

Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds

Edward O. Cannon · Ata Amini · Andreas Bender ·
Michael J. E. Sternberg · Stephen H. Muggleton ·
Robert C. Glen · John B. O. Mitchell

Received: 21 December 2006 / Accepted: 20 February 2007 / Published online: 27 March 2007
© Springer Science+Business Media B.V. 2007

Abstract We investigate the classification performance of circular fingerprints in combination with the Naive Bayes Classifier (MP2D), Inductive Logic Programming (ILP) and Support Vector Inductive Logic Programming (SVILP) on a standard molecular benchmark dataset comprising 11 activity classes and about 102,000 structures. The Naive Bayes Classifier treats features independently while ILP combines structural fragments, and then creates new features with higher predictive power. SVILP is a very recently presented method which adds a support vector machine after common ILP procedures. The performance of the methods is evaluated via a number of statistical measures, namely recall, specificity, precision, F-measure, Matthews Correlation Coefficient, area under the Receiver Operating Characteristic (ROC) curve and

enrichment factor (EF). According to the F-measure, which takes both recall and precision into account, SVILP is for seven out of the 11 classes the superior method. The results show that the Bayes Classifier gives the best recall performance for eight of the 11 targets, but has a much lower precision, specificity and F-measure. The SVILP model on the other hand has the highest recall for only three of the 11 classes, but generally far superior specificity and precision. To evaluate the statistical significance of the SVILP superiority, we employ McNemar's test which shows that SVILP performs significantly ($p < 5\%$) better than both other methods for six out of 11 activity classes, while being superior with less significance for three of the remaining classes. While previously the Bayes Classifier was shown to perform very well in molecular classification studies, these results suggest that SVILP is able to extract additional knowledge from the data, thus improving classification results further.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-007-9113-3) contains supplementary material, which is available to authorized users.

E. O. Cannon · A. Amini · A. Bender ·
M. J. E. Sternberg · S. H. Muggleton ·
R. C. Glen · J. B. O. Mitchell (✉)
Unilever Centre for Molecular Science Informatics,
Department of Chemistry, University of Cambridge,
Lensfield Road, Cambridge CB2 1EW, UK
e-mail: jbom1@cam.ac.uk

A. Amini · M. J. E. Sternberg · S. H. Muggleton
Structural Bioinformatics, Division of Molecular Biosciences,
Faculty of Natural Sciences, Imperial College,
London SW7 2AZ, UK

Present Address:

A. Bender
Novartis Institutes for Biomedical Research,
Lead Discovery Informatics, 250 Mass Ave., 3C463,
Cambridge, MA 02139, USA

Keywords Classification · Feature selection ·
Machine learning · Molecular similarity · Screening

Introduction

Molecular similarity searching is based on the “similar property principle” [1], which states that molecules with similar structures should exhibit similar properties [2]. This principle is reflected in neighbourhood behaviour [3], whereby structurally similar molecules to a molecule of known bioactivity are likely to exhibit the same activity. Similarity searching is widely used in classification studies in the academic and industrial communities to identify compounds for further testing (often referred to as “virtual screening”) [4].

The first step in a classification study is to locate molecules in chemical space using an informatics-based description of the molecule. Descriptors are usually classified by their dimensionality. One-dimensional descriptors use properties such as volume and log P [5]. Two-dimensional descriptors are derived from the connection table [6], whilst three-dimensional descriptors use geometric information from molecular structures in 3D space [7]. Probably the most commonly used descriptors are those found in 2D fingerprints [8, 9]. Such descriptors are usually binary in nature and typically encode the presence or absence of substructural fragments. More recently, there has been a shift towards circular substructure descriptors such as Extended Connectivity Fingerprints (ECFPs) and Functional Class Fingerprints (FCFPs), as provided by Scitegic in Pipeline Pilot [10]. One study using all of these fingerprints was carried out by Hert et al. [11]. They used a dataset of 102,000 molecules, composed of 11 different activity classes taken from the MDL Drug Data Report (MDDR) [12]. Their work focused on a comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. They concluded that circular substructure fingerprints were more effective than fingerprints based on hashing, dictionaries or topological pharmacophores. Circular fingerprints have also been applied to the prediction of pK_a values and metabolic sites, of which a review was recently compiled [13].

On a similar theme, Bender et al. [14] developed a circular substructure fingerprint. It describes the environment of each atom in the molecule by generating a fingerprint that takes into account the Sybyl atom types of its neighbouring atoms at one and two bonds' distance. A count vector of the occurrences of the atom types at a given distance from the central atom is constructed and results in a fingerprint with the number of count vector entries equal to the number of atoms in the molecule. The generation of this fingerprint is shown in Fig. 1.

Feature selection is considered the second step. The aim of this step is to obtain the features which are most relevant to the study. Examples of feature selection methods include genetic algorithms, information gain, gain ratio and the Gini Index [15]. Feature selection has been found to reduce

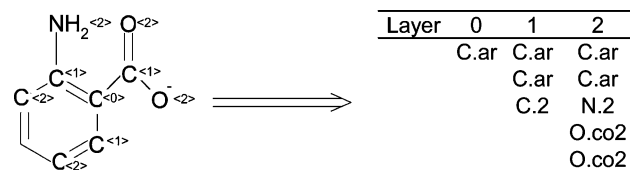


Fig. 1 Illustration of the ‘‘MOLPRINT 2D’’ circular fingerprints used in this study. For each heavy atom of the molecule, the type and number of atoms at a given number of separating bonds is kept in the descriptor

noise in datasets and improve classification performance [16].

In the third step, the molecular structures are then partitioned in chemical space using algorithms. Some of the most successful classification results in chemoinformatics have come from the field of machine learning. This field is the study of computer algorithms that improve automatically through experience [15]. In the domain of chemoinformatics, these methods are used to build models based on molecules assigned to a training set; these models are then used to predict properties or classify test set molecules into categories. A large number of machine learning methods exist. Our work concerns MOLPRINT 2D [14]. The MOLPRINT 2D method generates circular substructure fingerprints and uses information gain based feature selection; the selected features are then used by the Naive Bayes Classifier to predict the expected class of a molecule in a test set. These circular substructure fingerprints have also been used as descriptors for input into an inductive logic programming method (ILP) and a hybrid support vector machine/inductive logic programming method (SVILP) for the purpose of classification [17, 18].

To date, MOLPRINT 2D has been applied successfully to a variety of datasets and applications in the literature. Its debut was in 2004 [14], where it was used to classify molecules taken from the Briem/Lessel dataset (957 ligands) [19]. The performance of MOLPRINT 2D was compared to several other methods, such as feature trees and Daylight fingerprints, in its ability to retrieve five sets of active molecules seeded in the MDDR. In 2004, MOLPRINT 2D was applied to the Hert/Willett (102K) set [11, 20]. The objective of that work was to compare the performance of MOLPRINT 2D against alternative search methods [21] in their ability to retrieve active molecules. It was found that MOLPRINT 2D achieved considerably better results than binary kernel discrimination in combination with Unity 2D fingerprints.

More recently, in 2006, the method has been applied to two different datasets. The first is the World Anti-Doping Agency’s (WADA) 2005 Prohibited List [22], where MOLPRINT 2D was used to classify molecules taken from WADA [23] and the corresponding MDDR activity classes. This work compared MOLPRINT 2D with two other machine learning algorithms, random forest (RF) and k Nearest Neighbour (k NN). It was found that MOLPRINT 2D had the highest recall of positives but the lowest precision (recall and precision are defined in section ‘‘Measures of performance’’). The second example was the use of MOLPRINT 2D to classify a dataset of ~20 K molecules taken from the MDDR into different categories based on bitterness [24]. The aim was to identify important substructural features within the molecules necessary in discriminating bitter from non-bitter molecules. MOLPRINT

2D was able to predict 72% of the bitter compounds correctly. It is important to note that in neither of these studies was the the active/inactive MOLPRINT 2D cut-off score optimised, an improvement we introduce in this paper (see section “MOLPRINT 2D (MP2D)”).

One of the first successes of ILP in chemoinformatics was its application to the prediction of structure–activity relationships [25]. That work used a learning program called GOLEM, which enabled the authors to obtain better prediction results than the Hansch linear regression for the classification of trimethoprim analogues binding to *Escherichia coli* dihydrofolate reductase. In 1996, King et al. [26] used ILP with 2D descriptors on 229 aromatic and heteroaromatic nitro-compounds to predict mutagenic activity. ILP has also been used very recently to predict mutagenic activity in Xa inhibitors [27]. Ensemble methods are now becoming more popular in ILP, where the advantages of two methods are combined together to compensate for any shortcomings present in individual models. An earlier example of this ensemble methodology was demonstrated by Pompe et al. [28]. They used an ILP system in conjunction with a naive Bayesian classifier, and compared the results to a traditional ILP method. They showed that in certain domains the ILP-Bayesian methodology clearly outperformed more traditional methods. They believed this was due to the learner being able to detect strong probabilistic dependencies within the data if they existed. Other examples include: bagging [29], bootstrapping and boosting [30]. Over the last few years there has, however, been a growth in the use of support vector machines as a general purpose classification tool. In the field of molecular informatics, [31, 32] it has been shown to outperform other machine learning methods such as artificial neural networks and C5.0 decision trees. The facts that SVMs are very sophisticated robust classification tools and that there has been little coverage in the literature on SVILP models in the domain of chemoinformatics are just two reasons why we use ILP models and SVILP in this work.

In this work, we compare MOLPRINT 2D circular substructure fingerprints in combination with the Naive Bayes Classifier, ILP and SVILP using the Hert/Willett dataset. The following section “Methods”, gives details of the materials and methods used in this work, the results are presented and discussed in section “Results and discussion”. Section “Conclusion” summarises our conclusions.

Methods

Dataset and descriptor generation details

The Hert/Willett dataset [11] was taken from the MDDR (2003.01 version) and contains molecules from 11 different

activity classes. The active molecules were augmented with 94,290 inactive molecules. The inactives are molecules taken from the MDDR which are listed in the Hert/Willett dataset, but are not considered to be a member of any of the 11 activity classes. This lead to a dataset of 101,828 molecules. The breakdown of the data can be found in Table 1.

The molecules were retrieved in SDF format and converted to Sybyl mol2 format using OpenBabel 2.0.1 [33] with the `-d` option to delete hydrogen atoms. The datasets were divided into three training and test sets for each activity class (F1–F3). In each case 100 active and 100 inactive molecules were randomly selected for the training set. The remaining actives and inactives were used as two separate test sets. MOLPRINT 2D fingerprints were generated for these molecules. This process was repeated for all 11 activity classes.

Circular fingerprint description

An illustration of the descriptor generation step, applied to an aromatic carbon atom is shown in Fig. 1. The distances (“layers”) from the central atom are given in brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C(0)) up to a given distance (here two bonds from the central atom) are constructed. MOLPRINT 2D fingerprints are then binary presence/absence indicators of count vectors of atom types.

MOLPRINT 2D (MP2D): Feature selection

The second step of MP2D after the fingerprints have been generated is to select features. The information content of individual atom environments was computed using the information gain measure of Quinlan [34, 35]. In essence,

Table 1 MDDR activity classes used in the study

Activity code	Activity class	Actives
A06233	5HT3 Antagonists	752
A06235	5HT1A Agonists	827
A06245	5HT Reuptake	359
A07701	D2 Antagonists	395
A31420	Renin inhibitors	1130
A31432	Angiotensin II AT 1 antagonists	943
A37110	Thrombin inhibitors	803
A42731	Substance P antagonists	1245
A71523	HIV Protease inhibitors	750
A78331	Cyclooxygenase inhibitors	636
A78374	Protein kinase C inhibitors	452

higher information gain is related to better separation between active and inactive structures. The information gain, I , can be written as:

$$I = S - \sum_v \frac{|D_v|}{|D|} S_v$$

where

$$S = - \sum_{i=0}^1 p_i \log_2 p_i$$

S is the information entropy, S_v is the information entropy in data subset v , $|D|$ is the total number of data points, $|D_v|$ is the number of data points in subset v , and p_i is the proportion of positive (if $i = 1$) and negative examples (if $i = 0$) in S . In each run 40 features were selected. This work was then repeated using 250 features.

MOLPRINT 2D (MP2D): Classification

The Naive Bayes Classifier relies on the assumption that features are independent [15]. Given a training set of feature vectors (\mathbf{F}) which contain features f_i and class labels (CL), a Bayesian classifier is able to predict the class a new feature vector belongs to, based on which class has the highest conditional probability $P(CL_v|\mathbf{F})$.

$$P(CL_v|\mathbf{F}) = \frac{P(CL_v)P(\mathbf{F}|CL_v)}{P(\mathbf{F})}$$

The equation shown below is the one we use in this work to perform our classification study. Molecules are represented by their feature vectors \mathbf{F} , and the logarithm of the resulting ratio $P(CL_1|\mathbf{F})/P(CL_2|\mathbf{F})$ is used to determine the class label of the molecule. The default MP2D score assigns molecules to class 1 if this logarithm is greater than

zero and to class 2 if less than zero. If a given feature from a molecule is not present in one of the data subsets CL_1 or CL_2 , the probability of class membership would drop to zero, meaning either class is equally likely.

$$\frac{P(CL_1|\mathbf{F})}{P(CL_2|\mathbf{F})} = \frac{P(CL_1)}{P(CL_2)} \prod_i \frac{P(f_i|CL_1)}{P(f_i|CL_2)}$$

In previous work [22], one of the main problems found with MP2D is that it overpredicts false positives. To reduce the number of false positive predictions, in this work, we optimised the MP2D cut-off score for the active/inactive partition in each of the training sets based on obtaining the highest Matthews Correlation Coefficient (MCC).

$$MCC = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

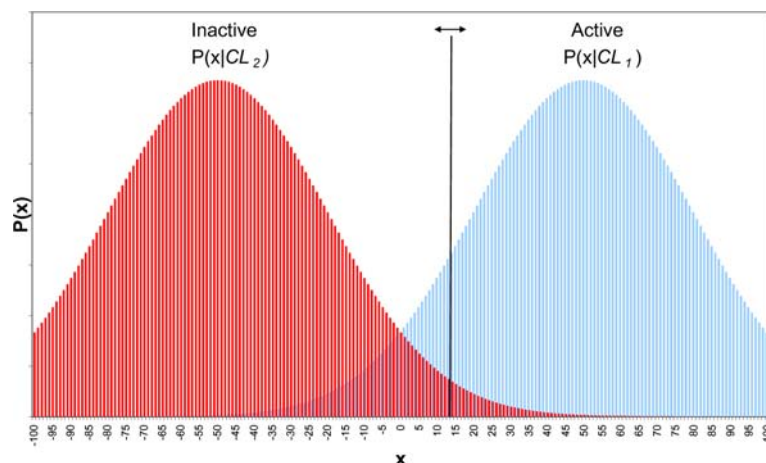
The optimised cut-off scores were subsequently used on the test set to determine whether a molecule was predicted active or inactive.

Figure 2 simply shows two distributions of MP2D scores. The first distribution CL_2 , is shown in red and represents scores for inactive molecules. The second distribution CL_1 , in blue, is for actives. The black line represents the optimum MP2D cut-off score.

ILP

Inductive Logic Programming learns logic rules from background knowledge and experimental observations [36]. The observations are the positive and negative examples that are, respectively the more active and less active molecules. The background knowledge is chemical fragments which could be either two- or three-dimensional. The ILP algorithm had been implemented in CProgol [37]. CProgol randomly selects a positive example and con-

Fig. 2 Optimising MOLPRINT 2D active/inactive cut-off scores. Generally a trade-off between true positives (at the expense of false positives) and true negatives (at the expense of false negatives) is given when the Bayes Classifier is employed



structs hypotheses using the provided background knowledge. For each hypothesis it calculates the compression which is defined as follows:

$$C = P - N - L$$

C is the compression, P and N are the number of positives and negatives covered by the hypothesis and L is the length of rule which is simply the number of fragments in the hypothesis. Compression is a measure of the power of the hypotheses constructed by the program. In the first step, the algorithm chooses only the best rule with maximum compression which is called the general rule. The calculation is continued on the next positive example, but the redundant (repetitive) examples relative to the new background knowledge are removed. After training is completed, the program applies the rules derived from the training set to the “external” (hypothetical) test set. Figure 3 schematically explains ILP for an example dataset of N positives.

One of the major strengths of logic-based structure activity relationships is that the program is able to construct chemical fragments in the form of logic relations using the basic atom and bond information. For instance, an OH group can be defined as:

oh(M):- atom(M, A, o, sp^3), atom (M, B, h, h), bond (M, A, B, 1).

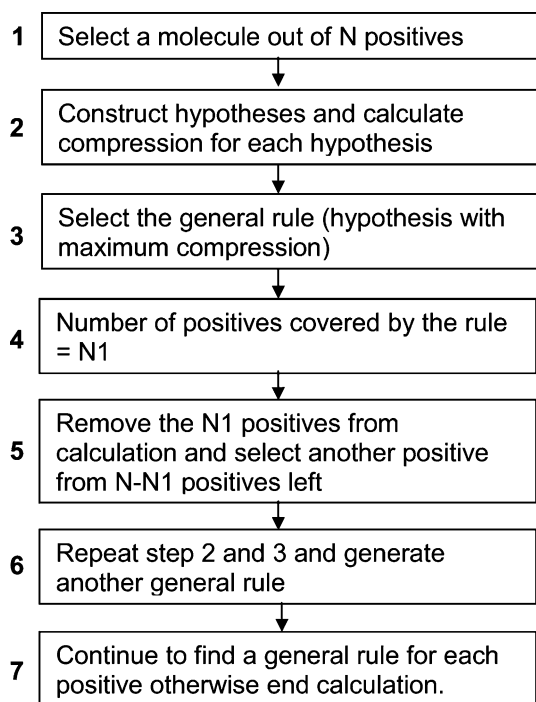


Fig. 3 Schematic representation of ILP. Starting with general rules covering as many instances as possible, rules are iteratively refined to cover as many instances of the training dataset as feasible in every step

That means: a molecule (M) has an OH group if it has a sp^3 oxygen atom and a hydrogen atom and the bond between these two atoms is single. A and B are used to label each atom. In the current study we have used the pre-processed background knowledge where all of the fragments have been prepared for the ILP calculations. Logic rules have one of the following formats: (1) active(A):-fragment(A) (2) active(A):-fragment1(A), fragment2(A) (3) active(A):-fragment(A), fragment2(A), fragment3(A). In type 1, molecule A is active if it has fragment1; in type 2, molecule A is active if it has fragment1 and fragment2; in type 3, molecule A is active if it has fragment1, fragment2 and fragment3.

We have done separate studies using ILP and SVILP for toxicology and activity predictions. One article on toxicology prediction is expected to be published soon. The logic rules are interpretable and this is one of the main advantages of the logic-based Structure Activity Relationships (SAR) with respect to the other methods. The description of the rules is beyond the scope of this article because of the number of rules and number of targets.

Support vector logic programming (SVILP)

Support vector inductive logic programming (SVILP) is a machine learning method that combines the advantages of a support vector machine (SVM) with ILP [18]. ILP is used to learn logic rules as explained in the last section, however, all of the hypotheses constructed by ILP that have positive compression are collected for quantification. Therefore SVILP, unlike ILP, is not dependent on the general rules. The number of collected rules is dependent on the size of the dataset, the complexity and the diversity of the molecules. In the current study, for the same number of molecules in the training sets, the number of rules varies from less than 100 to about 1,000 for different targets. Figure 4 shows the SVILP algorithm. Molecules are in rows ($M_1 \dots M_n$) where n is the number of molecules, and rules are in columns ($X_1 \dots X_m$) where m is the

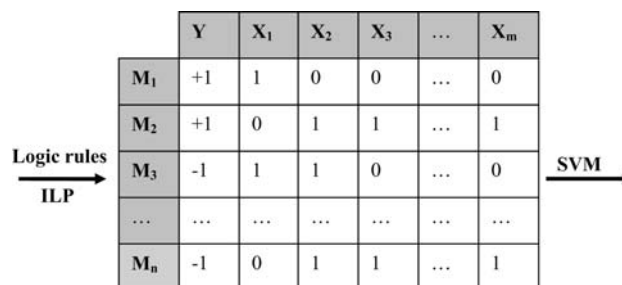


Fig. 4 Schematic representation of SVILP

number of features. The first column is the value of observed activities (if available) or in case of classification, +1 for positive and -1 for negatives (Y). Rules are considered as features and a value of ‘1’ is considered if the rule covers the molecule, otherwise a value of ‘0’ is designated. The result is a binary matrix, Fig. 4. The support vector machine is then used for classification or regression. In this study we used SVMLight for classification [38]. For all calculations in this work, we used an SVM with a linear kernel and all the default parameters provided in SVMLight.

Measures of performance

We generated a (2 × 2) confusion matrix for each of the three runs (F1–F3) and repeated this process for each activity class. The measures we have used to assess the classification results are recall, specificity, precision, the F-measure and the Matthews Correlation Coefficient (previously defined in sections ‘MOLPRINT 2D (MP2D): Feature selection and MOLPRINT 2D (MP2D): Classification’’).

$$\text{Recall} = \frac{t_p}{t_p + f_n} \quad \text{Specificity} = \frac{t_n}{t_n + f_p}$$

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad \text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

t_p is simply the number of true positives, that is molecules of a particular activity being classified as exhibiting that activity. t_n represents inactive molecules not of the activity under question correctly predicted. f_p and f_n represent the number of molecules incorrectly predicted to be active and inactive, respectively.

Kendall's W coefficient

We have also applied Kendall's W Coefficient of Concordance [39] to compare the inter-rater agreement between the 11 activity classes and the results for each classifier. In this study, each classifier is an object and each activity class is a rater. The Kendall's W Coefficient can be interpreted as a coefficient of agreement among raters. The coefficient W ranges from 0 → 1. A value of 1 indicates complete inter-rater agreement, 0 indicates complete disagreement.

Receiver operating characteristic curve

Here we have simply calculated the area under the ROC curve, for each activity class, run (F1–F3) and machine learning method. It is the most commonly used method in

the machine learning community, and gives an indication of the likelihood that a classifier will assign a higher score to a positive example than a negative example if one from each class were picked at random.

Enrichment factor

The enrichment factor (EF) is a measure of the ratio of the number of active molecules retrieved in a test set in comparison to what would have been expected if the molecules were selected at random at a given percentage of the ranked test set.

NOAMR = Number of active molecules retrieved

NEBR = Number of active molecules expected by random

$$\text{EF} = \left(\frac{\text{NOAMR}}{\text{NEBR}} \right)$$

In this work we have calculated the EF value for the top 1% and 5% of the ranked test set.

McNemar's test

McNemar's test [40] evaluates the significance of the difference between two methods. In comparison between method 1 and 2, A is the number of times both methods have correct predictions; B is the number of times method 1 has a correct prediction and method 2 has a wrong prediction; C is the number of times method 2 has a correct prediction and method 1 has a wrong prediction; D is the number of times both methods have incorrect predictions; the McNemar's χ^2 can then be calculated as shown below.

$$\chi^2 = \frac{(B - C)^2}{(B + C)}$$

Statistical significance is then evaluated by finding the probability associated with that χ^2 value using a reference table. The difference is considered to be significant if $p < 0.05$. B and C in the above equation are calculated separately for actives and inactives and since the number of inactives in this calculation (~94 K) is far more than the number of actives (200–1,300), larger B and C values are found for inactives. It is concluded that the significance test is dominated by the inactives. The values of B and C for inactives, however, could be scaled down using a function:

$$f = \frac{\text{number of actives}}{\text{number of inactives}}$$

by using this function, the actives and inactives now have an equal contribution to the McNemar significance test.

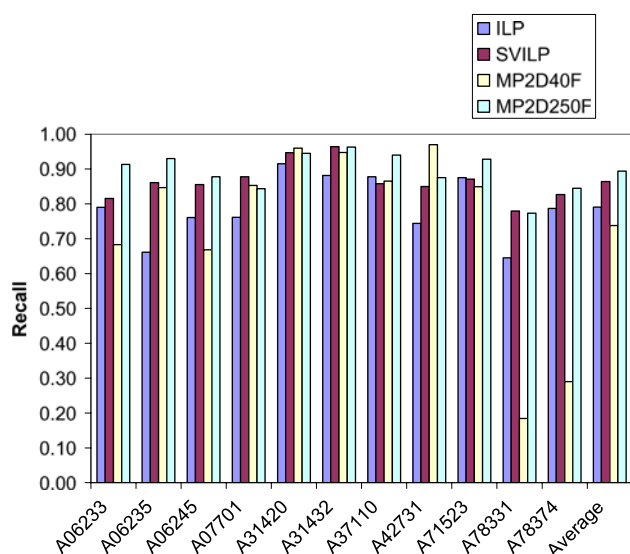


Fig. 5 Recall for each activity class and method. While ILP and SVILP show broadly similar performance across the datasets, the Naive Bayes Classifier shows much smaller recall on cyclooxygenase and protein kinase C inhibitors (classes A78331 and A78374)

Results and discussion

All results are summarised in Figs. 5–12 below and the breakdown is given in Table 2. The results presented here have been averaged over three runs (F1–F3) and have been calculated using the optimised thresholds with respect to MCC on the training set. Table 3 gives information on the

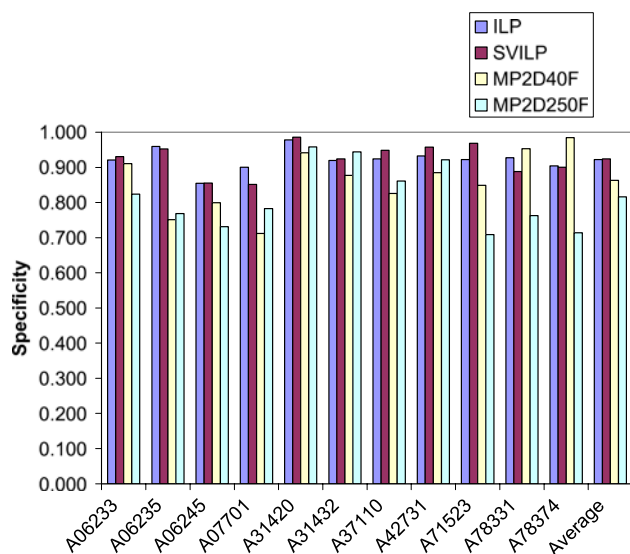


Fig. 6 Specificity for each activity class and method. Again, ILP and SVILP show broadly similar performance across all activity classes. As opposed to the previous recall plot, the cyclooxygenase and protein kinase C inhibitor datasets show very high specificity, thus the Naive Bayes Classifier represents a different trade-off position than the other methods

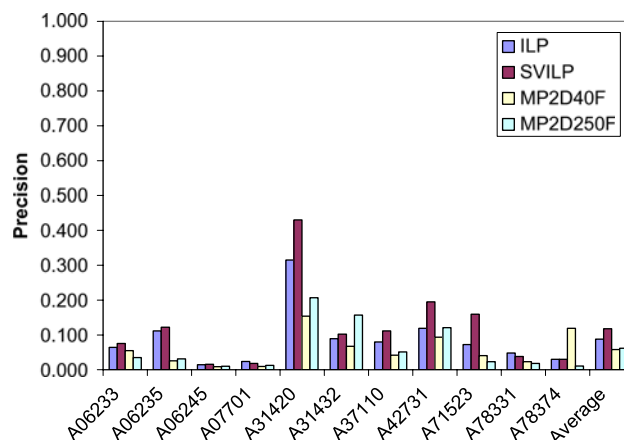


Fig. 7 Precision for each activity class and method. Here, wide variability is observed both across methods and datasets. Highest precision is observed for the renin inhibitor dataset (A31420), which has previously been shown to be atypical in size and thus easier to separate [41]

McNemar’s test results, comparing every method against every method for each activity class. Supplementary information on the optimised MP2D scores can be found in supplementary Table 4. Information on how well the Naive Bayes Classifier performed using a default cut-off value of zero can be found in supplementary Table 5. In supplementary Table 5 we present the recall, specificity, precision, F-measure and MCC values for the default Naive Bayes Classifier. The recall is higher on average using the default MP2D cut-off, however the specificity, precision, F-measure and MCC values are considerably lower on average. These trends are due to the lower MP2D cut-off score in the default model, which means that more

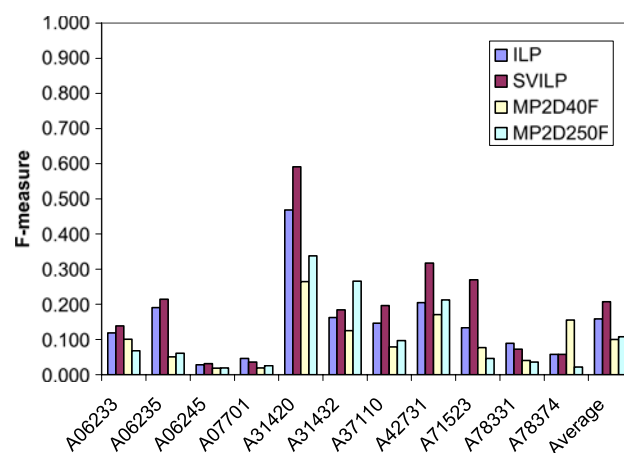


Fig. 8 F-measure for each activity class and method. While similar results are observed across classes and methods, the most profound drop is with the Naive Bayes Classifier applied to the 5HT1A agonists and HIV protease inhibitors

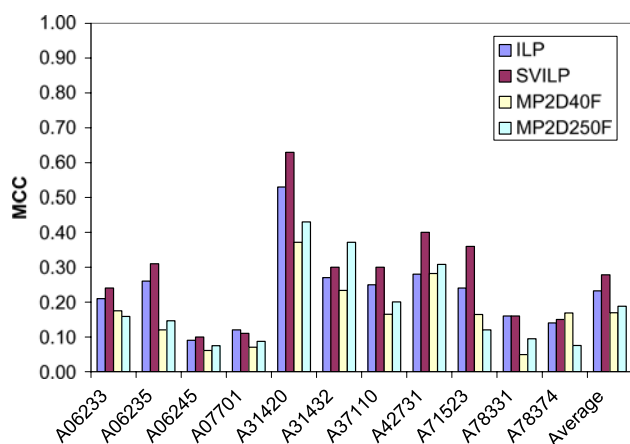


Fig. 9 Matthews Correlation Coefficient for each activity class and method. The results are fairly consistent with the previous performance measures, with the renin inhibitors having the spoil of the high values, whilst the cyclooxygenase and protein kinase C inhibitors show much more reserved estimates

molecules are classified as being active, resulting in an increase in the recall of positives.

MP2D shows the highest overall recall of positives, coming top in eight of the 11 classes (Fig. 5). The results show that more homogeneous classes such as thrombin inhibitors (A37110, 0.42), substance P antagonists (A42731, 0.40) and HIV protease inhibitors (A71523, 0.45) have a considerable enhancement in recall using MP2D (values in the brackets represent the activity class code and average intra-class Tanimoto similarity using the Unity 2D

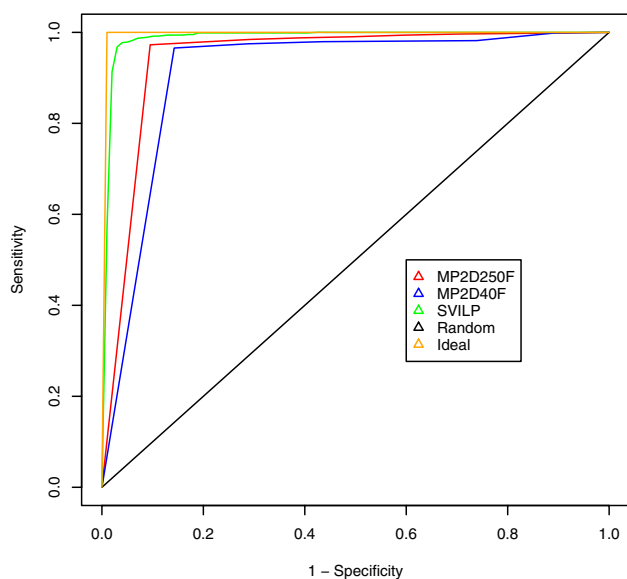


Fig. 10 Sample ROC plot for the second run of the A31432 activity class shows very steep ascents, in line with retrieving a large number of actives in the first few percent of the ranked database. These curves are close to the perfect case scenario

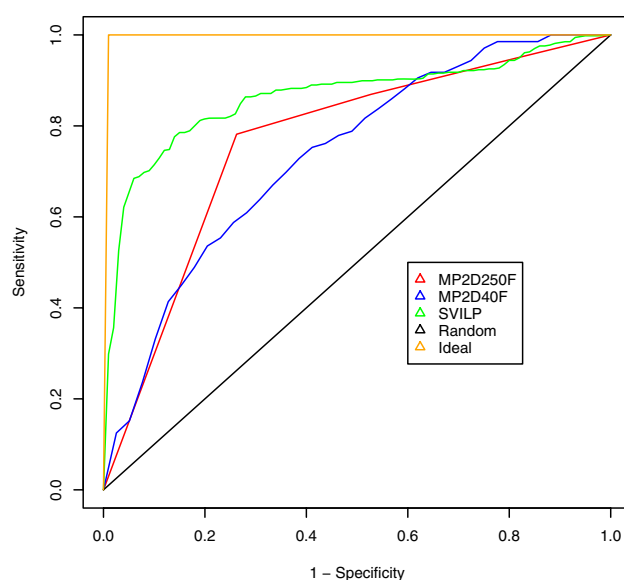


Fig. 11 ROC curve for the second run of the A78331 activity class. These curves are considerably less steep than those in Fig. 10, however the enrichment when compared to random is still great

fingerprints [11]). MP2D using only 40 features however performs particularly badly for the A78331 and A78374 classes. This may be due to the molecular skeletal diversity of the cyclooxygenase (COX) inhibitors and the protein kinase C inhibitors [11]. The poor performance in these two classes could however be attributed to the fact that the 40 features selected in the filtering stage may not be enough to differentiate active from inactive molecules using the Naive Bayes Classifier. In contrast, the SVILP method tends to perform second best in this measure with the ILP

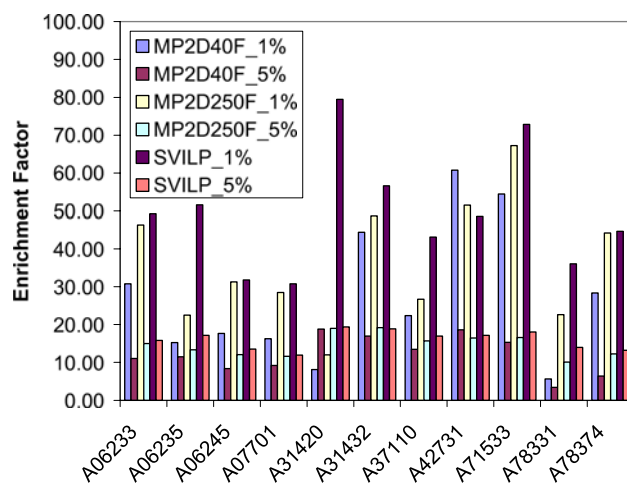


Fig. 12 Enrichment factor at the top one and five percent of the ranked database. Similar trends in enrichment factor are found as was the case for previous measures. The A06245, A07701 and A78331 classes have lower enrichment whilst higher values are noted from the A31420, A31432 and A71523 classes

Table 2 Recall, specificity, precision, F-measure, Matthews correlation coefficient, area under the ROC curve and enrichment factors taken at 1% and 5% of the ranked database all averaged over three

runs. AUC, EF at 1% and EF at 5% values have been omitted from this table for the ILP method on the basis that it predicts molecule class using a binary system with no score component

Measure	Method	A06233	A06235	A06245	A07701	A31420	A31432	A37110	A42731	A71523	A78331	A78374	Average
Recall	ILP	0.790	0.662	0.761	0.762	0.916	0.881	0.878	0.744	0.875	0.646	0.787	0.791
	SVILP	0.816	0.861	0.855	0.878	0.947	0.964	0.858	0.850	0.871	0.780	0.827	0.864
	MP2D40F	0.683	0.847	0.668	0.853	0.960	0.948	0.865	0.970	0.849	0.185	0.290	0.738
	MP2D250F	0.913	0.930	0.878	0.844	0.945	0.963	0.940	0.875	0.928	0.774	0.845	0.894
Specificity	ILP	0.921	0.959	0.855	0.900	0.978	0.920	0.924	0.933	0.922	0.927	0.904	0.922
	SVILP	0.931	0.952	0.856	0.852	0.986	0.924	0.949	0.957	0.968	0.888	0.900	0.924
	MP2D40F	0.911	0.751	0.800	0.712	0.941	0.877	0.826	0.885	0.849	0.953	0.985	0.863
	MP2D250F	0.824	0.769	0.731	0.782	0.958	0.944	0.861	0.922	0.709	0.763	0.714	0.816
Precision	ILP	0.064	0.111	0.014	0.024	0.315	0.089	0.080	0.119	0.072	0.048	0.030	0.088
	SVILP	0.076	0.122	0.016	0.018	0.430	0.102	0.111	0.195	0.159	0.038	0.030	0.118
	MP2D40F	0.055	0.026	0.009	0.010	0.154	0.067	0.042	0.093	0.041	0.023	0.119	0.058
	MP2D250F	0.035	0.031	0.010	0.013	0.207	0.157	0.051	0.121	0.023	0.018	0.011	0.062
F-Measure	ILP	0.119	0.191	0.028	0.046	0.468	0.162	0.146	0.205	0.133	0.089	0.057	0.158
	SVILP	0.138	0.214	0.031	0.036	0.591	0.184	0.197	0.318	0.269	0.073	0.058	0.208
	MP2D40F	0.101	0.050	0.018	0.019	0.265	0.125	0.079	0.170	0.077	0.040	0.155	0.100
	MP2D250F	0.067	0.061	0.019	0.025	0.338	0.266	0.097	0.212	0.046	0.036	0.022	0.108
MCC	ILP	0.211	0.260	0.091	0.123	0.530	0.266	0.251	0.280	0.239	0.162	0.141	0.232
	SVILP	0.235	0.314	0.105	0.114	0.633	0.301	0.298	0.395	0.365	0.156	0.146	0.278
	MP2D40F	0.175	0.120	0.061	0.071	0.371	0.233	0.165	0.282	0.165	0.049	0.169	0.169
	MP2D250F	0.159	0.146	0.075	0.087	0.430	0.371	0.200	0.308	0.120	0.094	0.076	0.188
AUC	SVILP	0.953	0.957	0.930	0.910	0.990	0.987	0.963	0.960	0.977	0.883	0.933	0.949
	MP2D40F	0.847	0.810	0.746	0.788	0.950	0.915	0.865	0.931	0.874	0.663	0.627	0.820
	MP2D250F	0.869	0.849	0.803	0.818	0.952	0.959	0.907	0.930	0.819	0.776	0.790	0.861
EF at 1%	SVILP	49.2	51.6	31.8	30.7	79.5	56.6	43.1	48.5	72.9	36.0	44.6	49.5
	MP2D40F	30.7	15.3	17.6	16.3	8.1	44.4	22.3	60.8	54.5	5.6	28.3	27.6
	MP2D250F	46.3	22.5	31.3	28.5	12.0	48.7	26.6	51.6	67.3	22.6	44.1	36.5
EF at 5%	SVILP	15.8	17.1	13.5	11.9	19.4	18.9	17.0	17.1	18.0	14.0	13.2	16.0
	MP2D40F	11.0	11.5	8.4	9.2	18.8	17.0	13.5	18.6	15.3	3.4	6.4	12.1
	MP2D250F	15.0	13.4	12.0	11.6	19.0	19.2	15.6	16.4	16.6	10.1	12.2	14.6

coming in last. The ILP method also found the A78331 and A78374 classes the hardest to predict, however only a little dip in performance was noted.

The SVILP method appears to have the highest specificity (Fig. 6), being ranked first on six out of the 11 occasions. The ILP method gave the best results for the A06235 and A07701 classes, whilst MP2D retrieved the most negatives for A78331 and A78374 activity classes. One possible explanation for the low recall but high specificity using MP2D with 40 features with the A78331 and A78374 classes is that it could be a result of the low MP2D scores assigned to the active molecules. Given that the optimised cut-off scores are positive, this means fewer molecules will be classified as being positive and more as negative; effectively this means MP2D with 40 features classifies a lot of molecules in these two classes as false negatives, hence a lower recall and higher specificity.

When considering precision (Fig. 7), SVILP performs better than the other two methods. Notable differences for the A31420 and A71523 classes in comparison to the other methods were found. The precision for MP2D is considerably lower than the other methods except for the A31432 and A78374 activity classes. The highest precision is observed for the renin inhibitor dataset (A31420), which has previously been shown to be atypical in size and thus easier to separate [41]. In general, the precision of predicting positives shown in Table 2 is very low. However, if one considers that the ratio between active and inactive molecules in this dataset is around 1:100, values exceeding 0.01 are better than random.

The F-measure is a measure that combines both recall and precision. Figure 8 shows that the SVILP method is the best overall using this measure. This method ranks first in seven out of the 11 classes. MP2D is the best for the

Table 3 Results of McNemar's significance test. p is the two-tailed probability; BM is the better method in a pair-wise comparison. A method is considered to be significantly better if $p < 0.05$

		ILP–SVILP	SVILP–MP2D40F	SVILP–MP2D250F	ILP–MP2D40F	ILP–MP2D250F	MP2D40F–MP2D250F
A06233	P	0.03	< 0.0001	0.64	< 0.0001	0.37	< 0.0001
	BM	SVILP	SVILP	SVILP	ILP	MP2D250F	MP2D250F
A06235	P	< 0.0001	< 0.0001	< 0.0001	0.49	0.008	< 0.0001
	BM	SVILP	SVILP	SVILP	ILP	MP2D250F	MP2D250F
A06245	P	0.01	< 0.0001	0.03	0.051	0.94	0.036
	BM	SVILP	SVILP	SVILP	ILP	ILP	MP2D250F
A07701	P	0.02	0.0003	0.02	0.03	0.41	0.12
	BM	SVILP	SVILP	SVILP	ILP	ILP	MP2D250F
A31420	P	< 0.0001	0.01	0.02	0.68	0.52	0.85
	BM	SVILP	SVILP	SVILP	MP2D40F	MP2D250F	MP2D250F
A31432	P	< 0.0001	0.0006	0.21	0.21	< 0.0001	< 0.0001
	BM	SVILP	SVILP	MP2D250F	MP2D40F	MP2D250F	MP2D250F
A37110	P	0.83	< 0.0001	0.85	< 0.0001	1	< 0.0001
	BM	SVILP	SVILP	SVILP	ILP	Even	MP2D250F
A42731	P	< 0.0001	< 0.005	0.35	< 0.0001	< 0.0001	< 0.018
	BM	SVILP	MP2D40F	MP2D250F	MP2D40F	MP2D250F	MP2D40F
A71523	P	0.01	< 0.0001	< 0.0001	0.0001	< 0.0001	< 0.02
	BM	SVILP	SVILP	SVILP	ILP	ILP	MP2D40F
A78331	P	< 0.0001	< 0.0001	0.0002	< 0.0001	0.35	< 0.0001
	BM	SVILP	SVILP	SVILP	ILP	ILP	MP2D250F
A78374	P	0.93	< 0.0001	0.0003	< 0.0001	0.0004	< 0.0001
	BM	SVILP	SVILP	SVILP	ILP	ILP	MP2D250F

A31432 and A78374 classes. There is no significant difference between SVILP and MP2D with 250 features for the A06233. These observations were supported by the McNemar's test results given in Table 3. The McNemar's test showed that in nine out of the 11 classes the SVILP method outperformed the other methods. MP2D with 250 features was favoured for the A31432 activity class using the McNemar's test. The probability was 0.21 indicating no significant difference between MP2D with 250 features and SVILP and was reflected in the F-measure scores of 0.266 and 0.184 for MP2D with 250 features and SVILP, respectively. The McNemar's test also supports the fact that there is little difference in performance across the A06233 class using SVILP and MP2D with 250 features (probability 0.64).

It can be seen from Fig. 9, that again the SVILP method appears to dominate the field in terms of MCC value. On seven occasions the SVILP performed the best. Particularly high results came from the A31420 class with the SVILP obtaining an MCC value of 0.63. The A06245 and A07701 classes were found to have the lowest MCC values. MP2D has the highest MCC values with the A31432 and A78374 classes. These results are in line with those for the F-measure, as would be expected, given the similar nature of the two performance measures.

In general, it is evident from these graphs that we see a divide in performance, with the SVILP and MP2D method with 250 features performing better than average, whilst MP2D with 40 features and the ILP models perform worse than average. These results are supported by the Kendall's W Coefficient of Concordance, with W coefficients of 0.71, 0.50, 0.81, 0.84 and 0.87 when using recall, specificity, precision, the F-measure and MCC as measures for the activity classes to act as raters to rank the four classification methods. The coefficients indicate there is a high level of inter-rater agreement between the ranks assigned to each classifier for each activity class and performance measure. The χ^2 probability associated with these values was $p < 0.05$ in all cases. This simply means the results in Table 2 are statistically significant for these performance measures. The McNemar's test results in Table 3 support the conclusion that the SVILP method is the best classifier, in nine out of the 11 classes the SVILP method outperformed the other two methods. A31432 and A42731 were the only activity classes where MP2D outperformed the SVILP method. In all cases bar the A42731 class, SVILP was significantly better than MP2D with 40 features. When compared to MP2D with 250 features, SVILP was the best method over all activity classes except A31432. In general the level of significance for the difference in performance

was a lot lower than in comparison between SVILP and MP2D with 40 features, indicating in general MP2D with 250 features is a better classifier than MP2D with 40 features.

Two example ROC curves are shown in Figs. 10 and 11. The first graph is for the A31432 dataset taken on the second run (F2). From Table 2 it can be seen that this dataset had the second highest average area under the curve (AUC), with values ranging from 0.92 to 0.99, which compare well with the ideal result of 1 (perfect classifier). A value of 1 indicates that no false positives and false negatives are present in the ranked test set. This effectively means that all the true positives were found at the top of the ranked database, with all true negatives coming lower; where the ranking is established based on a sorted list of scores (highest at the top). The area under the curve in this work was calculated using the trapezium rule. The values of 0.92 and 0.99 represent the proportion of the area of the ideal curve covered using our classifiers. Comparable area under the curve is achieved for the A31420 and A42731 datasets. Figure 11 shows a ROC curve for the A78331 dataset taken on the second run. The area under this curve for MP2D using 40 features is 0.73, the average AUC for MP2D using 40 features is 0.66. The A78331 and A78374 datasets were the most disappointing of all the runs, however the results are still far superior to an area of 0.50 expected if the molecules were to be sorted at random.

The EFs at 1% and 5% shown in Table 2 and Fig. 12 illustrate that the EF in general is higher at 1% than 5% and that the enrichment results for the SVILP method are better than either result by MP2D. The only anomalous result is the surprisingly low EF at 1% for MP2D using the renin inhibitor class. Here the average enrichment factor is only 12 compared to SVILP's 79, yet the AUC values are very comparable. The explanation is that firstly the renin test set is large (1,030), this would result in more active molecules being expected to be in the top 1% (10) rather than say four molecules with the A07701 activity class, hence the enrichment is partly dependent upon the size of the activity class. Secondly, after viewing the individual breakdown of the MP2D scores it became evident that some inactive molecules were assigned higher scores than any of the active molecules. This led to the actives falling lower in the rankings, hence we see a low enrichment at 1%. However, when we move to 5% we see the EF has increased to 19 in line with the 5% EF of SVILP, indicating that the vast majority of the active molecules in the renin test set fall in the top 2–5%. In fact, MP2D with 250 features retrieves 976 active molecules out of 1,030 in the top 5% of the test set (5% of 1,030 = 51.5, the EF at 5% is 18.96 therefore $18.96 \times 51.5 = 976$ active molecules). This result compares favourably to previous work [20], in that 94.44% of actives were retrieved for the renin dataset in the top 5% of the test set (which equates to 973 actives).

We had considered classifying the datasets using an SVM on its own. However, in an unpublished study, we found that the model based on the chemical fragments and SVM was unstable and some of the features acted as noise. It is thus imperative that a powerful feature selection method is used before testing all structural features using an SVM. ILP not only selects the best features, but it also combines the features and generates new information and in this article we have shown that it gives a substantial improvement across most measures of performance.

Conclusion

On the datasets examined here, which comprise 11 activity classes and about 102,000 structures, Support Vector Inductive Logic Programming outperforms both the optimised Naive Bayes Classifier and the standard Inductive Logic Programming with regard to the F-measure, which takes both recall and precision into account. SVILP is, for seven out of the 11 activity classes, the superior method, with six of the classification differences being significant in McNemar's test at a confidence level of $p < 5\%$. SVILP is a very recently presented method which adds a support vector machine after common ILP procedures. We thus present a combination of previous machine learning methods which improves results over considering individual features independently (such as the Naive Bayes Classifier) as well as over Inductive Logic Programming only. While previously the unoptimised Bayes Classifier was shown to perform very well in molecular classification studies, these results suggest that the optimised Bayes Classifier performs better. However, despite this improvement, the Support Vector Inductive Logic Programming is able to extract additional knowledge from the data, thus improving classification results further.

Acknowledgements E.O. Cannon, R.C. Glen and J.B.O. Mitchell thank Unilever plc and the EPSRC for funding. A. Bender thanks the Education Office of the Novartis Institutes for BioMedical Research for a postdoctoral fellowship. A. Amini, M.J.E. Sternberg and S.H. Muggleton thank the BBSRC for funding.

References

1. Johnson AM, Maggiora GM (1990) Concepts and applications of molecular similarity, eds. Wiley, New York
2. Bender A, Jenkins JL, Li Q, Adams SE, Cannon EO, Glen RC (2006) Molecular similarity: advances in methods, applications and validations in virtual screening and QSAR. In: Annual reports in computational chemistry, vol 2, pp 141–168
3. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) J Med Chem 39:3049
4. Bohm HJ, Schneider G (2000) Virtual screening for bioactive molecules ed. Wiley-VCH

5. Downs GM, Willett P, Fisanick W (1994) *J Chem Inf Comput Sci* 34:1094
6. Estrada E, Uriarte E (2001) *Curr Med Chem* 8:1573
7. Mason JS, Good AC, Martin EJ (2001) *Curr Pharm Des* 7:567
8. Leach AR, Gillet VJ (2003) *An introduction to chemoinformatics*. Kluwer, Dordrecht
9. Gasteiger J (2003) *Handbook of chemoinformatics*, eds. Wiley-VCH, Weinheim
10. Scitegic Inc. Retrieved from <http://www.scitegic.com/>
11. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) *Org Biomol Chem* 2:3256
12. Elsevier MDL, 2440 Camino Ramon, Suite 300, San Ramon, CA 94583, USA. <http://www.mdl.com/>
13. Glen RC, Bender A, Arnby CH, Carlsson L, Boyer S, Smith J (2006) *IDrugs* 9:199
14. Bender A, Mussa HY, Glen RC, Reiling S (2004) *J Chem Inf Comput Sci* 44:170
15. Mitchell TM (1997) *Machine learning*, ed. McGraw-Hill, New York
16. Liu YA (2004) *J Chem Inf Comput Sci* 44:1823
17. Muggleton SH, Lodhi H, Amini A, Sternberg MJE (2006) In: Holmes D, Jain LC (eds) *Innovations in machine learning*. Springer-Verlag, pp 113–135
18. Muggleton SH, Lodhi H, Amini A, Sternberg MJE (2005) *Proceedings of the 8th international conference on discovery science*. Springer-Verlag, 3735:163
19. Briem H, Lessel UF (2000) *Persepect Drug Discovery Design* 20:231
20. Bender A, Mussa HY, Glen RC, Reiling S (2004) *J Chem Inf Comput Sci* 44:1708
21. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) *J Chem Inf Comput Sci* 44:1177
22. Cannon EO, Bender A, Palmer DS, Mitchell JBO (2006) *J Chem Inf Model* 46:2369
23. World Anti-Doping Agency (WADA), Stock Exchange Tower, 800 Place Victoria, (Suite 1700), P.O. Box 120, Montreal, Quebec, H4Z 1B7, Canada. Retrieved from <http://www.wada.ama.org>
24. Rodgers S, Glen RC, Bender A (2006) *J Chem Inf Model* 46:569
25. King RD, Muggleton SH, Lewis R, Sternberg MJE (1992) *Proc Natl Acad Sci* 89:11322
26. King RD, Muggleton SH, Srinivasan A, Sternberg MJE (1996) *Proc Natl Acad Sci* 93:438
27. Buttingsrud B, Ryeng E, King RD, Alsberg BK (2006) *J Comput Aid Mol Des* 20:361
28. Pompe U, Kononenko I (1995) *Proceedings of the 5th international workshop on inductive logic programming*, pp 417–436
29. Dutra I, Page D, Santos Costa V, Shavlik J (2003) In: Matwin S, Sammut C (eds) *Proceedings of the 12th international conference on inductive logic programming*, vol 2583. *Lecture Notes in Computer Science*, Springer-Verlag, pp 48–65
30. Hoche S, Wrobel S (2001) In: Rouveirol C, Sebag M (eds) *Proceedings of the 11th interational conference on inductive logic programming*, vol 2157. *Lecture Notes In Computer Science*, Springer-Verlag, pp 51–64
31. Bender A, Glen RC (2004) *Org Biomol Chem* 2:3204
32. Barrett SJ, Langdon WB (2006) In: Tiwari A, Knowles J (eds) *Applications of soft computing: recent trends*, vol 19. Springer-Verlag, pp 99–110
33. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL (2006) *J Chem Inf Model* 46(3):991. The Open Babel Package (2006), version 2.0.1. Retrieved from <http://openbabel.sourceforge.net/>
34. Quinlan JR (1986) *Mach Learn* 1:81
35. A-Razzak M, Glen RC (1992) *J Comput Aided Mol Des* 6:349
36. Muggleton SH (1995) *New Generation Comput* 13:245
37. Muggleton SH, Bryant CH (2000) In: Cussens J, Frisch AM (eds) *Proceedings of the 10th international conference on inductive logic programming*. Springer-Verlag, pp 130–146
38. Joachims T (1999) *Making large-Scale SVM learning practical*. In: Schölkopf B, Burges C, Smola A (eds) *Advances in Kernel Methods-Support Vector Learning*, MIT-press, <http://svmlight.joachims.org>
39. Siegel S, Castellan NJ Jr (1988) *Nonparametric statistics for the behavioral sciences*. Boston, MA, McGraw-Hill
40. McNemar Q (1947) *Psychometrica* 12:153
41. Bender A, Glen RC (2005) *J Chem Inf Model* 45:1369