Ensemble of Linear Models for Predicting Drug Properties*

Tomasz Arodź†‡ David A. Yuen§ Arkadiusz Z. Dudek¶

Institute of Computer Science, AGH Univ. of Science and Technology, 30-059 Kraków, Poland
Minnesota Supercomputing Institute, Univ. of Minnesota, Minneapolis, MN 55455, USA
Univ. of Minnesota Medical School, Dept. of Medicine, Minneapolis, MN 55455, USA

Abstract We propose a new classification method for prediction of drug properties, called the Random Feature Subset Boosting for Linear Discriminant Analysis (LDA). The main novelty of this method is the ability to overcome the problems with constructing ensembles of linear discriminant models based on generalised eigenvectors of covariance matrices. Such linear models are popular in building classification-based structure–activity relationships. Introduction of ensembles of LDA models allows for analysis of more complex problems than by using single LDA, e.g. those involving multiple mechanisms of action. Using four datasets, we show experimentally that the method is competitive with other recently studied chemoinformatic methods, including support vector machines and models based on decision trees. We present an easy scheme for interpreting the model despite its apparent sophistication. We also outline theoretical evidence as to why, contrary to the conventional AdaBoost ensemble algorithm, this method is able to increase the accuracy of LDA models.

1 Introduction

Computational methods in drug design have been studied extensively in past decades1,2. Major efforts are focused on the activity of therapeutic agents, as well as on other essential properties of pharmaceuticals. Such studies, aiming at predicting in a quantitative or qualitative manner the activity or property of a compound from its structure are collectively known as structure-activity relationship (SAR) studies3. In particular, much attention in SAR research was directed towards predicting the pharmacokinetic properties of compounds, including absorption, distribution, metabolism and excretion, as well as toxicity, collectively referred to as ADME/Tox4,5. Adverse drug effects are also studied, e.g. drug resistance6.

The SAR problems pose some specific requirements for the computational methods. The biochemical mechanisms of the studied effect are often unknown or diverse. Such multiple mechanisms of action can result in highly non-linear relationship between descriptors of the compound structure and its activity. The descriptors themselves pose another challenge. The large number of available types of descriptors makes it hard to manually choose only those meaningful in context of a given activity. On the other hand, using too many descriptors results in chance correlations and in datasets with large number of inter-descriptor correlations. Next, the available information on activity of compounds is usually small and may be highly unbalanced, with much fewer number of active than non-active compounds. Moreover, as new compounds are created by combinatorial chemistry7 on the basis of previous analyses, they may not follow the same distribution as the compound datasets previously available. Finally, the models resulting from SAR analysis should be
at the same time accurate and easily interpretable. The SAR model should allow for reliable prediction of activity of compounds during screening. However, it should also give insight into the properties of compounds that exhibit high activities, allowing for guiding the further exploration of chemical space. These two goals often contradict each other.

Various computational methods have been applied to the SAR analysis. A number of simple statistical methods are in widespread use, including linear models such as multiple linear regression\(^8\), linear discriminant analysis\(^9,10\) or partial least squares\(^11\). These models offer good interpretability, yet their accuracy is not always optimal, in particular for larger, more diverse datasets. More complex models are also in use, including artificial neural networks\(^10\), support vector machines\(^12\) or decision trees\(^13\). Recently, a group of ensemble methods gained attention from the chemoinformatic community. These models are multiple classifiers usually composed of decision trees, as in e.g. random forests\(^13,14\) and stochastic gradient boosting of decision trees\(^15\) methods. However, other ensemble types have also been tried in SAR, including SVM ensemble\(^16\) and artificial neural network with boosting\(^17\) and bagging\(^18\).

Boosting methods, such as AdaBoost\(^19\), have been one of the most successful ensemble methods up to date, both empirically and in terms of theoretical foundations. However, they have only recently been introduced\(^15,17\) to SAR studies and have been used with relatively complex base models, namely decision trees\(^15\) or artificial neural networks\(^17\). We propose here to use a simple linear LDA base model in boosting framework. Previous analyses argued that boosting of LDA does not lead to higher accuracy than single LDA model\(^20,21\). Here, we introduce a new approach, the Random Feature Subset Boosting, which is able to overcome the problems with boosting of LDA.

The new method is evaluated for four datasets related to drug absorption, resistance and toxicity. These are predicting of human intestinal absorption of drugs, and prognosis whether compound is a P-glycoprotein substrate or not, and thus is susceptible to efflux from the cell. We have also studied the prediction if a compound induces Torsade de Points cardiac arrhythmia, an adverse side effect of some drugs, and predicting the multidrug resistance reversal activity of chemical agents.

The rest of the paper is arranged as follows. In section 2 the Random Feature Subset Boosting of LDA is introduced. Moreover, the datasets used in this study are described, along with the computation procedure of evaluation of the method. In section 3 we present and analyse the results and compare them with other SAR models used in literature. Finally, section 4 gives the conclusions.

## 2 Methods

The ensemble and boosting schemes in machine learning have been extensively described in literature, e.g. by Freund and Schapire\(^22\) and Meir and Rätsch\(^23\). Thus, we give only a brief description of boosting, focusing on the novel aspects of our method, its empirical and theoretical properties and on its interpretability. We discuss also the four SAR datasets analysed.

### 2.1 Random Feature Subset Boosting for LDA

Similar to other ensemble methods, boosting\(^19\) involves combining multiple models to obtain a more reliable model. Contrary to other ensembles, such as the random subspace method\(^24\) or bagging\(^25\), the base models may be erroneous to a large extent. The high accuracy of the ensemble model is achieved by constructing new ensemble members to correct the errors of previously trained ensemble members. During the course of the algorithm, the weights of the examples from the training set are altered, shifting the attention of the next base classifiers to the previously misclassified examples. Each consecutive base model is trained to minimize the weighted error on the training set. Thus, a base classifier is forced to classify correctly the examples with large weights, even at the cost of making mistakes for examples with small weights.

However, for such a scheme to be successful, the base classifier must exhibit some instability with respect to changes in the weights. If, in spite of these
changes, the base models are similar, the increase in accuracy is low in comparison to a single base model. The diversity of base models forming the ensemble is an important factor influencing the accuracy of the ensemble.

The linear discriminant analysis (LDA) used widely in SAR studies has been tested in boosting and other ensemble types. However, the results were discouraging. The ensemble of LDA models often gives results similar or worse than a single LDA model. This is because models trained in consecutive rounds of boosting are very similar. Therefore, in order to successfully apply the LDA as a base classifier, a method forcing the base models to be more diverse has to be used. Here, diversity is achieved by introducing the concept of random feature subsets to boosting. Specifically, in each turn, the new base classifier is trained using a different, randomly chosen subset of descriptors.

The algorithm for Random Feature Subset Boosting (RFSBoost) for LDA method is outlined below as Algorithm 1. First, the initial weights $D_1$ of all the samples in the training set are initialised uniformly for each class (line 1). Next, the algorithm operates sequentially in $T$ rounds (line 2), in each training a new base model. In every round, a group $FeatSset$ of $S$ features is chosen randomly from all features $X$ (line 3). In the training of the LDA the training set, with only the selected features, is used (line 4), finding the linear decision boundary with minimal weighted training error. The weighted version of LDA method is used to take account of the weights $D_1$ (line 5).

Once the new linear decision is obtained by LDA, the weighted classification error is evaluated (line 6) and the weight of the trained LDA model in the whole ensemble is calculated (lines 9-10). The weights of the correctly classified examples are lowered, while those of misclassified examples remain unchanged. Finally, at the end of each round, the weights are normalized so that sum of weights for each of the classes is equal (line 13). This modification of the boosting scheme is introduced specifically for SAR analysis, where unbalanced class representations are not uncommon.

**Algorithm 1.** The RFSBoost for LDA

**INPUT:**

- Training set $Tr$ of size $m$:
  $Tr = \{(\vec{x}_1, c_1), \ldots, (\vec{x}_m, c_m)\} \subset X \times \{-, +\}$
- $T$ - maximal number of training rounds
- $S$ - number of features to be used

**RandomFeatureSubsetBoost** – LDA ($Tr, T, S$)

1. $\forall 1 \leq i \leq m$ $D_1(i) = \frac{1}{2} \sum_{j \varepsilon = c_i} D_1(j)$

2. for $t = 1$ to $T$

3. $FeatSset = \text{choose randomly } S$ distinct features from $X$

4. $TrFsSset = \text{narrow } Tr$ to selected features $FeatSset$

5. $h_t = \text{LDA} (TrFsSset, D_t)$

6. $\varepsilon_t = \sum_{i \varepsilon h_t(\vec{x}) \neq c_i} D_t(i)$

7. if $\varepsilon_t = 0$ or $\varepsilon_t \geq 0.5$

8. exit loop

9. $\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$

10. $\alpha_t = \log \frac{1}{\beta_t}$

11. $\forall i : h_t(\vec{x}) = c_i, D_{t+1}(i) = \beta_t D_t(i)$

12. $\forall i : h_t(\vec{x}) \neq c_i, D_{t+1}(i) = D_t(i)$

13. $\forall 1 \leq i \leq m$ $D_{t}(i) = \frac{1}{2} \sum_{j : \varepsilon_t = c_i} D_t(j)$

14. return $h^{\text{fin}}(\vec{x}) = \frac{\sum_{i=1}^{T} \alpha_i h_t(\vec{x})}{\sum_{i=1}^{T} \alpha_t}$

Our previous analysis of the RFSBoost-LDA method in general classification setup has given insight into the reasons, which make the LDA work significantly better in the RFSBoost scheme than in conventional boosting. First, the diversity of the constructed RFSBoost ensemble, measured with the Q statistics or with the variance of the decision boundary coefficients, tends to be higher than that of the AdaBoost. High diversity has been argued to influence the accuracy of the ensemble classifiers.

Another justification stems from the margin-based analysis of boosting. The margin of the ensemble on an example describes the distance from decision boundary to that example. The sign of the margin shows whether the example has been classified correctly, while the magnitude of the margin represents the confidence in the decision for that example. One of the reasons of good performance of AdaBoost in general is the maximisation of the margins. This leads to increase in the confidence of the
classification and reduction of the number of errors. Compared to AdaBoost, the minimal margin of the RFSBoost ensemble of LDA models is in most cases larger, indicating better classification capabilities.

Finally, experimental results gain support from the statistical learning theory. The bound on the generalization error of boosting is dependent on the VC-dimension of the base classifier. For a linear classifier in the feature space defined by $f$ descriptors, this dimension is $\vartheta_{LDA} = f + 1$. In the RFSBoost-LDA, the number of features used in LDA training is significantly reduced comparing to the AdaBoost-LDA. Thus, the VC-dimension of each weak classifier is reduced and the bound on the generalization error of the whole ensemble is lower, leading to better performance during the classification of previously unseen samples.

The RFSBoost exhibits other interesting properties not directly related to its accuracy. First, the linear methods such as LDA are not directly applicable in cases where the within covariance matrix is ill-conditioned. This includes e.g. datasets with large descriptor to examples ratio. By using only a small fraction of the available descriptors in each round, RFSBoost allows the creation of ensemble of linear models in such cases. Another benefit of using subsets of descriptors is the significant reduction in the computational complexity of a linear base model, which compensates for the multiplicity of base models in ensemble. Moreover, relying on subsets of features makes the method independent to a large extent from the number of descriptors provided by the investigator.

### 2.2 Interpreting the Ensemble of Linear Models

While achieving high accuracy is an essential goal of a SAR model, its interpretability is no less important. In particular, linear models offer good insight into the relation between compound descriptors and the analysed drug activity or property. In more sophisticated models, based on machine learning, obtaining information on the importance of the descriptors in not as straightforward. In our method, each of the linear base models $h_t$ contributing to the ensemble can be analysed individually in a convenient way, by inspecting the coefficients $h^f_t$ corresponding to the descriptor $f$. Since in each round only a group of descriptors is chosen randomly to participate in creating the base models, only the coefficients for those descriptors may be nonzero. In addition to the coefficients, the weight $\alpha_t$ of the $t$-th base model is available for inspection.

While the above method allows for insight into small ensemble models, it is nonetheless prohibitively laborious for typical ensembles consisting of 100 or more base models. Therefore, we propose a scheme for holistic analysis of the ensemble. The method takes into account the random process of choosing the subset of features in each round $t$. It also uses the information on the features’ coefficients of each linear base model.

Let $p_t(f) : \mathbb{N} \rightarrow \{0, 1\}$, be a function taking 1 if a descriptor $f$ is selected by the random process in the base model $h_t$ and 0 otherwise. Then, $c(f) = \sum_{t=1}^{T} p_t(f)$ defines the number of rounds from the total $T$ rounds, in which the descriptor $f$ is used to build the base model. The importance of the descriptor $f$ in a base model $h_t$ is given by magnitude of its coefficient, $\|h^f_t\|$. Therefore, the estimate of the importance of a descriptor $f$ in an ensemble of $T$ base models is:

$$I(f) = \frac{1}{c(f)} \sum_{t=1}^{T} \alpha_t \|h^f_t\|. \quad (1)$$

The importance factor $I$ takes into account the importance of descriptors within base models and of the base models within the ensemble. By utilizing $c(f)$ factors, it also compensates for the random nature of the choice of descriptors in creating each base model.

### 2.3 Datasets

We have evaluated the RFSBoost for LDA ensemble model using four prediction problems encountered in drug design. The number of compounds in each dataset and their distribution into classes of compounds are summarized in Table 1.

**HIA** The Human Intestinal Absorption prediction focuses on absorption of orally administered drugs.
Table 1: Datasets used in tests. Sizes of test and training sets during cross-validation are presented, along with distribution of compounds between two classes to be predicted. See text for details on datasets and meaning of “+” and “−” classes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Sets</th>
<th>Training Sets</th>
<th>Class</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIA</td>
<td>157</td>
<td>39</td>
<td>131</td>
<td>65</td>
</tr>
<tr>
<td>P-gp</td>
<td>161</td>
<td>40</td>
<td>116</td>
<td>85</td>
</tr>
<tr>
<td>TdP</td>
<td>289</td>
<td>72</td>
<td>85</td>
<td>276</td>
</tr>
<tr>
<td>MDRR</td>
<td>396</td>
<td>132</td>
<td>298</td>
<td>230</td>
</tr>
</tbody>
</table>

into blood. High absorption rate is essential for bioavailability of the drug at its target site. For the purpose of prediction, the compounds with absorption rate above 70% were classified as absorbable (HIA +) and the rest as non-absorbable (HIA −). The dataset has been obtained from Xue et al. and contains 131 absorbable and 65 non-absorbable compounds.

P-gp The second dataset involves prediction of P-glycoprotein (P-pg) substrates from non-substrates. The P-gp is a membrane transporter capable of transporting out of the cell a wide and diverse range of chemical compounds, including many therapeutic agents. Thus, it is important in drug design to evaluate on an early stage whether a drug is not a P-gp substrate susceptible to P-gp mediated drug efflux from the cell, limiting its activity. The dataset under investigation has been collected from various literature sources by Xue et al. and includes 116 P-gp substrates (P-gp +) and 85 non-substrates (P-gp −).

TdP Torsade de Pointes is a potentially fatal polymorphic ventricular tachycardia. It is linked to genetic causes leading to long QT interval on the electrocardiogram, and has been studied both physiologically and computationally. It may also be induced as an adverse effect of drugs that cause QT prolongation. This effect is present in different categories of therapeutic agents, e.g. antihistamines, antidepressants or macrolide antibiotics. The TdP dataset has been collected by Xue et al. It includes 85 TdP inducing agents (TdP +) and 276 non-inducing compounds (TdP −).

MDRR The Multidrug Resistance Reversal activity prediction is linked to the P-gp transporter described above. In response to a single cytotoxic drug, the MDR1 gene encoding P-gp may become upregulated and thus lead to resistance to a range of structurally and functionally unrelated drugs by means of their efflux from the cell. This leads to major problem e.g. in cancer chemotherapy. One of the promising approaches to tackle this problem is the use of Multidrug Resistance Reversal agents. The set of 528 compounds originally studied by Kloppman et al. and Bakken and Jurs has been obtained from Svetnik et al. It includes 298 active (MDRR +) compounds, with ratio of reversing the leukaemia’s cells resistance to adriamycin above 4.2, and 230 not active (MDRR −) compounds, with this ratio below 2.0.

Molecular Descriptors For HIA, Pgp and TdP datasets a group of 159 molecular descriptors is used after Xue et al. It includes 18 simple molecular properties, 28 molecular connectivity and shape descriptors, 84 electrotopological state descriptors, 18 quantum chemical properties and 16 geometrical properties. In case of MDRR dataset, 342 DRAGON descriptors, previously used by Svetnik et al., are employed.

2.4 Computation Procedure

To evaluate the accuracy of the RFSBoost-LDA method, we have averaged the results over 10 runs of algorithm on different random partitions of the datasets into training and test sets. The exact number of compounds in the training and test sets are specified in Table 1. Based on the results on the test sets, we have calculated four values quantifying the model’s reliability. These are:

- accuracy $Acc$ – percentage of correctly predicted samples to total number of samples in the test set;
• sensitivity \( S_e \) – the percentage of correctly predicted samples from class "+" to total number of samples from class "+" in the test set;

• specificity \( S_p \) – the percentage of correctly predicted samples from class "-" to total number of samples from class "-" in the test set;

• balanced accuracy \( Acc_b \), as used in SAR study by Weston et al.\(^{41} \) – the average of \( S_e \) and \( S_p \).

During the evaluation, two sizes of the ensembles were used, consisting of \( T=100 \) or 200 base models. For a given \( T \), the RFSBoost behaviour is controlled by a free parameter \( S \), specifying the number of features to be randomly chosen in each round. For all four datasets, we have tested configurations with \( S=2, 4, 6, 8 \) and 10 descriptors. To select the optimal value of \( T \) for a given dataset, we have used \( T = 100 \) if the models for this \( T \) achieved training accuracies close to 100\%, or \( T = 200 \) otherwise. Thus, we give preference to the less complex ensembles, provided they are sufficiently accurate. From all configurations with different number of features, \( S \), we have selected the one yielding the highest accuracy on the training set as the final result of the tests. In the case of several configurations achieving equal training accuracy, the one yielding larger margin during training was chosen.

The Matlab 7.0 R14 was used to solve the generalized eigenproblem of the LDA method. The first generalized eigenvector of within- and between-class covariance matrices leads to the optimal linear transformation that separates the classes while reducing the within-class variances. The rest of the Random Feature Subset Boosting algorithm was developed in-house using C++ and Matlab 7.0 R14.

### Table 2: Results of the tests of Random Feature Subset Boosting of LDA. Ensembles of 100 base models for HIA, P-gp, TdP and 200 for MDRR. See text for details on datasets. \( Acc \) - accuracy, \( S_e \) - sensitivity, \( S_p \) - specificity, \( Acc_b \) - balanced accuracy. All results in percents

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Acc.</th>
<th>( S_e )</th>
<th>( S_p )</th>
<th>( Acc_b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIA</td>
<td>80.8</td>
<td>86.9</td>
<td>68.5</td>
<td>77.7</td>
</tr>
<tr>
<td>P-gp</td>
<td>76.3</td>
<td>83.0</td>
<td>67.1</td>
<td>75.1</td>
</tr>
<tr>
<td>TdP</td>
<td>74.2</td>
<td>72.9</td>
<td>74.5</td>
<td>73.7</td>
</tr>
<tr>
<td>MDRR</td>
<td>81.8</td>
<td>89.2</td>
<td>72.1</td>
<td>80.7</td>
</tr>
</tbody>
</table>

For HIA and P-gp, the value of \( S \) optimal on the training set was 10 descriptors, for TdP it was 4 and for MDRR it was 8 descriptors. Models with those values of \( S \) were used in further analysis and in comparisons. The ensembles for HIA, P-gp and TdP consisted of 100 base models. The ensemble for MDRR, which is a dataset with more compounds and features, was composed of 200 base models.

![Receiver Operator Characteristic Curve](image1)

**Figure 1:** Receiver Operator Characteristics Curve for all datasets, representing the trade-off between true positives (sensitivity) and false positives (equal to 1 – specificity) on the test sets.

To illustrate the trade-off between sensitivity and
specificity on the test set, we have used the Receiver Operator Characteristics Curve depicted in Fig. 1, recently used in chemoinformatic analysis by Müller et al.\textsuperscript{42}. The diagram is constructed using the model yielding median accuracy from 10 models trained during evaluation for the chosen $S$.

We also show the results of the prediction using Pharmacological Distribution Diagrams introduced by Galvez et al.\textsuperscript{43}. The diagrams for HIA and MDRR are presented in Fig. 2 and 3, respectively. The diagrams were prepared using the combined results of the evaluation on the test sets.

![Figure 2: Pharmacological Distribution Diagram for HIA dataset. Active and inactive expectancy as a function of the margin of compounds from the test set.](image1)

The ROC curve and the PDD diagrams show the possibility of manipulating the decision threshold in RFSBoost to increase the accuracy within one class at the expense of the other class. Such an analysis is especially useful in the context of uneven costs of misclassification or unbalanced nature of the dataset, both situations encountered in SAR analyses.

3.1 Effects of Training Set Size

Having small number of compounds with available assessed activity is a common situation in QSAR studies. In particular, datasets comprising analogues of a single hit are often composed of 100 or less compounds. While complex, non-linear methods are certainly useful in large libraries, for small, more homogeneous datasets linear models may suffice. To prove usefulness of the non-linear RFSBoost-LDA also in the latter case, we have evaluated its accuracy during the reduction of training set size. Based on the HIA dataset, we have constructed a series of prediction problems with reduced number of training compounds, ranging from 70\% to 30\% of the whole dataset. For each inspected training set size, we have generated 10 random subsets of each of the replicated training sets used in evaluation described above.

![Figure 3: Pharmacological Distribution Diagram for MDRR dataset. Coordinates as in Fig. 2](image2)

We have preserved the same test sets and used the optimal parameters of the classifiers as specified above, with the exception of the number of rounds, $T$. To take account of the smaller datasets, we have executed lower number of rounds in boosting, ranging from $T=80$ to 10. The results are depicted in Fig. 4. The accuracy of RFSBoost-LDA is compared with linear LDA model, operating in principal component space to prevent singularity due to the number of descriptors.

The results show that, while the accuracy of RFSBoost drops when training information is being limited, similar accuracy decline can be observed for linear method. The edge in prediction accuracy of the proposed method over simpler, linear models is thus not limited to larger datasets.
Figure 4: Comparison of the test set accuracies of RFSBoost and linear LDA models built using training sets of size reduced successively from 80% to 70%, 60%, 50%, 40% and 30% of the whole HIA dataset. The number of base models in ensemble followed the reduction in training set size, with $T=80, 60, 40, 20$ and 10, respectively.

3.2 Interpreting the Ensemble Model

To show the possibilities of interpreting the ensemble model, we have used the results obtained for the HIA dataset. We have evaluated the importance $I(f)$ of each descriptor $f$ as described in eq. (1). The training of the ensemble contains random elements. Thus, we have run the RFSBoost-LDA algorithm 10 times on the same training set. From each of the 10 trained models, we have selected 20 descriptors with highest importance factor $I$.

Despite the reliance of training on randomised procedure, within the 10 pools of 20 highest ranking descriptors, 10 descriptors were present in results of all 10 executions of training, 4 features in 9 runs and further 2 descriptors in 8 pools. Some other descriptors, e.g. number of H-bond donors, were present in less than 8 pools. This shows that importance factor of the descriptors, in the form defined in eq. (1), is independent to a large extent from the stochastic nature of training. The most important descriptors, along with information on the number of pools they were present in, is presented in Table 3.

<table>
<thead>
<tr>
<th>$N$</th>
<th>Descriptor</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>VS van der Waals surface area</td>
</tr>
<tr>
<td>10</td>
<td>$\mu$ Molecular dipole moment</td>
</tr>
<tr>
<td>10</td>
<td>nhyd count of hydrogen atoms</td>
</tr>
<tr>
<td>10</td>
<td>nhev count of heavy atoms</td>
</tr>
<tr>
<td>10</td>
<td>$^0\chi$ Simple molecular connectivity for path order 0</td>
</tr>
<tr>
<td>10</td>
<td>$^2\chi$ Simple molecular connectivity for path order 2</td>
</tr>
<tr>
<td>10</td>
<td>$^0\chi^v$ Valence molecular connectivity for path order 0</td>
</tr>
<tr>
<td>10</td>
<td>$^1\kappa$ Molecular shape $\kappa$ index for one boned fragments</td>
</tr>
<tr>
<td>10</td>
<td>$^1\kappa_\alpha$ $\kappa_\alpha$ index for one boned fragments</td>
</tr>
<tr>
<td>10</td>
<td>$^2\kappa_\alpha$ $\kappa_\alpha$ index for two boned fragments</td>
</tr>
<tr>
<td>9</td>
<td>$\varepsilon_a$ Hydrogen bond donor acidity</td>
</tr>
<tr>
<td>9</td>
<td>$A$ electron affinity</td>
</tr>
<tr>
<td>9</td>
<td>$^1\chi$ Simple molecular connectivity for path order 1</td>
</tr>
<tr>
<td>9</td>
<td>S(22) Atom-type H Estate sum for $&gt;$CH$_2$</td>
</tr>
<tr>
<td>8</td>
<td>$^1\chi^v$ Valence molecular connectivity for path order 1</td>
</tr>
<tr>
<td>8</td>
<td>S(38) Atom-type H Estate sum for :N:-</td>
</tr>
</tbody>
</table>

3.3 Comparison with Other Methods

The results obtained for the RFSBoost LDA method were compared with other methods. To facilitate such comparison, in our study we have used the same descriptors as in the referenced studies. The compounds used and the number of compounds in the training and test sets are also preserved. The results obtained by Xue et al.\textsuperscript{33} for HIA, P-gp and Tdp using Support Vector Machines are presented in Table 4, along with ones obtained by Svetnik et al.\textsuperscript{15} for MDRR using Decision Trees, Support Vector Machines, Partial Least Squares, Stochastic Gradient Boosting of Decision Trees and Random Forest.
Table 4: Results of other studies involving same compounds, descriptors and similar computational procedure as in our test. Results after Xue et al.\textsuperscript{33} for HIA, P-gp and TdP and after Svetnik et al.\textsuperscript{15} for MDRR. \textit{Acc} - accuracy, \textit{Se} - sensitivity, \textit{Sp} - specificity, \textit{Acc\textsubscript{b}} - balanced accuracy. All results in percents

<table>
<thead>
<tr>
<th>Model</th>
<th>HIA</th>
<th>Pgp</th>
<th>TdP</th>
<th>MDRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>77.0</td>
<td>68.3</td>
<td>82.0</td>
<td>81.5</td>
</tr>
<tr>
<td>SVM</td>
<td>83.4</td>
<td>68.9</td>
<td>54.5</td>
<td>81.7</td>
</tr>
<tr>
<td>SVM</td>
<td>63.2</td>
<td>68.2</td>
<td>90.6</td>
<td>82.6</td>
</tr>
<tr>
<td>SVM</td>
<td>73.3</td>
<td>68.4</td>
<td>72.6</td>
<td>83.1</td>
</tr>
<tr>
<td>SVM</td>
<td>73.3</td>
<td>68.4</td>
<td>72.6</td>
<td>83.1</td>
</tr>
</tbody>
</table>

For the HIA and P-gp datasets, our ensemble model has achieved accuracy higher by 3.8% and 8%, respectively, than the SVM operating on all descriptors. It yielded also balanced accuracy higher by 4.4% and 6.5%, respectively. While in the case of TdP the SVM yielded accuracy higher by 7.8% than our model, it shows a significant lack of balance between sensitivity and specificity. When balanced accuracies \textit{Acc\textsubscript{b}} are compared, our ensemble model fares better by 1.1%. Moreover, the methodology of Xue et al. involved choosing the best model during parameter tuning using the error on the prediction set, while we have chosen the more conservative scenario of using the training error.

The comparison of results for MDRR is restricted to accuracy, as Svetnik et al.\textsuperscript{15} did not quote the sensitivity and specificity. The RFSBoost LDA is comparable in performance to other methods, with differences in accuracy ranging from 3.9% in favour of RFSBoost in case of Decision Tree to 1.3% in favour of Random Forests.

These results show that the model we have proposed is among the best of classification models when evaluated on the same dataset and using all descriptors. In case of HIA, P-gp and TdP, Xue et al.\textsuperscript{33} tested also a method consisting of Recursive Feature Elimination (RFE) and SVM to reduce the number of descriptors. This method achieved accuracy of 86.7% for HIA, 79.4% for P-gp and 83.9% for TdP, which are higher than results of our models. One should note, that the RFSBoost for LDA method is capable of achieving results closer in accuracy to that of RFE-SVM, when significantly larger ensembles are used. To show this, we have evaluated our method for ensembles of \textit{T}=2000 base models and followed Xue et al.\textsuperscript{33} in showing the configuration yielding best results on the test set. We have achieved accuracy of 82.1% for HIA, 80.8% for P-gp and 79.2% for TdP.

The dedicated descriptor selection techniques, such as RFE, treat reduction of irrelevant features as one of their main objectives. On the other hand, random feature subsets technique serves a different purpose in our model. It is used for destabilizing the LDA base models. The design of a technique that allows for discarding irrelevant descriptors and, at the same time, destabilizes the LDA base models thus seems an interesting future research goal.

For HIA, P-gp and MDRR prediction other studies, not directly comparable with ours, have been performed. These studies use different sets of compounds and descriptors, and thus the accuracy values may only give some general orientation on their relation to our method. For HIA prediction, the accuracy of a probabilistic neural network\textsuperscript{44} was 80%. Much smaller training and testing sets were used, with lower threshold between absorbable and non-absorbable compounds. In case of P-gp substrate prediction, a multi-pharmacophore model\textsuperscript{45} on a slightly smaller dataset achieved accuracy of 63%. The same set of compounds, with minor modifications, has been studied also by Svetnik et al.\textsuperscript{13,15}. Using a different set of descriptors, namely the binarized atom-pair descriptors, accuracy reaching 75.5% for Stochastic Gradient Boosting and 80.4% for Decision Forests has been achieved. Finally, the MDRR problem was studied by Bakken and Jurs\textsuperscript{40} on the same set of compounds and interclass separation thresholds, with a
similar but not identical set of descriptors. Their method, utilizing genetic algorithm for descriptor selection and LDA method, achieved 83.1% accuracy. In general, all these studies show lower or similar accuracy to ours.

4 Conclusions

In this work we have proposed the novel Random Feature Subset Boosting of LDA. The method allows for overcoming the problem of low accuracy in ensembles composed of linear discriminant analysis as base models. This result is achieved by destabilising in each turn the LDA base model by using only a small, randomly chosen subset of features. Thus, our approach unites two concepts, the random subsampling and boosting, which on their own fail to create accurate LDA ensembles. The key to success in our method lies in the increase of diversity obtained by introducing random feature subsets, while keeping the strong drive for accuracy present in conventional boosting.

We have shown experimentally that our method is competitive in terms of accuracy with other classification methods recently proposed in SAR studies, such as Support Vector Machines and Random Forests. The method is computationally efficient, owing to the multifold reduction in the number of features used in creating each linear method. Moreover, accurate results are obtained using ensembles of moderate size, composed of 100 to 200 base models, whereas e.g. successful applications of Random Forests employed 500 base trees. In another study, boosting of decision trees utilised optimisation of the number of base models up to 500 or 1000 trees, depending on the size of dataset.

Compared to linear methods, RFSBoost remains competitive even for small datasets. Tests with reduced training set size indicated lower prediction error than LDA, even when few compounds were available to the methods. The strength of linear models in such cases stems significantly from their evident interpretability. Interpreting complex, non-linear models is usually not as straightforward. Yet, for small datasets, where minute ensembles suffice, RFSBoost allows for insight into the model by analysing its linear constituents. For larger ensembles, we have shown a computationally efficient method for assessing the importance of each descriptor in the combined model.

Finally, we have noted that the use of dedicated descriptor elimination method allows the SVM to elevate accuracy beyond that of our model on some datasets. Thus, we have proposed a future research goal of introducing a descriptor selection method to be used in conjunction with random feature subsets.

Acknowledgements The authors would like to thank Prof. Dr. Witold Dzwinel for inspiring discussions and for insightful comments during the preparation of the manuscript. The research is supported by CMG of National Science Foundation.

Supporting Information Available: The Pharmacological Distribution Diagrams for P-gp and TdP datasets. This material is available free of charge via the Internet at http://pubs.acs.org

References


