Virtual Screening Using PLS Discriminant Analysis and ROC Curve Approach: An Application Study on PDE4 Inhibitors

Andrea Rizzi* and Alessandro Fioni

Chiesi Farmaceutici, Chemical Synthesis Department, via San Leonardo 96, 43100 Parma, Italy

Received February 28, 2008

Virtual screening (VS) represents an important tool for the drug discovery process, in particular for the hit generation phase. Classifiers are often inserted as filters at the beginning of a VS path, and in the present paper the performances of several PLS-DA classifiers (QikProp, Dragon, EVA descriptors) are evaluated in the effort to distinguish PDE4 inhibitors from other druglike molecules. As benchmark also docking scores and the fitness to pharmacophore hypotheses were used to perform the same task, checking in this way if docking or 3D search can be anticipated in the VS process. The visual analysis of the Receiver Operating Characteristic (ROC) curve was useful to have an overall picture of the classification and to select the right threshold that marks the boundary between active and inactive classes. The best classification was obtained by a model based on the Dragon descriptors that are calculated from the molecular 2D structure. Its performance was good for the training set in terms of recall, enrichment factor, and area under the ROC curve and was confirmed in the prediction of the test set.

INTRODUCTION

In the modern pharmaceutical research virtual screening (VS) plays a key role in the early stages of drug discovery. The aim of this computational process is to select reduced sets of potential hits from large compound collections.[1] Nowadays the opportunity to access to the structural information of commercially available compounds allows the creation of comprehensive libraries with more than 2 million small organic molecules.[2–4] Starting from SMILES representations or 2D SDF files, the generation of the related 3D structures, including the calculation of the biological relevant ionization states and multiple tautomeric forms, is feasible, even if it is a time-consuming task. Such critical barrier may be avoided using ZINC, a free database that contains over 4.6 million commercially available compounds in ready-to-dock 3D formats.[5] The process of selection in databases with 10^5-10^6 structures needs the application of sequential filters, such as druglikeness,[6] statistical models based on molecular descriptors,[7] similarity[8] or pharmacophore searches,[9] and protein–ligand affinity scores.[10] Usually, in a hierarchical screening cascade[11] the level of complexity and the computational requirements define the position of the filter in the workflow, because it is advisable to perform the most time-consuming task at the end of the process, when the number of survived candidates is limited.

Classification models or similarity searches are fast methods and for this reason are often inserted at the beginning of the VS path with the aim of identifying and excluding potentially undesired structures in large compound databases.[12] Several statistical methods are employed to build classifiers: K nearest neighbors algorithms,[13] artificial neural networks,[14] support vector machines,[15] PLS discriminant analysis (PLS-DA),[16] etc. Classification can be performed according to two strategies: unsupervised, when the classes are not known “a priori”, and supervised, when the statistical method can learn from the objects of a training set.[17] The application of PCA (Principal Component Analysis)[18] is an example of unsupervised method because the directions of the orthogonal axes are placed where the variance of the distances among the objects is the largest. However, these orientations do not often coincide with the best separation among the classes, and PLS-DA, a supervised method, tends to perform better because new directions of the principal components are found with the purpose to enhance the class separation (“discrimination”).

PLS-DA has already been applied as supervised method in medicinal chemistry to discriminate active and inactive compounds,[11,19,20] well and poorly absorbed drugs.[21,22] Y-data vectors of dummy variables are normally used in order to encode a class identity, and, for instance, if active molecules are defined with the value 1 and inactive compounds with 0, a theoretical threshold of 0.5 must be set to define the boundary between the two classes, but if the classes are constituted by a different number of molecules, it is necessary to shift the threshold toward the value of the most populated class.[19] The definition of the most appropriate cutoff limit is not trivial since a high value produces a limited number of hits (conservative strategy: few true and false actives), while vice versa a low threshold selects several hits (liberal strategy: many true and false actives). In addition, an absolute indicator of the classification performance does not exist, and it is recommendable to analyze various measures together.

A helpful support to solve this uncertainty may be provided by the visual inspection of the “receiver operating characteristic” (ROC) curve, which has been used in various fields (medicine, meteorology, etc.) to make better decisions[23] and recently has been applied also in the drug discovery field.[24] On a ROC graph the false positive rate (1 − specificity) on

* Corresponding author phone: +39 0521 279912; fax: +39 0521 279880; e-mail: a.rizzi@chiesigroup.com.

10.1021/ci800072r CCC: $40.75 © 2008 American Chemical Society
Published on Web 08/01/2008

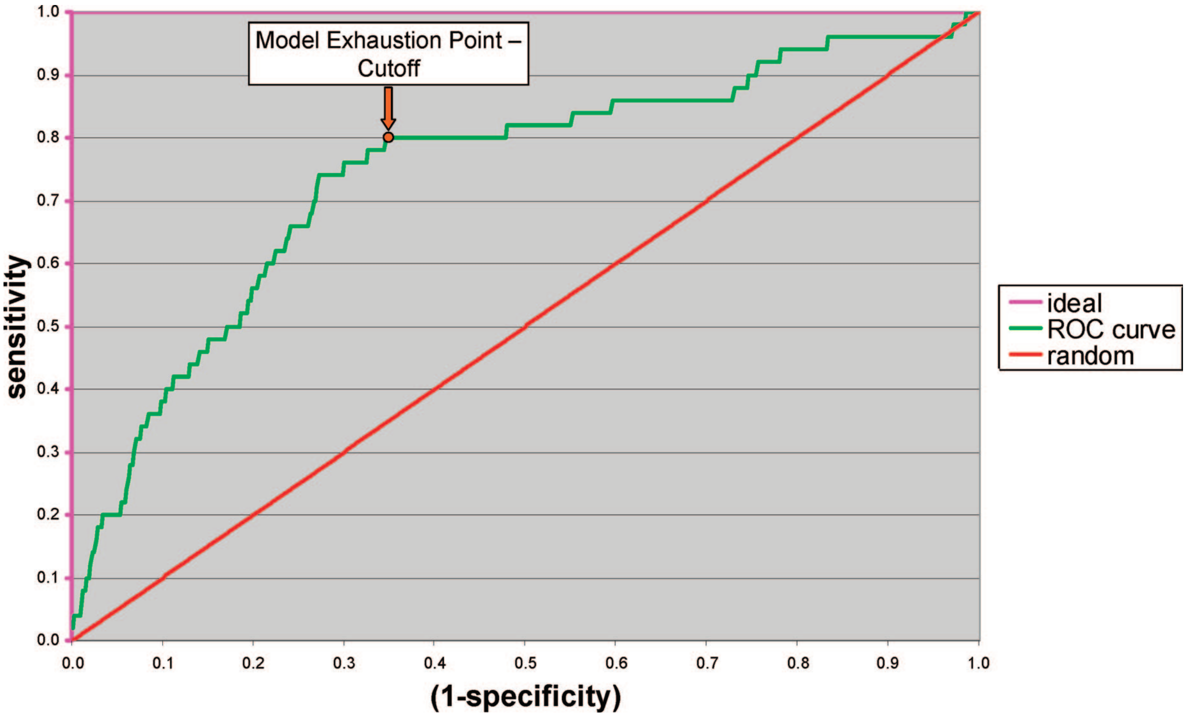
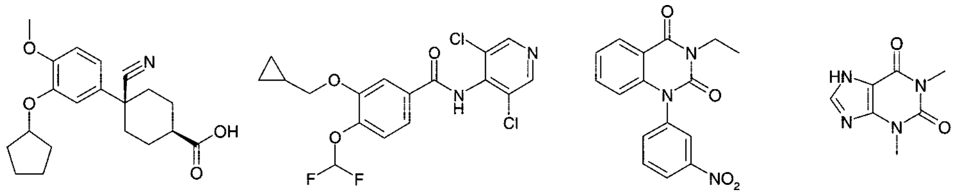


Figure 1. ROC curve: sensitivity is plotted vs (1 − specificity) for all cutoff values.



1. Cilomilast **2.** Roflumilast **3.** Nitraquazone **4.** Theophylline

Figure 2. PDE4 inhibitors: reference lead compounds.

the *X* axis and the true positive rate (sensitivity) on the *Y* axis are drawn for all possible threshold levels. The true positive and true negative rates are defined as follows:

True Positive Rate = Sensitivity = TP/(TP + FN)
False Positive Rate = 1 − Specificity = FP/(TN + FP)

where TP = true positives, FN = false negatives, FP = false positives, and TN = true negatives. The plot of a random classification would give a diagonal rising from the origin to the upper right corner (red line - Figure 1), while an ideal classification is represented by two segments, which join the origin, the upper left corner and then the upper right corner (magenta curve - Figure 1).

The area under the curve (AUC), which in several cases ranges between 0.5 (random classifier) and 1 (perfect classifier), is an objective measure of the overall method performance, independently of the selected threshold.

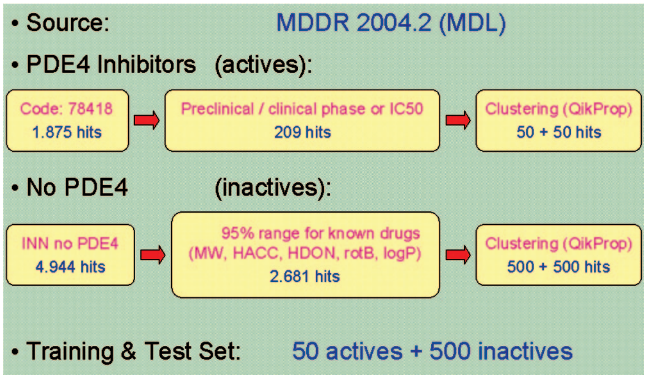
In this study, PLS-DA and ROC analysis were used to build and compare classifiers based on molecular descriptors and were applied in the identification of PDE4 inhibitors from other inactive druglike compounds. In the last years PDE4 inhibitors have been studied as drugs for the treatment of asthma and chronic obstructive pulmonary disease (COPD) for their combined anti-inflammatory and bronchodilatory profiles.[25] PDE4 inhibitors (Figure 2) are roughly classified into three main chemical families:[26] catechol ethers represented by cilomilast (**1**) and roflumilast (**2**), quinazolinediones related to nitraquazone (**3**), and xanthines to which theophylline (**4**) belongs.

Although none of the clinical candidates have reached the market for the difficulty to balance the therapeutic benefits and side effects, such as nausea, emesis, and headache, PDE4 inhibitors are still receiving a substantial interest by the scientific community. The goal of this paper is to analyze the performance of PLS-DA classifiers based on different sets of molecular descriptors (QikProp,[27] EVA,[28] Dragon[29]) and to compare their efficiency in retrieving PDE4 inhibitors to 3D pharmacophore searches and molecular docking, which represent two important tools for the virtual screening of compound libraries.

METHODS

Training and Test Sets. For comparing the performance and the robustness of the different discrimination models a training and a test set, both consisting of 50 active and 500 inactive compounds, were created, using the MDL Data Drug Report (MDDR)[30] as the source of the molecular structures. To generate active sets (Scheme 1) among the structures classified as PDE4 inhibitors (MDDR activity index: 78418), a preliminary group of 209 active molecules was extracted picking out both compounds in the preclinical-clinical phase and inhibitors characterized by $pIC_{50} > 6$.

Scheme 1. Selection Criteria Followed for the Composition of the Training and the Test Sets (50 Active and 500 Inactive Compounds)


- Source: MDDR 2004.2 (MDL)
- PDE4 Inhibitors (actives):

Code: 78418 1.875 hits → Preclinical / clinical phase or IC50 209 hits → Clustering (QikProp) 50 + 50 hits

- No PDE4 (inactives):

INN no PDE4 4.944 hits → 95% range for known drugs (MW, HACC, HDON, rotB, logP) 2.681 hits → Clustering (QikProp) 500 + 500 hits

- Training & Test Set: 50 actives + 500 inactives

Table 1. Classes of Descriptors Generated in Dragon[a]

constitutional descriptors (0D)
topological descriptors (2D)
walk and path counts (2D)
connectivity indices (2D)
information indices (2D)
2D autocorrelations (2D)
edge adjacency indices (2D)
Burden eigenvalues (2D)
topological charge indices (2D)
eigenvalue-based indices (2D)
Randic molecular profiles (3D)
geometrical descriptors (3D)
RDF descriptors (3D)
3D-MoRSE descriptors (3D)
WHIM descriptors (3D)
GETAWAY descriptors (3D)
functional group counts (1D)
atom-centered fragments (1D)
charge descriptors (others)
molecular properties (others)

[a] In brackets the type of molecular representation required for the calculation.

The final sets, consisting of 50 active compounds, were selected by structural diversity applying the Ward reciprocal nearest neighbor clustering method on the first two components of a PCA model based on 20 QikProp physicochemical descriptors. 4.944 compounds, characterized by an INN or USAN name and not registered as PDE4 inhibitors, were selected for the generation of inactive sets (Scheme 1) and to ensure druglikeness; this subset was reduced to 2.681 compounds applying the following criteria: MW (130/725), rotatable bonds (0/15), H-bond donors (0/6), H-bond acceptors (2/20), and logP (−2/6).[31] The last step in the selection process consisted of the same clustering procedure used for the PDE4 inhibitors.

PLS-Discriminant Analysis. PLS-DA is a multivariate technique that allows the generation of models able to discriminate active and inactive compounds for a considered target and in the present study was applied to different sets of molecular descriptors: QikProp, Dragon, and EVA. QikProp descriptors consist of physicochemical and pharmaceutically relevant properties,[32] which are calculated from the full 3D structure avoiding a fragment-based approach. Dragon calculates more than 1650 molecular descriptors, which are classified in 20 blocks (Table 1). Dragon descriptors were used to develop three different models: 1) all descriptors − 1661 variables; 2) with exclusion of 3D descriptors − 926 variables; 3) classes related to the molecular shape (topological descriptors, Randic molecular profiles, geometrical descriptors, WHIM descriptors, functional group counts, and walk and path counts) − 532 variables.

EVA descriptors are based on theoretically derived normal coordinate frequencies (IR spectrum), and in the current study the vibrational frequencies ($0 - 4000$ cm^{-1}) were calculated with Spartan[33] with the lines transformed in a spectrum applying a Gaussian smoothing function (fuzziness). Since the theoretical absorptions are sampled in a 5 cm^{-1} range, every EVA set consists of 800 spectral descriptors.

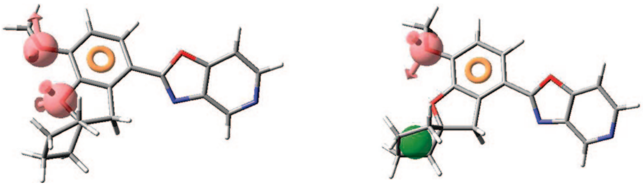


Figure 3. Pharmacophore hypotheses: AAR1 on the left and AHR11 on the right. The H-bond acceptors atoms are shown in red, while the hydrophobic centroids and the aromatic rings are in green and in orange, respectively. The AHR15 hypothesis has not been illustrated because it is rather similar to AHR11, except for a different orientation of the aromatic ring.

In the PLS-DA models the training of the system was performed assigning a class membership to PDE4 inhibitors ($Y = 1$) and to the inactive compounds ($Y = 0$). The number of significant components was determined in SIMCA[34] applying a cross-validation with seven groups, and the outliers were detected by visual inspection of the score plots looking at the elliptical confidence region based on Hotelling's T^2 (0.05). To obtain more robust models the number of variables was reduced following the indications of the variable importance in the projection (VIP) and the coefficient plot tools. Variables were sorted by order of importance, and after iterative exclusion of the less pertinent variables (VIP $<$ 1) the cross-validated R^2 (Q^2) was calculated for the new models. If the exclusion of a variable produced a decrease of the Q^2 in the new model, such a descriptor was reinserted; otherwise, it was eliminated definitively from the data set. The iterative process continued until Q^2 reached the maximum value, and any additional removal of variables did not improve the classification. After deletion of the unimportant variables the optimized PLS-DA models were applied to the data sets in prediction mode, and the calculated Y values were used as indicator of the class membership.

Docking Scores. The ability of structure-based methods to discriminate active and inactive compounds was assessed with flexible docking in the catalytic site of the PDE4B enzyme (PDB code: 1XM4). Glide docking scores,[35] calculated in SP (standard precision) or XP (extra precision) mode, were used in the first two models to rank and classify the compounds in the data sets. Knowledge about the ligand binding mode was included in the receptor grid generation and in the docking procedure, in order to obtain better enrichment factors. For example, the hydrogen bond between the ligand and the Gln 443 residue is known to be important for the activity,[36] and such interaction was imposed in Glide as constraint for the protein−ligand polar interactions. The issue regarding the solvation of the protein in the binding pocket was overcome retaining seven water molecules: six of which are directly coordinated to the catalytic metal ions (Zn^{2+}, Mg^{2+}), while the last one forms a hydrogen bond with the Piclamilast amide moiety in the 1XM4 complex. The most appropriate VdW radii of atoms were indicated by a preliminary experimental design: the best scaling factors were 0.9 for the ligand and 1.0 for the protein. The compounds, for which it was impossible to find a correct alignment in the binding pocket, were excluded in the score ranking. With the aim to improve the docking results, Glide SP scores were also treated in a third classification model applying a procedure called multiple active site corrections

Table 2. Performance Indicators of the PLS-DA Models Based on Physicochemical and ADME Descriptors (Training Set)

descriptors	Q^2	A[a]	K_{start}[b]	K_{end}[c]	accuracy	precision	recall	EF[d]	MCC[e]	AUC[f]
QikProp all	0.187	2	36	9	0.81	0.29	0.80	3.21	0.40	0.87
QikProp no ADME	0.104	2	16	6	0.83	0.30	0.68	3.34	0.37	0.85
EVA	0.179	2	800	72	0.85	0.36	0.74	3.91	0.44	0.89
Dragon all	0.476	3	1661	51	0.95	0.65	0.90	7.17	0.74	0.98
Dragon no 3D	0.470	3	926	101	0.96	0.68	0.98	7.49	0.80	0.99
Dragon shape	0.388	3	532	72	0.91	0.50	0.88	5.50	0.62	0.96

[a] Number of PLS components. [b] Number of starting variables. [c] Number of variables included in the final model. [d] Enrichment factor. [e] Matthews correlation coefficient. [f] Area under the ROC curve.

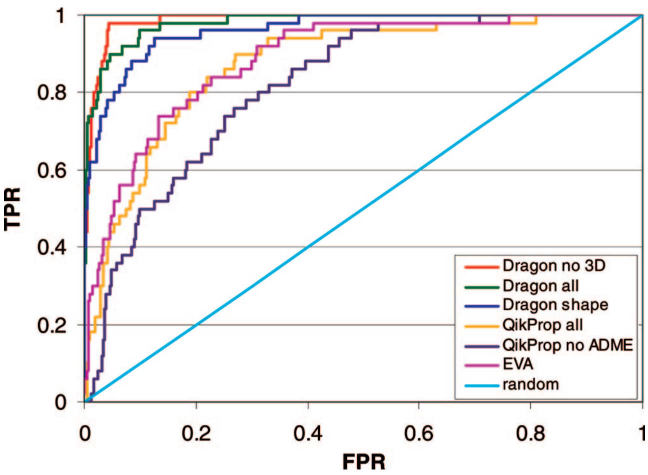


Figure 4. ROC curves of PLS-DA models based on physicochemical and ADME descriptors (training set).

(MASC).[37] The basic idea behind MASC is to measure the ligand specificity for a particular target site, lowering the aspecific contributions, which are evaluated with a multiple docking in at least seven diverse proteins. A fourth model was obtained building a PLS-DA classifier based on the following binding parameters that are included in the Glide report file: docking score (GScore), lipophilic contact (Lipo), hydrogen-bonding (HBond), metal-binding (Metal), penalty for buried polar groups (BuryP), van der Waals energy (vdW), Coulomb energy (Coul), penalty for freezing rotatable bonds (RotB), polar interactions in the active site (Site), internal torsional energy of the ligand conformer (Intern), the nonbonded interaction energy between the ligand and the receptor (CvdW = Coul + vdW), and Emodel (a specific combination of GScore, CvdW, Intern). PLS-DA was applied to the Glide parameters to examine whether multivariate analysis can improve the discriminatory power of molecular docking.[19]

3D Pharmacophore Models. In Phase[38] the generation of the pharmacophore hypotheses was performed without conformational sampling, taking into consideration only the poses of PDE4 inhibitors that were correctly aligned in the protein binding site by the docking procedure. Three hypotheses (AAR1, AHR11, AHR15) were considered effective because they recovered at least 70% of the PDE4 inhibitors included in the training set (Figure 3).

AAR1 consists of two H-bond acceptor atoms directly linked to an aromatic ring, while AHR hypotheses are formed by a combination of H-bond acceptor, hydrophobic group, and aromatic ring pharmacophore points. Both the training and the test sets were screened according to the three pharmacophore hypotheses, and the resulting descriptors generated by Phase (alignment score, vector score, volume score, overall fitness), which measure how well the best conformer of each molecular structure matches the pharmacophore features of the hypothesis, were used to build classification models in SIMCA. In the training and test sets all the compounds were ranked according to their Y predicted values calculated by the PLS-DA model, excluding the candidates not matching the pharmacophore hypothesis.

ROC Analysis and Performance Indicators. As mentioned in the Introduction, sensitivity and specificity are the main features allowing the design of the ROC curve. A preliminary indication about the quality of a model can be obtained by visual inspection of the plot, especially in the left region where relevant enrichments take place. The area under the ROC curve (AUC) is an important indicator of the model performance, independently from the selected cutoff value: the closer AUC is to 1, the better is the performance of the classification.

In our study we used the ROC curve also for the determination of the most appropriate cutoff value. In the plot the point from where the slope of the jagged curve starts to become substantially lower than the random model straight line, is defined the "model exhaustion point", which indicates that the model has just finished its ability to classify active compounds more efficiently than a random selection (Figure 1). Since any given point on the curve corresponds to a given threshold, the cutoff value associated with the "model exhaustion point" has been used in our study to define the boundary between active and inactive objects in the evaluation of the classification performance.

For the evaluation of the classification performance several indices (accuracy, precision, recall, enrichment factor, and Matthews correlation coefficient) were calculated for each model. Accuracy is the overall classification accuracy of the model, including both the active and inactive compounds. It is defined by the formula

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. In our case, since the ratio between actives and inactives is 1:10, the measure of accuracy is particularly influenced by inactive compounds. Precision is a measure of the capability of predicting active compounds. It is defined by

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

Recall is a measure of the ability to keep hits of a certain class (i.e., PDE4 inhibitors). It is defined by

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

These parameters should be seen together because none of these three numbers is an absolute indicator of classifica-

Table 3. Performance Indicators of the Models Based on Docking Scores (Training Set)

	no docked structures								
docking mode	PDE4	no PDE4	total	accuracy	precision	recall	EF[a]	MCC[b]	AUC[c]
SP	1	56	57	0.74	0.22	0.70	2.39	0.28	0.77
XP	11	251	262	0.74	0.22	0.74	2.44	0.30	0.75
MASC	0	31	31	0.84	0.31	0.64	3.45	0.37	0.78
PLS-DA	1	56	57	0.88	0.39	0.58	4.25	0.41	0.84

[a] Enrichment factor. [b] Matthews correlation coefficient. [c] Area under the ROC curve.

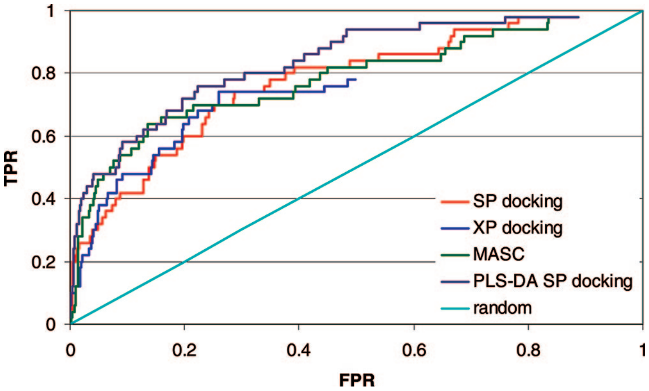


Figure 5. ROC curves of models based on docking scores (training set).

tion performance by itself. In the docking studies enrichment factor (EF) is the most used measure of virtual screening effectiveness, and, following Pearlman and Charifson,[39] the enrichment factor is defined as

$$EF = (TP/(TP + FP)) * ((TP + FP + TN + FN)/(TP + FN))$$

The formula can be broken into two factors: on the left the "precision" indicator can be identified, while on the right is written the ratio between the whole population of the set (in our work 550 compounds) and the active molecules (50 PDE4 inhibitors). The enrichment factor of a random classifier is 1, while for our data set the upper limit, which can be reached only by an ideal model, is represented by the value 11. A complementary measure of the prediction accuracy is the Matthews correlation coefficient,[40] which is used to reflect the correlation between the predicted and the observed result. Matthews correlation coefficient (MCC) is calculated with the formula

$$MCC = (TP * TN - FP * FN)/((TN + FP) * (TN + FN) * (TP + FP) * (TP + FN)) \wedge 0.5$$

The values range from −1 to 1, and a perfect prediction obtains a MCC of 1.

RESULTS

PLS-DA: Physicochemical and ADME Descriptors. QikProp, EVA, and Dragon descriptors were used to create classification models (Table 2).

QikProp descriptors led to a two component model with a good recall (0.80), but the misclassification of several inactive compounds was responsible for a low precision (0.29). When descriptors related specifically to the ADME profile (i.e., Caco-2 cell permeability, blood brain partition coefficient, and skin permeability) were eliminated from the whole QikProp set, the performance of the model got worse, particularly for its lower recall (0.68 vs 0.80).

EVA descriptors, which are indirectly related to the presence of chemical groups (i.e., amides), generated a two component model. QikProp and EVA classifiers approximately showed a similar profile in terms of performance indicators (Table 2).

A considerable improvement in the prediction capability was obtained using Dragon descriptors in all their variants: all descriptors, exclusion of 3D variables, and shape related classes. All Dragon models consisted of three PLS components, and the best performance was obtained by the model where 3D descriptors were not considered (Dragon no 3D - Figure 4).

The model based on descriptors related to the molecular shape showed a loss in efficiency (Dragon shape - Figure 4).

Docking Scores. Two different docking procedures were applied: standard precision (SP) and extra precision mode (XP).

XP docking was extremely selective: nearly half of the candidates (11 actives + 251 inactives − Table 3) did not receive favorable docking scores. This result is not surprising because XP mode is devised to identify ligand poses that would be expected to have unfavorable energies.

Multiple active site corrections (MASC) and PLS-DA based on a set of Glide energy parameters were separately evaluated to check if score postprocessing may improve the classification ability. After having ranked the compounds according to their Glide, MASC, or Y predicted (PLS-DA model) scores, the analysis of the performance indicators (Table 3) pointed out that standard precision (SP) and extra precision (XP) docking show the worst performance profile. MASC correction is beneficial in terms of precision and accuracy, but the recall of the classifier decreases. Both XP mode and MASC procedure have the serious drawback of being time-expensive methods. The PLS-DA model, which makes use of an appropriate and optimized combination of different energy contributions, performs better than other docking-based methods. A similar improvement in discrimination has been reported by Jacobsson in a study where classifiers showed a superior performance compared to consensus scoring and single scoring functions.[19] In SIMCA VIP analysis and coefficient plots highlight the importance of Glide score, lipophilic contact term, and Emodel in exerting a heavy influence on the model.

The improved discrimination power of the PLS-DA model compared to the single docking scores is confirmed also by ROC curve profile (Figure 5), which is cutoff independent.

3D Pharmacophore Models. Training and test sets were screened against three pharmacophore hypotheses: AAR1, AHR11, and AHR15. After 3D search about 75−80% of the active compounds were retrieved to agree with each of the three hypotheses.

Table 4. Performance Indicators of the PLS-DA Models Based on Pharmacophore Fitness Descriptors (Training Set)

hypothesis	Q^2	A[a]	K_{end}[b]	N[c]	accuracy	precision	recall	EF[d]	MCC[e]	AUC[f]
AAR1	0.182	1	3	315	0.84	0.32	0.74	3.57	0.42	0.84
AHR11	0.296	1	2	320	0.86	0.37	0.80	4.04	0.48	0.88
AHR15	0.245	1	4	315	0.84	0.34	0.80	3.73	0.45	0.86

[a] Number of PLS components. [b] Number of final variables. [c] Number of matched compounds. [d] Enrichment factor. [e] Matthews correlation coefficient. [f] Area under the ROC curve.

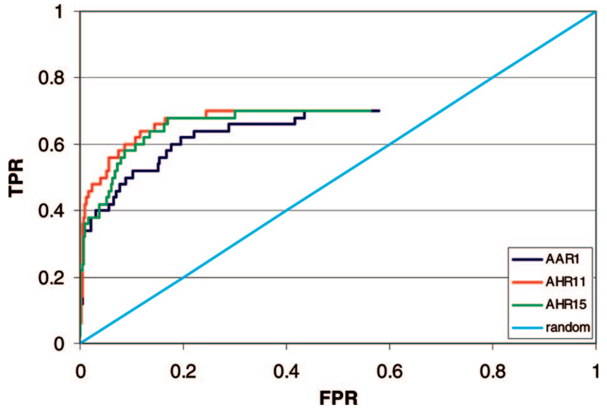


Figure 6. ROC curves of models based on pharmacophore fitness scores (training set).

In AAR1 the presence of two H-bond acceptor points allowed for the retrieval of all the catechol inhibitors but caused the detriment of finding other active compounds, while AHR hypotheses were less specific and also captured some xanthine and quinazolinedione derivatives (Figure 2). As reported in Table 4 and in Figure 6 the pharmacophore models gave similar results for all the performance indicators.

In general terms, the pattern of ROC curves indicates that 3D search based on pharmacophore gives excellent results if a reduced part (1−2%) of the set is sampled. If the pharmacophore search is enlarged at the level of 5% or more of the whole database, the classification performance drops. Nevertheless, if a substantial reduction is needed to reduce large commercial compound databases, highly selective pharmacophore models may be useful in the first steps of a virtual screening process.

DISCUSSION AND CONCLUSION

In this study a comparative analysis of different virtual screening methods has been performed for identifying PDE4 inhibitors from other druglike compounds. The comparison was facilitated by the application of the ROC curve approach, which is a method based on the plot representation of the sensitivity and the specificity of a given classifier at different threshold levels. ROC analysis offers the advantage of a direct comparison among different models in a simple graphical way and a useful feature of the ROC plot is the area under the curve (AUC), which measures the overall method performance, independently of the selected cutoff.

Furthermore, we have found that from the profile of the curve it is possible to establish an appropriate cutoff value of the classification model, which corresponds to the point where the slope of the curve starts to become lower than the straight line representing a random classifier. This particular point (model exhaustion point) defines an objective level, from which the classifier stops performing efficiently in the discrimination between active and inactive compounds (Figure 1).

Both ligand-based and structure-based methods have been applied in our virtual screening experiments, and, although our conclusion is based solely on a single case study, we have found that ligand-based approaches might be preferred to molecular docking, especially when at the beginning of a virtual screening cascade the starting molecular database needs to be reduced in size excluding many inactive compounds. Since the PDE4 catalytic site is characterized by a large binding cavity and shows a wide region exposed to the solvent,[36] it represents a difficult target for structure-based methods, and consequently high enrichments should not be expected. In addition, the accuracy of scores still represents a serious problem for molecular docking because the free energy of binding cannot be predicted accurately with the current scoring functions.[41] In spite of this drawback, molecular docking is important for the information that can be derived from the mapping of the protein binding site and thus remains an irreplaceable method in virtual screening,

Table 5. Performance Indicators of All Classifiers (Training and Test Sets)

	training set					test set				
classifiers	recall	precision	EF[a]	MCC[b]	AUC[c]	recall	precision	EF[a]	MCC[b]	AUC[c]
Dragon all	0.90	0.65	7.17	0.74	0.98	0.88	0.49	5.38	0.61	0.96
Dragon no 3D	0.98	0.68	7.49	0.80	0.99	0.90	0.54	5.89	0.66	0.95
Dragon shape	0.88	0.50	5.50	0.62	0.96	0.90	0.40	4.38	0.54	0.94
QikProp all	0.80	0.29	3.21	0.40	0.87	0.74	0.22	2.37	0.29	0.82
QikProp no ADME	0.68	0.30	3.34	0.37	0.85	0.62	0.23	2.58	0.28	0.82
EVA	0.74	0.36	3.91	0.44	0.89	0.70	0.36	4.01	0.44	0.83
3D-pharm AAR1	0.74	0.32	3.57	0.42	0.84	0.60	0.39	4.29	0.42	0.80
3D-pharm AHR11	0.80	0.34	3.73	0.45	0.86	0.64	0.42	4.57	0.46	0.78
3D-pharm AHR15	0.80	0.37	4.04	0.48	0.88	0.66	0.34	3.70	0.40	0.78
SP docking	0.70	0.22	2.39	0.28	0.77	0.62	0.20	2.17	0.23	0.77
XP docking	0.74	0.22	2.44	0.30	0.75	0.68	0.19	2.10	0.24	0.73
MASC	0.64	0.31	3.45	0.37	0.78	0.56	0.28	3.08	0.31	0.79
PLS-DA	0.58	0.39	4.25	0.41	0.84	0.62	0.34	3.75	0.39	0.84

[a] Enrichment factor. [b] Matthews correlation coefficient. [c] Area under the ROC curve.

especially if the selection should be performed within focused sets. The performance of 3D pharmacophore searches was not so remarkable because the structural diversity of the PDE4 inhibitors included in our training set did not allow the generation of pharmacophore hypotheses with more than 3 points. We assume that if a 4-point hypothesis had been found, the discrimination between active and inactive compounds would have been more consistent. Mason compared the performance of 3-point and 4-point pharmacophore fingerprints and found that 4-point pharmacophores are able to discriminate better for their enhanced selectivity.[9]

In our study the best classification was obtained with a PLS-DA model based on the Dragon descriptors that are derived from the 2D molecular structure. Although some Dragon descriptors are not easily interpretable, their capability to classify PDE4 inhibitors is considerable in terms of enrichment factor and recall. Furthermore, the model is robust because the difference in performance between the training and the test set is not large (Table 5), even if a risk of chance correlation exists in consequence of the high number of variables included in the classifier. The exclusion of 3D descriptors seems to indicate that from the connection table it is possible to extract sufficient information for the discrimination of PDE4 inhibitors.

Our conclusion is that the application of multivariate statistical analysis to a large set of molecular descriptors can generate efficient classifiers, which are useful in the first steps of a virtual screening process when fast computational methods are needed to reduce the set of candidates with a controlled loss of active compounds.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening: an overview. *Drug Discovery Today* **1998**, *3*, 160–178.

(2) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 643–651.

(3) Sirois, S.; Wei, D. Q.; Du, Q.; Chou, K. C. Virtual screening for SARS-CoV protease based on KZ70888 pharmacophore points. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1111–1122.

(4) Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model.* **2006**, *46* (2), 708–716.

(5) (a) Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182. (b) A free database for virtual screening: ZINC is not commercial: http://zinc.docking.org.

(6) Li, Q.; Bender, A.; Pei, J.; Lai, L. A. Large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification. *J. Chem. Inf. Model.* **2007**, *47* (5), 1776–1786.

(7) Plewczynski, D.; Spieser, S. A.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (3), 1098–1106.

(8) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1275–1281.

(9) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42* (17), 3251–3264.

(10) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46* (12), 2287–2303.

(11) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual screening of biogenic amine-binding G-protein coupled receptors: comparative evaluation of protein- and ligand-based virtual screening protocols. *J. Med. Chem.* **2005**, *48* (17), 5448–5465.

(12) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64–70.

(13) Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.

(14) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, 1999.

(15) Vapnik, V. N. *The nature of statistical learning theory*; Springer Verlag: New York, 1995.

(16) Wold, S.; Johansson, E.; Cocchi, M. PLS-Partial least-squares projections to latent structures. In *3D QSAR in Drug Design*; ESCOM: Leiden, 1993; pp 523−550.

(17) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics - Part B*; Elsevier Science: Amsterdam, 1997.

(18) Stahle, L.; Wold, S. Multivariate data analysis and experimental design in biomedical research. *Prog. Med. Chem.* **1988**, *25*, 291–338.

(19) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003**, *46* (26), 5781–5789.

(20) Pirard, B.; Pickett, S. D. Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Model.* **2000**, *40* (6), 1431–1440.

(21) Sun, H. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J. Chem. Inf. Model.* **2004**, *44* (2), 748–757.

(22) Adenot, M.; Lahana, R. Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 239–248.

(23) Swets, J. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293.

(24) Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O. Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* **2005**, *48* (7), 2534–2547.

(25) Burnouf, C.; Pruniaux, M. P. Recent advances in PDE4 inhibitors as immunoregulators and anti-inflammatory drugs. *Curr. Pharm. Des.* **2002**, *8*, 1255–96.

(26) Dal Piaz, V.; Giovannoni, M. P. Phosphodiesterase 4 inhibitors, structurally unrelated to Rolipram, as promising agents for the treatment of asthma and other pathologies. *Eur. J. Med. Chem.* **2000**, *35*, 463–480.

(27) *QikProp, version 2.1*; Schrödinger, LLC: New York, NY, 2005.

(28) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409–422.

(29) *DRAGON Plus*, *version 5.4*; Talete srl: Milano, Italy, 2005.

(30) *MDL Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA, 2005.

(31) *Schrödinger QikProp 2.1 Operating Manual*; Schrödinger Press: 2005. The 95% range of similar values for known drugs is in brackets.

(32) Jorgensen, W. L.; Duffy, E. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355–356.

(33) *Spartan 02 Linux/Unix*; Wavefunction Inc.: Irvine, CA, 2003.

(34) *SIMCA-P*, *version 10.* 0; Umetrics AB: Umea, Sweden, 2005.

(35) *Glide*, *version 3.5*; Schrödinger, LLC: New York, NY, 2005.

(36) Card, G. L.; England, B. P.; Suzuki, Y.; Fong, D.; Powell, B.; Lee, B.; Luu, C.; Tabrizizad, M.; Gillette, S.; Ibrahim, P. N.; Artis, D. R.; Bollag, G.; Milburn, M. V.; Sung-Hou, K.; Schlessinger, J.; Zhang, K. Structural basis for the activity of drugs that inhibit phosphodiesterases. *Structure* **2004**, *12*, 2233–2247.

(37) Vigers, G. P.; Rizzi, J. P. Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **2004**, *47* (1), 80–89.

(38) *Phase*, *version 1.0*; Schrödinger, LLC: New York, NY, 2005.

(39) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44* (4), 502–511.

(40) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(41) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49* (20), 5851–5855.

CI800072R