 879−887.
(17) Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494−3504.
(18) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P.; Olsen, L.; Jorgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37*, 658−664.
(19) González-Díaz, H.; Bonet, I.; Teran, C.; De Clercq, E.; Bello, R.; García, M. M.; Santana, L.; Uriarte, E. ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. *Eur. J. Med. Chem.* **2007**, *42*, 580−585.
(20) de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245−1254.
(21) Rodríguez, J. J.; Kuncheva, L. I.; Alonso, C. J. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal.* **2006**, *28*, 1619−1630.
(22) Kuncheva, L. I. *Combining Pattern Classifiers, Methods and Algorithms*; Wiley Interscience: New York, NY, 2004.
(23) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123−140.
(24) Freund, Y.; Schapire, R. E. Decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119−139.
(25) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.
(26) Novotarskyi, S.; Sushko, I.; Korner, R.; Pandey, A. K.; Tetko, I. V. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *J. Chem. Inf. Model.* **2012**, *51*, 1271−1280.
(27) Lanctot, J. K.; Putta, S.; Lemmen, C.; Greene, J. Using ensembles to classify compounds for drug discovery. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2163−2169.
(28) Dutta, D.; Guha, R.; Wild, D.; Chen, T. Ensemble feature selection: Consistent descriptor subsets for multiple QSAR models. *J. Chem. Inf. Model.* **2007**, *47*, 989−997.
(29) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−595.
(30) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* **2010**, *29*, 476−488.
(31) Pérez-Castillo, Y.; Lazar, C.; Taminau, J.; Froeyen, M.; Cabrera-Pérez, M. Á.; Nowé, A. GA(M)E-QSAR: A novel, fully automatic genetic-algorithm-(meta)-ensembles approach for binary classification in ligand-based drug design. *J. Chem. Inf. Model.* **2012**, *52*, 2366−2386.
(32) Todeschini, R.; Consonni, V.; Pavan, M. A distance measure between models: a tool for similarity/diversity analysis of model populations. *Chemom. Intell. Lab. Syst.* **2004**, *70*, 55−61.
(33) Kuncheva, L. I.; Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **2003**, *51*, 181−207.
(34) Helguera, A. M.; Pérez-Garrido, A.; Gaspar, A.; Reis, J.; Cagide, F.; Vina, D.; Cordeiro, N.; Borges, F. Combining QSAR classification models for predictive modeling of human monoamine oxidase inhibitors. *Eur. J. Med. Chem.* **2012**, *59*, 75−90.
(35) Baraldi, P. G.; Bovero, A.; Fruttarolo, F.; Romagnoli, R.; Tabrizi, M. A.; Preti, D.; Varani, K.; Borea, P. A.; Moorman, A. R. New strategies for the synthesis of A_3 adenosine receptor antagonists. *Bioorg. Med. Chem.* **2003**, *11*, 4161−4169.
(36) Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Klotz, K. N.; Spalluto, G.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine derivatives as adenosine receptor ligands: A starting point for searching A_{2B} adenosine receptor antagonists. *Drug Dev. Res.* **2001**, *53*, 225−235.
(37) Baraldi, P. G.; Tabrizi, M. A.; Fruttarolo, F.; Bovero, A.; Avitabile, B.; Preti, D.; Romagnoli, R.; Merighi, S.; Gessi, S.; Varani, K.; Borea, P. A. Recent developments in the field of A_3 adenosine receptor antagonists. *Drug Dev. Res.* **2003**, *58*, 315−329.
(38) Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Moro, S.; Klotz, K. N.; Leung, E.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Pyrazolo[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine derivatives as highly potent and selective human A_3 adenosine receptor antagonists: Influence of the chain at the N^8 pyrazole nitrogen. *J. Med. Chem.* **2000**, *43*, 4768−4780.
(39) Baraldi, P. G.; Cacciari, B.; Moro, S.; Romagnoli, R.; Ji, X.; Jacobson, K. A.; Gessi, S.; Borea, P. A.; Spalluto, G. Fluorosulfonyl- and bis-(β-chloroethyl)amino-phenylamino functionalized pyrazolo-[4,3-e]1,2,4-triazolo[1,5-c]pyrimidine derivatives: Irreversible antagonists at the human A_3 adenosine receptor and molecular modeling studies. *J. Med. Chem.* **2001**, *44*, 2735−2742.
(40) Baraldi, P. G.; Cacciari, B.; Romagnoli, R.; Spalluto, G.; Monopoli, A.; Ongini, E.; Varani, K.; Borea, P. A. 7-Substituted 5-amino-2-(2-furyl)pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidines as A_{2A} adenosine receptor antagonists: A study on the importance of modifications at the side chain on the activity and solubility. *J. Med. Chem.* **2002**, *45*, 115−126.
(41) Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A_{2A} and A_3 adenosine receptor antagonists. *Eur. J. Med. Chem.* **2003**, *38*, 367−382.
(42) Pastorin, G.; Da Ros, T.; Spalluto, G.; Deflorian, F.; Moro, S.; Cacciari, B.; Baraldi, P. G.; Gessi, S.; Varani, K.; Borea, P. A. Pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives as adenosine receptor antagonists. Influence of the N5 substituent on the affinity at the human A_3 and A_{2B} adenosine receptor subtypes: A molecular modeling investigation. *J. Med. Chem.* **2003**, *46*, 4287−4296.
(43) Baraldi, P. G.; Tabrizi, M. A.; Preti, D.; Bovero, A.; Romagnoli, R.; Fruttarolo, F.; Zaid, N. A.; Moorman, A.; Varani, K.; Gessi, S.;

Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of new 8-heterocyclic xanthine derivatives as highly potent and selective human A_{2B} adenosine receptor antagonists. *J. Med. Chem.* **2004**, *47*, 1434−1447.
(44) Baraldi, P. G.; Cacciari, B.; Moro, S.; Spalluto, G.; Pastorin, G.; Da Ros, T.; Klotz, K. N.; Varani, K.; Gessi, S.; Borea, P. A. Synthesis, biological activity, and molecular modeling investigation of new pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]pyrimidine derivatives as human A_3 adenosine receptor antagonist. *J. Med. Chem.* **2002**, *45*, 770−780.
(45) Okamura, T.; Kurogi, Y.; Nishikawa, H.; Hashimoto, K.; Fujiwara, H.; Nagao, Y. 1,2,4-Triazolo[5,1-i]purine derivatives as highly potent and selective human adenosine A_3 receptor ligands. *J. Med. Chem.* **2002**, *45*, 3703−3708.
(46) Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of C^9- and C^2-substituted pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c]-pyrimidines as new A_{2A} and A_3 adenosine receptors antagonists. *J. Med. Chem.* **2003**, *46*, 1229−1241.
(47) Stefanachi, A.; Brea, J. M.; Cadavid, M. I.; Centeno, N. B.; Esteve, C.; Loza, M. I.; Martinez, A.; Nieto, R.; Raviña, E.; Sanz, F.; Segarra, V.; Sotelo, E.; Vidal, B.; Carotti, A. 1-, 3- and 8-substituted-9-deazaxanthines as potent and selective antagonists at the human A_{2B} adenosine receptor. *Bioorg. Med. Chem.* **2008**, *16*, 2852−2869.
(48) Stefanachi, A.; Nicolotti, O.; Leonetti, F.; Cellamare, S.; Campagna, F.; Loza, M. I.; Brea, J. M.; Mazza, F.; Gavuzzo, E.; Carotti, A. 1,3-Dialkyl-8-(hetero)aryl-9-OH-9-deazaxanthines as potent A_{2B} adenosine receptor antagonists: Design, synthesis, structure−affinity and structure−selectivity relationships. *Bioorg. Med. Chem.* **2008**, *16*, 9780−9789.
(49) *Dragon for Window (Software for Molecular Descriptors Calculations)*, version 5.4; Talete srl: Italy, 2006.
(50) *STATISTICA (data analysis software system)*, version 8.0; StatSoft Inc: Tulsa, USA, 2007.
(51) *MOPAC*, version 2007; Stewart Computational Chemistry: Colorado Springs, USA, 2007.
(52) *Molecular Operating Environment*, version 2007.09; Chemical Computing Group: Montreal, Canada, 2007.
(53) Consonni, V.; Todeschini, R.; Pavan, M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.
(54) Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693−705.
(55) Estrada, E. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844−849.
(56) Estrada, E. Spectral moments of the edge-adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320−328.
(57) Estrada, E. Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23−27.
(58) Witten, I.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.; Diane Cerra: San Francisco, 2005; p 525.
(59) Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507−2517.
(60) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, NY, 1997; p 432.
(61) Quinlan, J. R. *C4.5: Programs for Machine Learning;* Morgan Kaufmann: San Mateo, CA, 1993.
(62) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, NY, 1995.
(63) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition*; MIT Press: Cambridge, MA, 1986; Vol. *1*, pp 318−362.
(64) Huang, J.; Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299−310.
(65) Daskalaki, S.; Kopanas, I.; Avouris, N. Evaluation of classifiers for an uneven class distribution problem. *Appl. Artif. Intell.* **2006**, *20*, 381−417.
(66) Chawla, N. V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: New York, NY, 2010; pp 875−886.
(67) Hand, D.; Till, R. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **2001**, *45*, 171−186.
(68) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, NY, 1993.
(69) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.