Descriptor Biclustering Improves Binary Compound Classification with Partial Least Squares

Megon J. Walker[1], *Terrence Wu*[1], *Andreas Bender*[2], *Simon Kasif*[1,*]

[1] Bioinformatics Program, Boston University, Life Sciences and Engineering Building, 24 Cummington Street, Boston, MA 02215, USA

[2] Discovery Technologies, Lead Discovery Informatics, Novartis Institutes for Biomedical Research, Inc., Cambridge, MA 02139, USA.

* Corresponding author e-mail: kasif@bu.edu, Phone: +1 (617) 358-1845, Fax: +1 (617) 353-6766

Running title: Biclustering for Binary Compound Classification

Keywords: Compound classification, virtual screening, quantitative structure activity relationships, property prediction, biclustering, partial least squares regression

Abstract

Compound classification is commonly performed by mathematical models based on the presence and absence of structural features and the numerical value of physicochemical features. Recently, biclustering has proven useful in the world of bioinformatics. Biclustering performs class-specific feature subset selection. By simultaneously clustering along both the class membership and the feature axes of a data matrix, biclustering identifies descriptor subsets that can be used as a cluster to predict class membership. Here biclustering is applied for the first time to small molecule property prediction of three datasets: factor Xa inhibitors, monoamine oxidase (MAO) inhibitors, and chemical mutagens. Employing MOE and Leadscope descriptors, it is found that descriptor biclustering significantly improves classification accuracy of the minority class of compounds, along with precision and F-measure, for two of the datasets. For the MAO dataset, a combination of MOE and biclustered Leadscope descriptors improves sensitivity from 42.9% to 54.7%, while specificity remains constant. Biclustered descriptors improve specificity for the mutagenicity dataset. Classification of the factor Xa dataset is near-ideal with and without biclustering. Biclustered descriptors also reduce overfitting to the training data. Differences between model performance on training and testing sets decrease 60% to 65% when full Leadscope fingerprints are reduced to only biclustered features. Overall, it can be concluded that biclustering identifies meaningful descriptor combinations that improve compound classification.

Introduction

Computational predictions of compound properties are based on the principle that structurally similar compounds are more likely to exhibit similar properties than

structurally dissimilar compounds.[1-4] Two steps are of importance here: firstly the derivation of descriptors and secondly the mathematical relationship between those descriptors and the property to be predicted. Well-established traditional statistical methods such as principal component analysis (PCA) and partial least squares regression (PLS) are readily interpretable,[5-7] but accuracy is limited in particular for large diverse data sets characterized by highly nonlinear structure-activity relationships.[8] Machine learning constructs have proven useful complements to mathematical transformation. Decision trees,[9-11] rule-based approaches,[12] neural networks,[13] support vector machines,[14] and self-organizing maps[15] are among the best-known examples. Computer-aided drug design and virtual screening often involve data mining large chemical spaces. While the choice of descriptors and modeling methods usually depends on the dataset and empirical factors, the general principle is that the descriptors should capture information relevant to the problem while introducing little noise. The mathematical model should be able to describe the relationship between descriptor vectors and the property to be predicted.[16, 17]

Often variable selection techniques are employed to distinguish variables associated with the phenomenon being studied from unrelated features. Genetic algorithms and evolutionary programming,[9, 11, 18, 19] particle swarms,[20] simulated annealing,[21, 22] and artificial ant colony systems[23, 24] have been employed to this end. RFSBoost uses random feature subset boosting to build ensembles of linear discriminant analysis models.[25] Such techniques combine feature selection and classification, reduce model complexity, and increase interpretability by reducing the number of molecular descriptors used in the predictive model.

This project introduces the biclustering algorithm to select subsets of mutually

dependent features. The merit of the approach is corroborated by the fact that chemical features do not contribute to bioactivity independently.[26] Quite the contrary, there are strong non-linear (not simply additive) effects at work. As will be outlined in the following section, this phenomenon is addressed by the biclustering algorithm, which considers multiple descriptors and local "pools" of features at once.

Data subsetting strategies have been proposed for chemical data mining. When compounds in the database are clustered and classifiers are built from each cluster, majority voting by the classifiers can improve classification accuracy.[27-30] Algorithmic clustering is common to those previous approaches and to the biclustering method presented here, but previous studies clustered only structures. In the approach presented here, clustering is performed simultaneously along both the compound as well as the feature axes.

Materials and Methods

a) Introduction to biclustering

Biclustering is a distinct class of clustering algorithms that simultaneously cluster the rows (instances, such as molecules) and columns (such as features) of data matrices. Biclustering is also called co-clustering, projective clustering, two-way clustering, block clustering, and subspace clustering. It was initially developed for image retrieval and for analysis of word-document co-occurrence tables, webpage-user browsing data, and market-basket data.[31, 32] The predominant biological application is identification of groups of genes that show similar activity patterns under a specific subset of the experimental conditions.[33] The term “biclustering” was first introduced by Cheng and Church[34] in the context of gene expression data analysis and pattern recognition in

microarray data.

This project applies the biclustering methodology to small molecule classification for the first time. Biclustering improves compound classification by identifying meaningful descriptor subsets that are relevant to bioactivity. Biclustering selects class-specific feature combinations. In this context, a bicluster is defined as a homogeneous subset of compounds characterized by

1. the same combination of features, and
2. the same class label (here active or inactive).

The difference between conventional and biclustering-based determination of quantitative structure activity relationships (QSAR) is shown in Figures 1 and 2a. In conventional QSAR, a large number of features are calculated and subjected to a mathematical modeling procedure (such as in this case PLS, Figure 1). In this project, biclustering precedes QSAR. Biclustering identifies "blocks" (or biclusters) of compounds from either class (active/inactive) which contain identical features and class labels (Figure 2a). Then PLS modeling is performed using those descriptors (Figure 1).

Figure 2a illustrates the biclustering procedure. Inactive bicluster #2 is a pure inactive bicluster. All member compounds share the same class label (I), and no active compounds contaminate the bicluster purity. Inactive bicluster #2 is characterized by the combination of features 1, 3, and 6. This particular subset of features is present only in inactive compounds. The feature combinations returned by biclustering form positive conjunction rules. All compounds that are members of a particular bicluster must contain all features characterizing the bicluster. If any of the features are absent in a compound, this compound cannot belong to the bicluster. Otherwise it will be included. For example,

compound 1 is excluded from active bicluster #1 because it lacks features 1 and 3.

The feature combinations defining each bicluster are ideally both recurrent and discriminant. Feature patterns observed in the training set are expected to occur in the testing set (recurrence). Testing set compounds characterized by the same feature combinations as a training set bicluster are expected to also show the same class label (discrimination). Each bicluster's feature subset forms a positive conjunction rule that can be used during data mining to delineate new compounds characterized by the same combination of features (and hopefully by the same class label in accordance with the molecular similarity principle). These positive conjunction rules are readily interpretable, and the chemical structures composing them contribute in combination to bioactivity.

The GEMS[35-37] biclustering implementation used in this project finds axis parallel biclusters. Unlike other traditional unsupervised clustering and biclustering methods that disregard class labels completely, GEMS is a supervised clustering analysis. GEMS was modified to maximize bicluster purity and size such that each bicluster's feature pattern offers extensive coverage of compounds all belonging to one single class. Not only is each bicluster's distinct feature signature specific to its pure subset of compounds, it also includes all compounds of the same class that contain the feature subset, and as many features as possible.

The most similar study to date also involved class-specific feature selection, but did not consider as many features as possible. Tarasov *et al*.[38] conducted a mutagenicity study with QSAR analysis of "univocal ensemble descriptors" consisting of two unrelated structural fragments present only in active or only in inactive compounds. This approach improved compound classification accuracy.

b) Datasets

Three datasets of varying size and chemical diversity were used. The dataset sizes are given in Table 1.

I. Factor Xa dataset. Fontaine *et al*.[39] released a small dataset of 279 factor Xa inhibitors and 156 non-inhibitors. This benzamidine series is characterized by the amidine group that binds the factor Xa pocket.

II. MAO dataset. Brown and Martin[40] divided a set of 1650 monoamine oxidase inhibitors into four classes of MAO inhibitory activity (3, 2, 1, and 0), with 3 being the most active class and 0 the least active. For this study, the 1360 compounds with 0 activity labels were classified as inactive. 290 compounds were labeled active: 115 class 1 compounds, 87 class 2 compounds, and 88 class 3 compounds. This set of compounds spans a number of diverse structural classes and has been analyzed extensively.[18, 41] However, the dataset is biased, as the structures were synthesized to follow up on a lead.[42]

III. Mutagenicity dataset. Kazius *et al*.[43] analyzed 4337 compounds in a mutagenicity study. The dataset contained 2401 mutagens and 1936 nonmutagenic compounds.

c) Descriptor calculation

Two dimensional structures were provided in SD format.

I. MOE Descriptors. All available MOE[44] 1D and 2D topological descriptors of the compounds were employed. Hydrogens were added and PEOE partial charges were assigned. Constant columns exhibiting the same value for more than 80% of the compounds were removed due to their minor information content. The remaining data columns were normalized to mean 0 and variance 1. (As a representative example, Table

2 lists molecular descriptors remaining after pre-processing of the first cross-validation partition training set for each dataset.) The dimension of the resulting property vector was 169, 112, and 162 for the factor Xa, MAO, and mutagenicity datasets, respectively.

II. Leadscope Descriptors. The standard Leadscope[45] structural feature hierarchy is based on over 27000 structural features and combinations of features typically found in small molecule drug candidates and creates a binary representation via structural keys. For each dataset, only Leadscope features occurring in at least five training set compounds were selected for subsequent analysis. As a representative example, the dimension of the resulting property vector was 484, 710, and 1597 for the first cross-validation partition training sets of the factor Xa, MAO, and mutagenicity datasets, respectively. The Leadscope feature vector length correlates with dataset size and diversity. (See Tables 1, 2, and 4.) Following elucidation of each fingerprint using the Leadscope hierarchy, there was no attempt to keep the substructural features mutually exclusive or unrelated. Thus, a molecule or bicluster may be characterized by both methyl-pyridine and 3-alkyl-pyridine descriptors even though these descriptors refer to an overlapping sets of atoms. Such related features are often included together in the same bicluster (Tables 5 and 6).

d) Biclustering and biclustering parameters

Biclustering was only applied to Leadscope descriptors. Class labels and Leadscope binary fingerprints for all training set compounds were input to a modified implementation of the GEMS biclustering software.[35-37] GEMS finds axis parallel biclusters. It is a supervised clustering analysis. GEMS had been modified to maximize bicluster purity and size such that as many columns and rows as possible were included

in each bicluster.

Given a data matrix and a subset of features, GEMS computes the maximum set of compounds that are contained in the bicluster defined by these features. The resulting submatrices produced as output are composed entirely of 1's (features present). This reduces the problem of computing biclusters to the problem of searching for subsets of features. The first step in the GEMS biclustering algorithm is the deployment of a sampling algorithm to find a subset of substructural features corresponding to a maximal subset of compounds. The goal of the Gibbs sampler is to optimize the chosen substructural feature subset so that the number of conserved compounds containing these substructures is the largest. Next, GEMS uses a local search step to refine the bicluster. When the largest bicluster has not been optimized in several iterations, a local procedure is invoked on this largest bicluster to further increase (if possible) the number of compounds in the bicluster by scanning every feature not included in the bicluster to see if this feature can be recruited and added to the bicluster without violating conservation criterion. The resulting refined bicluster is optimal in the sense that the number of compounds (bicluster size) cannot be increased by a single change in the number of features.

Biclustering of the binary fingerprint descriptors was performed using parameters and system commands detailed in Table 3. The number of biclusters discovered during each search procedure was most dramatically affected by three parameters. Biclusters were required to cover at least five compounds (-rN=5) since feature subsets covering only a few compounds would have little predictive power on the testing dataset. Biclusters were characterized by 2 to 50 features (-c=2, -cM=50). This high upper limit maximized

individual bicluster size and coverage. Each newly discovered bicluster was required to include at least one previously uncovered compound (-r=1). Training set compounds were weighted to optimize bicluster purity. During discovery of pure active biclusters, active compounds were weighted 1 and inactives were weighted -5, or -15. During discovery of pure inactive biclusters, active compounds were weighted -10 or -15 and inactives were weighted 1. Bicluster purity was rarely compromised. The maximum number of biclusters rendered (-n=200) was high enough that pure biclusters were exhausted and the search did not terminate prematurely.

e) Partial least squares (PLS) and PLS parameter tuning

PLS regression is a powerful, linear statistical modeling procedure where underdetermined systems are encountered. A PLS regression model will try to find the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space by transforming input data into an uncorrelated space of latent variables, similar to PCA (but including the output variable). Instead of finding the hyperplanes of maximum variance, it finds a linear model describing some predicted variables in terms of other observable variables. The variability in the response variable is considered in the attempt to maximize the covariance between predictor and response variables. Cross-validated model generation and regression analysis were performed using Wehrens's and Mevik's R-language software PLS (partial least squares),[46] an R package freely available from CRAN.[47] The nonlinear iterative partial least squares (NIPALS) algorithm[5-7, 48, 49] that determines the principal components (PCs) of a data matrix X in an iterative manner was used for this analysis. NIPALS

Two parameters were optimized during model fitting: the number of principal

components used in the regression model and the active/inactive classification threshold used to convert real-value regression output to binary class label predictions of active (1) or inactive (0) for both the training and testing set. The optimal number of principal components is often determined by selecting the number of principal components resulting in the smallest cross-validated root mean square error (RMSE). This implementation retained the number of principal components and the classification threshold that yielded the highest F-measure for the training set when used in combination. The F-measure performance metric was preferable to the overall accuracy or class-averaged accuracy because it resists selecting the classification boundary that merely predicts the majority class label for most the dataset instances.

f) Molecular diversity analysis

The molecular diversity of each dataset was determined using averaged Tanimoto coefficients and binary Leadscope fingerprints. The diversity index (DI) of each individual dataset was calculated by averaging the value of the Tanimoto coefficient between all pairs of compounds in the dataset. The representative index (RI) value is the mean Tanimoto coefficient of the compounds in two datasets being compared. For each dataset, the active and inactive compounds were compared to each other, as were the training set and testing set.[50] Results are displayed in Table 4.

g) Computational details

Each of the three datasets was randomly divided into a training set (63%) and a testing set (37%) of compounds such that the proportion of compounds in each of the original activity classes (4 classes for MAO, 2 classes for the factor Xa and mutagenicity datasets) was the same in the overall entire dataset, the training set, and the testing set. Five-fold

cross-validation was employed. It has demonstrated to be superior to leave-one-out cross validation with respect to model generalizability, though a truly external validation set would be the ideal situation.[51] Table 1 details the composition of a representative cross-validation partition for each compound dataset.

PLS models were generated for the following descriptor combinations:

1) all MOE descriptors

2) all Leadscope descriptors

3) all MOE descriptors and all Leadscope descriptors

4) only those Leadscope descriptors included in biclusters

5) all MOE descriptors and only those Leadscope descriptors included in biclusters

6) all MOE descriptors and Leadscope ensemble descriptors.

The first three descriptor combinations establish a baseline for performance. MOE physicochemical descriptors are indicative of whole-molecule properties such as solubility and cell permeability. Leadscope descriptors represent the presence and absence of structural features that determine ligand-target interactions. Both descriptors are also combined into a single feature vector.

The latter three descriptor combinations gauge whether descriptor biclustering improves compound classification beyond that achieved with either MOE or Leadscope descriptors alone or with all MOE and all Leadscope descriptors in combination. It is important to distinguish Leadscope bicluster features from Leadscope ensemble features (Figure 2b). Leadscope bicluster descriptors are derived when the full Leadscope fingerprint is reduced to only those descriptors included in biclusters. Leadscope ensemble descriptors are derived as follows. Each ensemble descriptor represents a

bicluster. Compounds containing all of the bicluster features have a corresponding ensemble descriptor value of 1. Compounds lacking any of the bicluster features have a corresponding ensemble descriptor value of 0. Thus, the sixth descriptor combination involved both MOE descriptors and individual features representing the compound's presence or absence in each bicluster.

Results and Discussion

a) Molecular diversity analysis

Prediction accuracy was previously shown to be strongly affected by the diversity of samples used in the training set.[50, 52, 53] Also, bicluster coverage of compounds was expected to correlate inversely with diversity among actives and inactives. Compounds that bind a target protein's active site and share common substructures required for bioactivity will be covered with only a few large biclusters. Inactive compound coverage is impeded by structural variance. Diverse inactives that fail to bind the target protein and have few common substructures will be covered by many smaller biclusters. (This phenomenon is widely applicable. Leo Tolstoy observed that **"**Happy families are all alike; every unhappy family is unhappy in its own way."[54]) For this reason, the molecular diversity of each dataset was analyzed: the diversity index (DI) of the overall dataset, the representative index (RI) between the active and inactive compounds, and the RI between each training set and its corresponding testing set. DI values are inversely proportional to diversity, and datasets with high RI are more representative of each other.

Table 4 displays molecular diversity data and bicluster information from a representative cross-validation partition for each dataset. The factor Xa dataset is the most cohesive, with the highest DI (0.215) and thus least diversity. Less than 500

Leadscope fragments (out of 27,000; less than 2%) were required to represent all 433 structures. The actives resemble each other (DI=0.236), and the inactives are cohesive (DI=0.251), but the relatively low RI between active and inactive compounds (0.185) suggests that the two classes of compounds are readily separable. The roughly equivalent average Tanimoto coefficients between and within the training and testing datasets suggests that the same patterns are represented in each. This dataset illustrates the correlation between dataset separability and biclustering coverage. Because of the homogeneity within each class and the lack of intra-class similarity, a few large biclusters achieve substantial compound coverage. Five active biclusters cover 94% of the actives, and ten inactive biclusters cover 70% of the inactives. This may be due to inclusion in the dataset of compounds with high and low binding activity in the factor Xa dataset at the exclusion from the dataset of compounds with intermediate activity. Overall, the factor Xa dataset should represent an "easy" dataset for classification studies.

The MAO inhibitor dataset displays intermediate size, chemical diversity, and separability. Roughly 700 Leadscope fragments were required to represent all 1650 MAO structures. The MAO dataset has intermediate diversity index values (active DI=0.104) equaling less than half the corresponding factor Xa index values. The actives and inactives are not readily separable, as neither class is clustered in chemical space. The inactive DI (0.077) is exceeded by the RI between active and inactive compounds (0.081), meaning that on average active and inactive compounds are (slightly) more similar to each other than the internal similarity of the set of inactive structures. Biclustering covers only the extreme bioactivity classes 0 and 3. The preponderance of inactives in the dataset explains the greater number of inactive biclusters. 90 inactive

biclusters are required to cover 79% of the inactives. Five active biclusters cover only 25% of the actives. Most of the MAO inhibitors included in active biclusters were originally classified with the highest bioactivity class 3 (30% of the actives). None of the bioactivity class 1 actives were biclustered. Class 1 and 2 MAO inhibitors exhibit lower bioactivity and were not readily separable due to substructural features shared with inactive compounds. When bioactivity was binned to binary class labels, information was lost and arbitrary thresholds were chosen, leading to borderline effects observed.[18]

The mutagenicity dataset is the largest, most chemically diverse, and least separable dataset. More than 1500 Leadscope fragments were required to represent all 4337 structures. Although the RI between actives and inactives was the lowest of all three datasets (0.063), so was the active DI (0.075). The heterogeneity within each class and the number of features common to both classes compromised biclustering coverage (as well as classification performance, shown later). 62 active biclusters cover 55% of the actives and 28 inactive biclusters cover 28% of the inactives. Unlike actives in the other two datasets, there is no single protein targeted by the mutagens. Substructural features responsible for mutagenicity can be as varied as features characterizing nonmutagens.

b) Bicluster feature interpretation

The full Leadscope fingerprint was reduced to only those Leadscope descriptors included in biclusters. Leadscope feature vectors for the factor Xa, MAO, and mutagenicity datasets were reduced in length from 473.2, 717.2, 1587 to 138.6, 218.4, 252, respectively (averages after cross-validation). Finding meaningful descriptor subsets is important when developing interpretable classification rules and reducing expense involved in assaying entire databases. Tables 5 and 6 illustrate compound clusters and

features derived during biclustering of the training set compounds.

The factor Xa dataset is a benzamidine series[39] with benzene, amidine, iminomethyl and amine groups common to both actives and inactives. Thus, presence or absence of other functional groups and different pathlength distances between substructures determine the bioactivity of compounds. Such benzamidine series and mimics are potential antithrombotic agents.[55, 56] The training set actives were covered by five or six biclusters during each round of cross validation. The four largest active biclusters are distinguishable from each other and from the inactives because of a functional group specific to each bicluster: an aromatic-path7-hacceptor, an alcohol group, a sulfonyl group, and a carboxylate group.

For the MAO dataset, most of the compounds included in active biclusters were irreversible inhibitors[57] originally classified with the highest bioactivity class 3. Biclustering separates 22 hydrazine-based inhibitors into the largest bicluster of actives. Three smaller active biclusters each contain five to nine propargylamine derivatives. Eight cyclopropylamines are in the remaining active bicluster.

Kazius *et al*.[43] analyzed the mutagenicity dataset in order to identify small substructural representations that detected over 70 mutagens with accuracy of at least 70%. These discriminant substructures, called toxicophores, were purported to "indicate an increased potential for mutagenicity." Eight general toxicophores were determined: aromatic nitro groups, aromatic amines, three-membered heterocycles, nitroso moieties, unsubstituted heteroatom-substituted heteroatom arrangements, azo bonds, aliphatic halides, and polycyclic aromatic systems. Taken together, these eight substructures detect 75% of all mutagens in the dataset. Biclustering identifies many of the same mutagenic functional

groups. For example, the aromatic nitro and amine groups are well established as toxicophores whose mechanisms of mutagenicity are explained by partially overlapping metabolic activation pathways. Compounds with large polycyclic aromatic systems of three or more fused rings, such as members of the second active bicluster described in Tables 5 and 6, can intercalate into DNA.[43] Biclustering of the training set resulted in more than sixty biclusters covering roughly 55% of the active mutagens. These feature combinations were more specific than the moieties described by Kazius *et al*. On the other hand, and feature specificity limited individual bicluster coverage of compounds.

c) Biclustering performance

Results obtained for the different descriptors, descriptor combinations, and biclustering procedures are shown in Tables 7-10.

As the diversity analysis indicated, the factor Xa dataset is the most readily separable. Factor Xa regression analysis required consistently fewer principal components for model building and resulted in smaller RMSE in comparison to analysis of the other datasets (Table 7). The structural homogeneity within each class and the lack of inter-class overlap enabled extensive coverage with only a few large biclusters. Analyzed separately, both MOE descriptors and Leadscope descriptors yield high accuracy (above 99% for the training and above 93% for the testing set) and F-measure (above 99% for the training and above 95% for the testing set). Classification is not substantially improved by any of the other four combinations of the descriptors, although MOE and Leadscope ensemble descriptors achieve the highest testing set specificity, precision, and F-measure (Table 8). Differences are minimal, which is in good part due to the near-optimal performance of all of the methods employed.

For the MAO and mutagenicity datasets, the structural heterogeneity within each class makes class separation more difficult. The substructures common to both actives and inactives result in lower bicluster coverage.

Overall, binary substructural fingerprints do not usually render the best QSAR fits and predictions when compared to real-value physicochemical descriptors.[58] In both the MAO and mutagenicity datasets, MOE descriptor analysis results in high classification accuracy of the majority class, while combining MOE and Leadscope descriptors was required to obtain the highest classification accuracy of the minority class.

The MAO dataset includes compounds with high and low binding activity (levels 3 and 0), along with compounds of intermediate activity (activity levels 1 and 2). Analysis with MOE descriptors required the most principle components of any descriptor combination (Table 7) and resulted in the highest specificity of the majority class MAO noninhibitors (Table 9). PLS analysis of the full Leadscope fingerprints increased minority class MAO inhibitor sensitivity 13% higher than when MOE descriptors were analyzed (from 42.9% to 55.6% testing set sensitivity), but specificity and precision suffered. All Leadscope descriptors used alone or in combination with all MOE descriptors yielded the highest RMSE (Table 7) and the lowest accuracy (Table 9). The full Leadscope fingerprints identified more true positives but also more false positives. Indeed, the full Leadscope fingerprints are not necessary for classification, as higher overall accuracy is achieved with the fewest principle components when the regression models is built from only the biclustered Leadscope features. The combination of MOE descriptors and the biclustered Leadscope features increases sensitivity 10%, precision 5%, and F-measure 8% higher than when MOE descriptors are used for regression analysis. The combination of MOE

and Leadscope ensemble descriptors does not grant a comparable advantage.

Harper *et al*.[41] performed the most comparable analysis. The MAO dataset was classified using kernel discrimination, neural network, and merged similarity search algorithms. Molecular descriptors included atom pairs and topological torsions. When 825 compounds (~145 actives) were randomly selected to train each classifier, the 200 top-ranked compounds in the testing set included 70-75 actives (at most 51.7% sensitivity and 37.5% precision, assuming the actives were also halved). When 200 compounds (~35 actives) were randomly selected to train each classifier, the 200 top-ranked compounds in the testing set included 82-98 actives (at most 38.4% sensitivity and 49% precision, assuming the actives were split proportionally). In comparison, the NIPALS model used here is built from a training set of 1024 compounds (179 actives) and correctly predicts 59 actives in the testing set (55% sensitivity and 43-46% precision).

The mutagenicity dataset does not contain compounds targeting a single protein. Analysis of MOE descriptors alone results in high sensitivity of the majority class mutagens (Table 10). Addition of the full Leadscope fingerprints increases minority class nonmutagen specificity by 11% but reduces sensitivity by misclassifying more mutagens as nonmutagens. This is reflected in the highest RMSE for the full Leadscope fingerprints alone and in combination MOE descriptors (Table 7). The combination of MOE descriptors with biclustered Leadscope features increases specificity 9%, precision 4%, and F-measure 3% higher than when MOE descriptors alone are used for regression analysis. Sensitivity is not compromised. The classification performance achieved by the PLS model and biclustered descriptors exceeds the 75% sensitivity obtained using the eight toxicophores identified by Kazius *et al*.[43]

Another trend is interesting to note. PLS models generated from biclustered Leadscope descriptors overfit the training data less and generalize to testing data better than models generated from full Leadscope fingerprints. The difference between model performance on the training set and model performance on the testing set was averaged across all three datasets (Tables 8-10). The full Leadscope fingerprints gave the following performance gaps: accuracy 0.15, sensitivity 0.16, specificity 0.26, precision 0.26, and F-measure 0.22. The corresponding differences in performance are 60% to 65% smaller for the biclustered Leadscope descriptors: accuracy 0.05, sensitivity 0.07, specificity 0.07, precision 0.10 and the F-measure 0.09. While this effect was only partially shown on the datasets employed here, we are currently investigating the application of the algorithm to new datasets.

Conclusion

Biclustering performs class-specific feature subset selection by simultaneously clusterting both substructural features and compounds to reveal feature combinations that describe pure clusters of compounds. Chemical structures included in biclusters contribute in combination to biological activity or inactivity. The positive conjunction rules resulting from biclustering are recurrent, disciminant, and readily interpretable.

The project has applied biclustering for the first time to small molecule property prediction of three datasets: factor Xa inhibitors, monoamine oxidase (MAO) inhibitors, and chemical mutagens. It was found that biclustering significantly improves classification for two of the datasets (MAO inhibitors and mutagens). For the MAO dataset, a combination of MOE and biclustered Leadscope descriptors improves sensitivity from 42.9% to 54.7%, while specificity remains constant (86.8% vs. 86.2%).

Biclustered descriptors increase specificity for the mutagenicity dataset. In both cases, improved precision, F-measure, and classification of the minority class of compounds are achieved by combining MOE and biclustered Leadscope descriptors or by using biclustered Leadscope descriptors alone. Classification of the factor Xa dataset is near-ideal with and without biclustering.

Biclustered descriptors also reduce overfitting to the training data. Differences between model performance for training and testing sets decrease 60% to 65% when full Leadscope fingerprints are reduced to only biclustered features. This is an important advantage when machine learning methods are employed.

Overall, it can be concluded that biclustering identifies meaningful descriptor combinations that improve prediction performance. The results indicate that biclustered descriptors combined with topological information provide better classification than binary substructural or 2D topological descriptors alone. Demonstration of descriptor biclustering on such diverse data promotes confidence that the method will generalize well. Future extensions to the analysis will involve the use of regression to predict real-value biological activity, toxicity, and mutagenicity instead of binned class labels.

Acknowledgements

The authors would like to thank Dr. Yvonne C. Martin of Abbott for providing the monoamine oxidase inhibitor data set and Paul Blower of Leadscope for helpful discussion of Leadscope fingerprints. Special thanks to Taner Kaya and Sandor Vajda for assistance in deriving the MOE descriptors. This work was funded by the National Science Foundation under grant no. 8691-5. Andreas Bender thanks the Education Office of Novartis for funding.

Figures and Tables

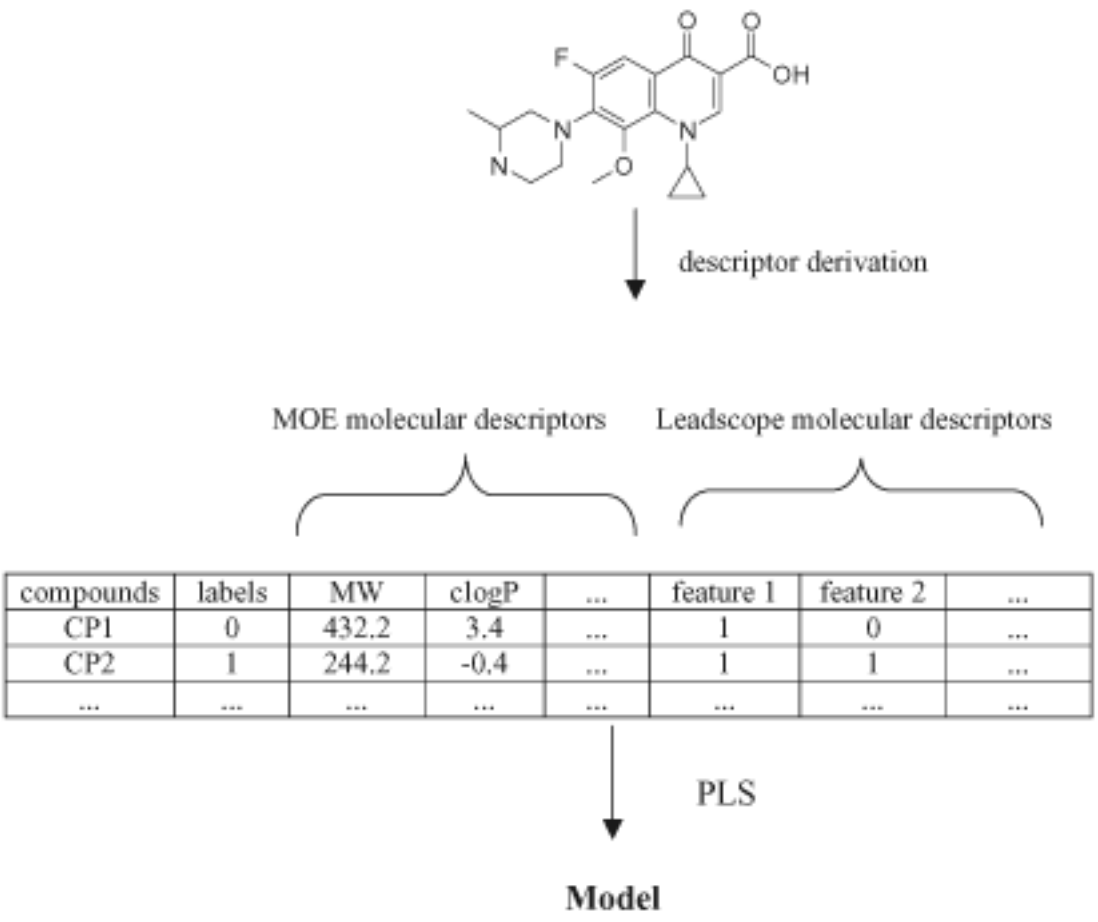


Figure 1. Flowchart for QSAR based on partial-least squares (PLS). Firstly, a large set of molecular descriptors is calculated using the programs MOE and Leadscope. In the second step, partial least squares regression is performed on the descriptors, using 37% hold out sets in five random validations.

a)

entire dataset

		features					
compounds	labels	**1**	**2**	**3**	**4**	**5**	**6**
CP1	**1**	0	1	0	1	1	0
CP2	**1**	1	1	1	0	1	0
CP3	**1**	1	1	1	1	1	0
CP4	**0**	1	1	1	0	0	1
CP5	**0**	1	0	1	1	0	1
CP6	**0**	1	0	1	0	1	1
CP7	**1**	1	1	1	0	1	0
CP8	**0**	0	1	1	0	1	0
CP9	**0**	0	1	0	1	0	1
CP10	**1**	1	0	0	0	0	1

biclusters of active compounds (label = 1)

biclusters of inactive compounds (label = 0)

bicluster #1

		features			
compounds	labels	**1**	**2**	**3**	**5**
CP2	**1**	1	1	1	1
CP3	**1**	1	1	1	1
CP7	**1**	1	1	1	1

bicluster #2

		features		
compounds	labels	**1**	**3**	**6**
CP4	**0**	1	1	1
CP5	**0**	1	1	1
CP6	**0**	1	1	1

b)

bicluster features

		features				
compounds	labels	**1**	**2**	**3**	**5**	**6**
CP1	**1**	0	1	0	1	0
CP2	**1**	1	1	1	1	0
CP3	**1**	1	1	1	1	0
CP4	**0**	1	1	1	0	1
CP5	**0**	1	0	1	0	1
CP6	**0**	1	0	1	1	1
CP7	**1**	1	1	1	1	0
CP8	**0**	0	1	1	1	0
CP9	**0**	0	1	0	0	1
CP10	**1**	1	0	0	0	1

ensemble features

		features	
compounds	labels	**bicluster #1**	**bicluster #2**
CP1	**1**	0	0
CP2	**1**	1	0
CP3	**1**	1	0
CP4	**0**	0	1
CP5	**0**	0	1
CP6	**0**	0	1
CP7	**1**	1	0
CP8	**0**	0	0
CP9	**0**	0	0
CP10	**1**	0	0

Figure 2. Flow chart for biclustering (a) and illustration of biclustered features and ensemble features (b). Biclustering was only applied to Leadscope descriptors. Bicluster features are derived when the full Leadscope fingerprint is reduced to only those descriptors included in biclusters. Each ensemble feature represents a bicluster.

Table 1. Compound datasets for biclustering and PLS classification. Number of compounds in the training and testing sets for the factor Xa, MAO, and mutagenicity datasets (first partition of random validation).

Dataset	Subset	Active	Inactive	Total
Factor Xa	training	174	97	271
	testing	104	58	162
MAO	training	179	845	1024
	testing	107	502	609
Mutagenicity	training	1505	1205	2710
	testing	893	717	1610

Table 2. MOE molecular descriptors describing each dataset. First partition of random validation.

Factor Xa dataset (169 MOE descriptors)	MAO dataset (112 MOE descriptors)	Mutagenicity dataset (162 MOE descriptors)
diameter	petitjean	diameter
petitjean	petitjeanSC	petitjean
petitjeanSC	radius	petitjeanSC
radius	VDistEq	radius
VDistEq	VDistMa	VDistEq
VDistMa	weinerPath	VDistMa
weinerPath	weinerPol	weinerPath
weinerPol	a_aro	weinerPol
BCUT_PEOE_0	a_count	BCUT_PEOE_0
BCUT_PEOE_1	a_IC	BCUT_PEOE_1
BCUT_PEOE_2	a_ICM	BCUT_PEOE_2
BCUT_PEOE_3	a_nH	BCUT_PEOE_3
BCUT_SLOGP_0	b_1rotN	BCUT_SLOGP_0
BCUT_SLOGP_1	b_1rotR	BCUT_SLOGP_1
BCUT_SLOGP_2	b_ar	BCUT_SLOGP_2
BCUT_SLOGP_3	b_count	BCUT_SLOGP_3
BCUT_SMR_0	b_double	BCUT_SMR_0
BCUT_SMR_1	b_rotN	BCUT_SMR_1
BCUT_SMR_2	b_rotR	BCUT_SMR_2
BCUT_SMR_3	b_single	BCUT_SMR_3
GCUT_PEOE_0	chi0v	GCUT_PEOE_0
GCUT_PEOE_1	chi0v_C	GCUT_PEOE_1
GCUT_PEOE_2	chi1v	GCUT_PEOE_2
GCUT_PEOE_3	chi1v_C	GCUT_PEOE_3
GCUT_SLOGP_0	reactive	GCUT_SLOGP_0
GCUT_SLOGP_1	Weight	GCUT_SLOGP_1
GCUT_SLOGP_2	a_heavy	GCUT_SLOGP_2
GCUT_SLOGP_3	a_nC	GCUT_SLOGP_3
GCUT_SMR_0	a_nN	GCUT_SMR_0
GCUT_SMR_1	a_nO	GCUT_SMR_1
GCUT_SMR_2	b_heavy	GCUT_SMR_2
GCUT_SMR_3	chi0	GCUT_SMR_3
a_aro	chi0_C	a_aro
a_count	chi1	a_count
a_IC	chi1_C	a_IC
a_ICM	VAdjEq	a_ICM
a_nH	VAdjMa	a_nH
b_1rotN	zagreb	b_1rotN
b_1rotR	balabanJ	b_1rotR
b_ar	PEOE_PC.	b_ar
b_count	PEOE_PC..1	b_count
b_double	PEOE_RPC.	b_double
b_rotN	PEOE_RPC..1	b_rotN
b_rotR	PEOE_VSA.0	b_rotR
b_single	PEOE_VSA.1	b_single
chi0v	PEOE_VSA.2	chi0v

chi0v_C	PEOE_VSA.3	chi0v_C
chi1v	PEOE_VSA.4	chi1v
chi1v_C	PEOE_VSA.5	chi1v_C
chiral	PEOE_VSA.0.1	chiral
rings	PEOE_VSA.1.1	reactive
Weight	PEOE_VSA.2.1	rings
a_heavy	PEOE_VSA.4.1	Weight
a_nC	PEOE_VSA.5.1	a_heavy
a_nF	PEOE_VSA.6.1	a_nC
a_nN	PEOE_VSA_FHYD	a_nN
a_nO	PEOE_VSA_FNEG	a_nO
a_nS	PEOE_VSA_FPNEG	b_heavy
b_heavy	PEOE_VSA_FPOL	chi0
chi0	PEOE_VSA_FPOS	chi0_C
chi0_C	PEOE_VSA_FPPOS	chi1
chi1	PEOE_VSA_HYD	chi1_C
chi1_C	PEOE_VSA_NEG	FCharge
FCharge	PEOE_VSA_PNEG	VAdjEq
VAdjEq	PEOE_VSA_POL	VAdjMa
VAdjMa	PEOE_VSA_POS	zagreb
zagreb	PEOE_VSA_PPOS	balabanJ
balabanJ	Q_VSA_HYD	PEOE_PC.
PEOE_PC.	Q_VSA_POS	PEOE_PC..1
PEOE_PC..1	Kier1	PEOE_RPC.
PEOE_RPC.	Kier2	PEOE_RPC..1
PEOE_RPC..1	Kier3	PEOE_VSA.0
PEOE_VSA.0	KierA1	PEOE_VSA.1
PEOE_VSA.1	KierA2	PEOE_VSA.2
PEOE_VSA.2	KierA3	PEOE_VSA.3
PEOE_VSA.3	KierFlex	PEOE_VSA.4
PEOE_VSA.4	apol	PEOE_VSA.5
PEOE_VSA.5	bpol	PEOE_VSA.6
PEOE_VSA.6	mr	PEOE_VSA.0.1
PEOE_VSA.0.1	a_acc	PEOE_VSA.1.1
PEOE_VSA.1.1	a_don	PEOE_VSA.5.1
PEOE_VSA.3.1	a_hyd	PEOE_VSA.6.1
PEOE_VSA.4.1	vsa_acc	PEOE_VSA_FHYD
PEOE_VSA.5.1	vsa_don	PEOE_VSA_FNEG
PEOE_VSA.6.1	vsa_hyd	PEOE_VSA_FPNEG
PEOE_VSA_FHYD	vsa_other	PEOE_VSA_FPOL
PEOE_VSA_FNEG	vsa_pol	PEOE_VSA_FPOS
PEOE_VSA_FPNEG	SlogP	PEOE_VSA_FPPOS
PEOE_VSA_FPOL	SlogP_VSA0	PEOE_VSA_HYD
PEOE_VSA_FPOS	SlogP_VSA1	PEOE_VSA_NEG
PEOE_VSA_FPPOS	SlogP_VSA2	PEOE_VSA_PNEG
PEOE_VSA_HYD	SlogP_VSA3	PEOE_VSA_POL
PEOE_VSA_NEG	SlogP_VSA4	PEOE_VSA_POS
PEOE_VSA_PNEG	SlogP_VSA5	PEOE_VSA_PPOS
PEOE_VSA_POL	SlogP_VSA7	PC.
PEOE_VSA_POS	SlogP_VSA8	PC..1
PEOE_VSA_PPOS	SlogP_VSA9	Q_PC.
PC.	SMR	Q_PC..1
PC..1	SMR_VSA0	Q_RPC.
Q_PC.	SMR_VSA1	Q_RPC..1
Q_PC..1	SMR_VSA2	Q_VSA_FHYD
Q_RPC.	SMR_VSA3	Q_VSA_FNEG
Q_RPC..1	SMR_VSA4	Q_VSA_FPNEG

Q_VSA_FHYD
Q_VSA_FNEG
Q_VSA_FPNEG
Q_VSA_FPOL
Q_VSA_FPOS
Q_VSA_FPPOS
Q_VSA_HYD
Q_VSA_NEG
Q_VSA_PNEG
Q_VSA_POL
Q_VSA_POS
Q_VSA_PPOS
RPC.
RPC..1
lip_acc
lip_don
lip_violation
opr_brigid
opr_nring
opr_nrot
opr_violation
Kier1
Kier2
Kier3
KierA1
KierA2
KierA3
KierFlex
logS
apol
bpol
mr
a_acc
a_base
a_don
a_hyd
vsa_acc
vsa_base
vsa_don
vsa_hyd
vsa_other
vsa_pol
SlogP
SlogP_VSA0
SlogP_VSA1
SlogP_VSA2
SlogP_VSA3
SlogP_VSA4
SlogP_VSA5
SlogP_VSA7
SlogP_VSA8
SlogP_VSA9
SMR
SMR_VSA0
SMR_VSA1
SMR_VSA2
SMR_VSA3
SMR_VSA5
SMR_VSA6
SMR_VSA7
TPSA
density
vdw_area
vdw_vol
logP.o.w.
Q_VSA_FPOL
Q_VSA_FPOS
Q_VSA_FPPOS
Q_VSA_HYD
Q_VSA_NEG
Q_VSA_PNEG
Q_VSA_POL
Q_VSA_POS
Q_VSA_PPOS
RPC.
RPC..1
lip_acc
lip_don
opr_brigid
opr_nring
opr_nrot
Kier1
Kier2
Kier3
KierA1
KierA2
KierA3
KierFlex
logS
apol
bpol
mr
a_acc
a_don
a_hyd
vsa_acc
vsa_don
vsa_hyd
vsa_other
vsa_pol
SlogP
SlogP_VSA0
SlogP_VSA1
SlogP_VSA2
SlogP_VSA3
SlogP_VSA4
SlogP_VSA5
SlogP_VSA7
SlogP_VSA8
SlogP_VSA9
SMR
SMR_VSA0
SMR_VSA1
SMR_VSA2
SMR_VSA3
SMR_VSA4
SMR_VSA5
SMR_VSA6
SMR_VSA7
TPSA
density
vdw_area

SMR_VSA4
SMR_VSA5
SMR_VSA6
SMR_VSA7
TPSA
density
vdw_area
vdw_vol
logP.o.w.

vdw_vol
logP.o.w.

Table 3. Biclustering parameters. Parameter fields left blank were not specified during biclustering of the dataset. Example command line format:

./gems32 mao_Leadscope_train_1 -r=1 -rN=3 -c=2 -cM=50 -w=0.5 -n=200 -m=r -a=maoi_Leadscope_train_1_weights -l=200 -tt -x=0 -mi -s=1 -v

Parameter	Definition	Factor Xa	MAO	mutagenicity
-m=R	mask the rows (compounds) already discovered during search			
-m=c	mask columns (features) in earlier biclusters			
-m=r	mask rows in earlier biclusters	-m=r	-m=r	-m=r
-mi	restore the masked rows in the output	-mi	-mi	-mi
-r	if used with -m=R and –mi: minimum number of previously uncovered rows to be included in a new bicluster	1	1	1
-rN	minimum rows covered after restoration of masked compounds	5	5	10
-c	minimum number of columns	2	2	2
-cM	maximum number of columns	50	50	50
-w	bin width or range of max and min real-value features in columns	0.5	0.5	0.5
-n	maximum number of biclusters discovered	200	200	200
-a	weights file specifiying weights of actives and inactives during discovery of pure biclusters	active bicluster discovery A =1	active bicluster discovery A =1	active bicluster discovery A =1

		I = -5 inactive bicluster discovery A = -10 I = 1	I = -5 inactive bicluster discovery A = -15 I = 1	I = -15 inactive bicluster discovery A = -15 I = 1
-tt	find constant columns, not constant rows	-tt	-tt	-tt
-u	find as few features as possible and do not expand the bicluster to include more columns			
-v	verbose mode	-v	-v	-v
-s=1	score the bicluster by the sum of weighting	-s=1	-s=1	-s=1
-x=0	define '0' as exclusion (feature absence)	-x=0	-x=0	-x=0
-l	speed up the Gibbs sampler by compromising the optimality	200	200	200

Table 4. Diversity analysis and biclustering coverage of compound datasets.

Dataset	DI entire dataset	RI between actives and inactives	DI actives	DI inactives	RI between training and testing	DI Training	DI testing	Active biclusters (% coverage)	Inactive biclusters (% coverage)
Factor Xa	0.215	0.185	0.236	0.251	0.215	0.212	0.219	5 (94%)	10 (70%)
MAO	0.079	0.081	0.104	0.077	0.079	0.079	0.079	5 (25%)	90(79%)
Mutagenicity	0.068	0.063	0.075	0.068	0.068	0.069	0.067	62 (55%)	28 (28%)

Table 5. Sample active biclusters for the factor Xa, MAO, and mutagenicity datasets. Two active biclusters are represented for each dataset. The representative compounds display structural features characterizing all the compounds included in each respective bicluster.

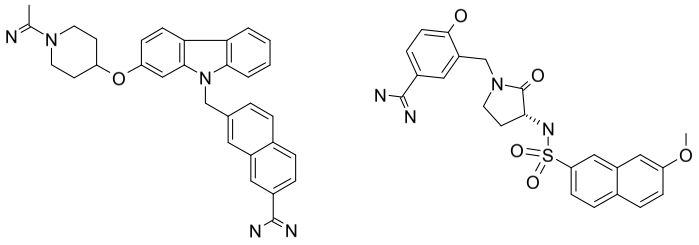
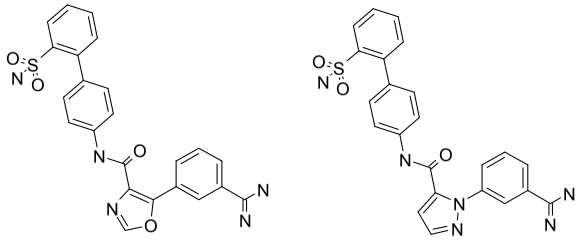
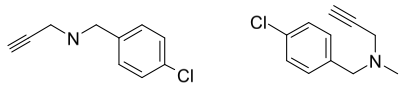
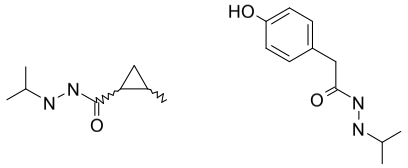
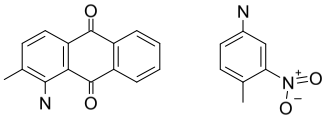
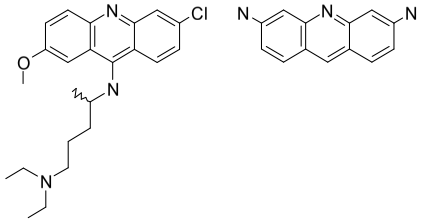
Dataset	Active bicluster 1	Active bicluster 2
Factor Xa		
MAO		
Mutagenicity		

Table 6. Leadscope features defining the sample active biclusters for the factor Xa, MAO, and mutagenicity datasets. Two active biclusters are represented for each dataset. These structural features characterize all the compounds included in each respective bicluster.

Dataset	Features
Factor Xa Active bicluster 1	hacceptor-path6-hacceptor amine, alkyl, cyc- hacceptor-path6-hdonor aromatic-path7-hacceptor iminomethyl, phenyl- benzene, 1-iminomethyl- amidine, phenyl-
Factor Xa Active bicluster 2	hacceptor-path6-hacceptor benzene, 1,2,4-acyc hacceptor-path6-hdonor hdonor-path6-pcharge hdonor-path6-hdonor aromatic-path8-hdonor aromatic-path8-hacceptor benzene, 1-hydroxy- alcohol, phenyl- alcohol, aryl- alcohol hacceptor-path6-pcharge aromatic-path6-hacceptor iminomethyl, phenyl- benzene, 1-iminomethyl- amidine, phenyl-
MAO Active bicluster 1	methane, 1-alkylamino-,1-aryl- benzene, 1-arylthio-,3-chloro- methane, 1-alkylamino-,1-phenyl- methane, 1-alkynyl-,1-amino- amine, benzyl- benzene, 1-aminomethyl(CH2)- alkyne, monosubst benzene, 1-alkylaminomethyl- methane, 1-amino-,1-aryl- methane, 1-alkylamino-,1-alkynyl- amine, propargyl- benzene, 1-aminomethyl- benzene, 1-(alkyl, acyc)- alkyne aromatic pcharge amine, alkyl, acyc- benzene, 1-halo- halide, phenyl- halide, aryl- halide

MAO Active bicluster 2	hydrazide, N-(i-propyl)- hydrazide, N-(alkyl, acyc)- hydrazide, N-(s-alkyl)- hydrazide, N-alkyl- hydrazine, i-propyl- hydrazine, alkyl, acyc- hydrazine, s-alkyl- hydrazine, alkyl- hydrazine(RNHNHR) hydrazine carbonyl, hydrazino- hydrazide hydrazine, N-carbonyl- hacceptor-path3-hdonor hacceptor carbonyl
Mutagenicity Active bicluster 1	aromatic hacceptor-path5-pcharge hacceptor benzene, 1-methyl- benzene, 1-(alkyl, acyc)-
Mutagenicity Active bicluster 2	aromatic pyridine, 3-fused ring- pyridine, 2-fused ring- quinoline quinoline, 3-fused ring- quinoline, 2-fused ring- hacceptor-path8-pcharge pyridine hacceptor hacceptor-path4-pcharge hacceptor-path8-hdonor hdonor hacceptor-path4-hdonor

Table 7. PLS parameters and RMSE for the three datasets using NIPALS regression models on various descriptor combinations. Results are averaged over five random validations and standard deviations are provided.

Dataset	Descriptor	#PC	Activity threshold	RMSE
Factor Xa	MOE	52.8 ± 22.219	0.39 ± 0.042	0.336 ± 0.070
	Leadscope	17.6 ± 3.782	0.33 ± 0.045	0.221 ± 0.023
	MOE & Leadscope	14.8 ± 3.033	0.32 ± 0.027	0.215 ± 0.013
	Biclusters	16.4 ± 6.580	0.39 ± 0.065	0.241 ± 0.027
	MOE & biclusters	18.0 ± 2.646	0.36 ± 0.042	0.236 ± 0.018
	MOE & ensembles	34.0 ± 15.379	0.44 ± 0.022	0.270 ± 0.045
MAO	MOE	95.6 ± 6.986	0.34 ± 0.022	0.387 ± 0.008
	Leadscope	92.0 ± 9.798	0.41 ± 0.055	0.743 ± 0.042
	MOE & Leadscope	93.0 ± 5.568	0.44 ± 0.022	0.741 ± 0.050
	biclusters	30.0 ± 6.083	0.32 ± 0.027	0.366 ± 0.015
	MOE & biclusters	75.6 ± 9.290	0.38 ± 0.045	0.398 ± 0.017
	MOE & ensembles	60.2 ± 28.385	0.32 ± 0.027	0.352 ± 0.011
Muta-genicity	MOE	91.4 ± 10.644	0.44 ± 0.022	0.406 ± 0.004
	Leadscope	93.2 ± 9.311	0.48 ± 0.027	0.598 ± 0.039
	MOE & Leadscope	93.8 ± 5.63	0.50 ± 0.000	0.562 ± 0.016
	biclusters	54.2 ± 16.037	0.44 ± 0.022	0.399 ± 0.004
	MOE & biclusters	85.4 ± 17.358	0.45 ± 0.000	0.405 ± 0.013
	MOE & ensembles	71.2 ± 17.441	0.46 ± 0.022	0.374 ± 0.003

Table 8. Classification results for the factor Xa dataset using NIPALS regression models on various descriptor combinations. Results are averaged over five random validations. Standard deviations are provided.

Descriptor	Data subset	Accuracy	Sensitivity	Specificity	Precision	F-measure
MOE	training	0.994 ± 0.002	0.999 ± 0.003	0.985 ± 0.009	0.992 ± 0.005	0.995 ± 0.002
	testing	0.935 ± 0.021	0.973 ± 0.032	0.866 ± 0.070	0.930 ± 0.033	0.950 ± 0.015
Leadscope	training	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	testing	0.937 ± 0.034	0.988 ± 0.013	0.845 ± 0.103	0.922 ± 0.048	0.953 ± 0.024
MOE & Leadscope	training	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	testing	0.951 ± 0.019	0.988 ± 0.008	0.883 ± 0.061	0.939 ± 0.030	0.963 ± 0.014
Biclusters	training	0.993 ± 0.005	0.990 ± 0.007	0.998 ± 0.004	0.999 ± 0.003	0.994 ± 0.004
	testing	0.955 ± 0.020	0.979 ± 0.012	0.914 ± 0.076	0.955 ± 0.038	0.966 ± 0.014
MOE & Biclusters	training	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	testing	0.955 ± 0.019	0.988 ± 0.004	0.896 ± 0.052	0.946 ± 0.026	0.966 ± 0.014
MOE & Ensembles	training	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
	testing	0.957 ± 0.008	0.971 ± 0.014	0.931 ± 0.037	0.962 ± 0.019	0.967 ± 0.005

Table 9. Classification results for the MAO dataset using NIPALS regression models on various descriptor combinations. Results are averaged over five random validations. Standard deviations are provided.

Descriptor	Data subset	Accuracy	Sensitivity	Specificity	Precision	F-measure
MOE	training	0.848 ± 0.015	0.532 ± 0.050	0.915 ± 0.028	0.577 ± 0.050	0.550 ± 0.013
	testing	0.791 ± 0.032	0.429 ± 0.073	0.868 ± 0.048	0.418 ± 0.062	0.419 ± 0.043
Leadscope	training	0.961 ± 0.006	0.882 ± 0.040	0.977 ± 0.013	0.896 ± 0.050	0.887 ± 0.014
	testing	0.736 ± 0.024	0.556 ± 0.058	0.774 ± 0.039	0.346 ± 0.023	0.425 ± 0.018
MOE &	training	0.976 ± 0.003	0.924 ± 0.009	0.987 ± 0.005	0.938 ± 0.025	0.931 ± 0.009
Leadscope	testing	0.731 ± 0.019	0.575 ± 0.035	0.764 ± 0.022	0.342 ± 0.025	0.429 ± 0.027
Biclusters	training	0.883 ± 0.008	0.726 ± 0.052	0.917 ± 0.020	0.653 ± 0.042	0.685 ± 0.005
	testing	**0.793 ± 0.023**	**0.547 ± 0.033**	**0.845 ± 0.034**	**0.434 ± 0.044**	**0.482 ± 0.021**
MOE &	training	0.920 ± 0.004	0.764 ± 0.046	0.953 ± 0.015	0.778 ± 0.047	0.769 ± 0.004
Biclusters	testing	**0.807 ± 0.015**	**0.547 ± 0.041**	**0.862 ± 0.026**	**0.461 ± 0.029**	**0.498 ± 0.010**
MOE &	training	0.863 ± 0.016	0.705 ± 0.049	0.896 ± 0.027	0.595 ± 0.050	0.642 ± 0.018
Ensembles	testing	0.765 ± 0.025	0.472 ± 0.061	0.827 ± 0.040	0.372 ± 0.035	0.414 ± 0.029

Table 10. Classification results for the mutagenicity dataset using NIPALS regression models on various descriptor combinations. Results are averaged over five random validations. Standard deviations are provided.

Descriptor	Data subset	Accuracy	Sensitivity	Specificity	Precision	F-measure
MOE	training	0.804 ± 0.007	0.889 ± 0.017	0.698 ± 0.036	0.786 ± 0.016	0.835 ± 0.003
	testing	0.771 ± 0.008	0.869 ± 0.015	0.648 ± 0.027	0.755 ± 0.012	0.808 ± 0.006
Leadscope	training	0.948 ± 0.003	0.964 ± 0.008	0.928 ± 0.010	0.944 ± 0.007	0.953 ± 0.003
	testing	0.785 ± 0.017	0.819 ± 0.027	0.741 ± 0.011	0.797 ± 0.010	0.808 ± 0.017
MOE & Leadscope	training	0.959 ± 0.001	0.966 ± 0.003	0.951 ± 0.002	0.961 ± 0.002	0.963 ± 0.001
	testing	0.791 ± 0.014	0.818 ± 0.017	0.758 ± 0.012	0.808 ± 0.010	0.813 ± 0.013
Biclusters	training	0.831 ± 0.010	0.880 ± 0.014	0.770 ± 0.041	0.827 ± 0.022	0.853 ± 0.006
	testing	0.799 ± 0.013	**0.862 ± 0.020**	0.721 ± 0.033	**0.794 ± 0.017**	**0.826 ± 0.011**
MOE & Biclusters	training	0.867 ± 0.005	0.916 ± 0.004	0.807 ± 0.010	0.855 ± 0.007	0.885 ± 0.004
	testing	0.809 ± 0.016	**0.876 ± 0.009**	0.727 ± 0.025	**0.800 ± 0.016**	**0.836 ± 0.013**
MOE & Ensembles	training	0.852 ± 0.010	0.892 ± 0.016	0.800 ± 0.037	0.849 ± 0.022	0.870 ± 0.006
	testing	0.795 ± 0.010	0.847 ± 0.030	0.731 ± 0.031	0.797 ± 0.014	0.821 ± 0.011

Bibliography

1. Bender, A.; Glen, R. C., Molecular Similarity: A Key Technique in Molecular Informatics. *Organic and Biomolecular Chemistry* 2004, 2, 3204-3218.

2. Johnson, M. A.; Maggiora, G. M., *Concepts and Applications of Molecular Similarity*. John Wiley & Sons: New York, 1990.

3. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K., Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences* 2004, 44, (6), 1912-28.

4. Willett, P. B., J. M.; Downs, G. M. , Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* 1998, 38, 983-996.

5. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S., *Introduction to multi- and megaVariate data analysis using projection methods (PCA & PLS)*. Umetrics: Umeå, 1999.

6. Geladi, P.; Kowalski, B., Partial Least-Squares Regression: A Tutorial. *Analytica Chimca Acta* 1986, 185, 1-17.

7. Noonan, R. D.; Wold, H., *Educational Research, Methodology, and Measurement: An International Handbook*. Pergamon Press: Oxford, 1988; p 710-716.

8. Hirst, J. D., Nonlinear quantitative structure-activity relationship for the inhibition of dihydrofolate reductase by pyrimidines. *Journal of Medicinal Chemistry* 1996, 39, (18), 3526-32.

9. DeLisle, R. K.; Dixon, S. L., Induction of decision trees via evolutionary programming. *Journal of Chemical Information and Computer Sciences* 2004, 44, (3), 862-70.

10. Rusinko, A., 3rd; Young, S. S.; Drewry, D. H.; Gerritz, S. W., Optimization of focused chemical libraries using recursive partitioning. *Combinatorial Chemistry and High Throughput Screening* 2002, 5, (2), 125-33.

11. Wang, X. Z.; Buontempo, F. V.; Young, A.; Osborn, D., Induction of decision trees using genetic programming for modelling ecotoxicity data: adaptive discretization of real-valued endpoints. *SAR and QSAR in Environmental Research* 2006, 17, (5), 451-71.

12. A-Razzak, M.; Glen, R. C., Applications of rule-induction in the derivation of quantitative structure-activity relationships. *Journal of Computer-Aided Molecular Design* 1992, 6, (4), 349-83.

13. Schneider, G.; Wrede, P., Artificial neural networks for computer-based molecular design. *Progress in Biophysics and Molecular Biology* 1998, 70, (3), 175-222.

14. Byvatov, E.; Schneider, G., SVM-based feature selection for characterization of focused compound collections. *Journal of Chemical Information and Computer Sciences* 2004, 44, (3), 993-9.

15. Schneider, G.; Nettekoven, M., Ligand-based combinatorial design of selective purinergic receptor (A2A) antagonists using self-organizing maps. *Journal of Combinatorial Chemistry* 2003, 5, (3), 233-7.

16. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S., Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *Journal of Chemical Information and Computer Sciences* 2004, 44, (1), 170-8.

17. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S., Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of Chemical Information and Computer Sciences* 2004, 44, (5), 1708-18.

18. Gao, H.; Lajiness, M. S.; Van Drie, J., Enhancement of binary QSAR analysis by a GA-based variable selection method. *Journal of Molecular Graphics and Modelling* 2002, 20, (4), 259-68.

19. Kubinyi, H., Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quantitative Structure-Activity Relationships* 1994, 13, 285-294.

20. Lu, J. X.; Shen, Q.; Jiang, J. H.; Shen, G. L.; Yu, R. Q., QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. *Journal of Pharmaceutical and Biomedical Analysis* 2004, 35, (4), 679-87.

21. Sutter, J. M.; Jurs, P. C., Automated Descriptor Selection for Quantitative Structure-Activity Relationships Using Generalized Simulated Annealing. *Journal of Chemical Information and Computer Sciences* 1995, 35, 77-84.

22. Zheng, W.; Tropsha, A., Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *Journal of Chemical Information and Computer Sciences* 2000, 40, (1), 185-94.

23. Izrailev, S.; Agrafiotis, D., A novel method for building regression tree models for QSAR based on artificial ant colony systems. *Journal of Chemical Information and Computer Sciences* 2001, 41, (1), 176-80.

24. Izrailev, S.; Agrafiotis, D. K., Variable selection for QSAR by artificial ant colony systems. *SAR and QSAR in Environmental Research* 2002, 13, (3-4), 417-23.

25. Arodz, T.; Yuen, D. A.; Dudek, A. Z., Ensemble of linear models for predicting drug properties. *Journal of Chemical Information and Modeling* 2006, 46, (1), 416-23.

26. Williams, D. H.; Stephens, E.; O'Brien, D. P.; Zhou, M., Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angewandte Chemie. International Edition in English* 2004, 43, (48), 6596-616.

27. Cho, S. J.; Hermsmeier, M. A., Genetic Algorithm guided Selection: variable selection and subset selection. *Journal of Chemical Information and Computer Sciences* 2002, 42, (4), 927-36.

28. Engels, M. F.; Thielemans, T.; Verbinnen, D.; Tollenaere, J. P.; Verbeeck, R., CerBeruS: a system supporting the sequential screening process. *Journal of Chemical Information and Computer Sciences* 2000, 40, (2), 241-5.

29. Mattioni, B. E.; Kauffman, G. W.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M., Predicting the genotoxicity of secondary and aromatic amines using data subsetting to generate a model ensemble. *Journal of Chemical Information and Computer Sciences* 2003, 43, (3), 949-63.

30. Wild, D. J.; Blankley, C. J., Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *Journal of Chemical Information and Computer Sciences* 2000, 40, (1), 155-62.

31. Dhillon, I. S.; Mallela, S.; Modha, D. S., Information-Theoretic Co-Clustering. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2003, 89-98.

32. Hartigan, J. A., Direct clustering of a data matrix. *Journal of the American Statistical Association* March 1972, 67, (337), 123-129.

33. Madeira, S. C.; Oliveira, A. L., Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2004, 1, (1), 24-45.

34. Cheng, Y.; Church, G. M., Biclustering of expression data. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* 2000, 8, 93-103.

35. GEMS: Gene Expression Mining Server. http://genomics10.bu.edu/terrence/gems/.

36. Wu, C. J.; Kasif, S., GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Research* 2005, 33, (Web Server issue), W596-9.

37. Wu, J.; Kasif, S.; DeLisi, C., Identification of functional links between genes using phylogenetic profiles. *Bioinformatics* 2003, 19, (12), 1524-30.

38. Tarasov, V. A.; Mustafaev, O. N.; Abilev, S. K.; Mel'nik, V. A., [Use of ensemble structural descriptors for increasing the efficiency of QSAR study]. *Genetika* 2005, 41, (7), 997-1005.

39. Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F., Anchor-GRIND: filling the gap between standard 3D QSAR and the GRid-INdependent descriptors. *Journal of Medicinal Chemistry* 2005, 48, (7), 2687-94.

40. Brown, R. D.; Martin, Y. C., Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *Journal of Chemical Information and Computer Sciences* 1996, 36, (3), 572 - 584.

41. Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V.; Leach, A. R., Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences* 2001, 41, (5), 1295-300.

42. Martin, Y. C.; Kofron, J. L.; Traphagen, L. M., Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry* 2002, 45, (19), 4350-8.

43. Kazius, J.; McGuire, R.; Bursi, R., Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry* 2005, 48, (1), 312-20.

44. Chemical Computing Group Inc. MOE 2004.03, 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.

45. Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E., Jr., LeadScope: software for exploring large sets of screening data. *Journal of Chemical Information and Computer Sciences* 2000, 40, (6), 1302-14.

46. Wehrens, R.; Mevik, B.-H., The pls package. Reference manual available at http://cran.r-project.org/.

47. The Comprehensive R Archive Network, http://cran.r-project.org/.

48. Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G., ChemSpaceShuttle: A tool for data mining in drug discovery by classification, projection, and 3D visualization. *QSAR and Combinatorial Science* 2003, 5, 549-559.

49. Helland, I. S., On the structure of partial least squares. *Communications in Statistics – Simulation* 1998, 17, 1581-607.

50. Yap, C. W.; Chen, Y. Z., Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *Journal of Chemical Information and Modeling* 2005, 45, (4), 982-92.

51. Hawkins, D. M.; Basak, S. C.; Mills, D., Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences* 2003, 43, (2), 579-86.

52. Rajer-Kanduc, K.; Zupan, J.; Majcen, N., Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemometrics and Intelligent Laboratory Systems* 2003, 65, (9), 221-229.

53. Schultz, T. W.; Netzeva, T. I.; Cronin, M. T., Selection of data sets for QSARs: analyses of Tetrahymena toxicity from aromatic compounds. *SAR and QSAR in Environmental Research* 2003, 14, (1), 59-81.

54. Tolstoy, L., In *Anna Karenina*, 1917 ed.; The Harvard Classics Shelf of Fiction: pp part 1, chapter 1.

55. Pauls, H. W.; Ewing, W. R., The design of competitive, small-molecule inhibitors of coagulation factor Xa. *Current Topics in Medicinal Chemistry* 2001, 1, (2), 83-100.

56. Quan, M. L.; Wexler, R. R., The design and synthesis of noncovalent factor Xa inhibitors. *Current Topics in Medicinal Chemistry* 2001, 1, (2), 137-49.

57. Strolin Benedetti, M.; Dostert, P., Overview of the present state of MAO inhibitors. *Journal of Neural Transmission. Supplementum* 1987, 23, 103-19.

58. Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q., Boosting: an ensemble learning tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling* 2005, 45, (3), 786-99.