Comparative Analysis of Machine Learning Methods in Ligand-Based Virtual Screening of Large Compound Libraries

Xiao H. Ma[1], Jia Jia[1], Feng Zhu[1], Ying Xue[1,2], Ze R. Li[1,2] and Yu Z. Chen[*,1]

[1]*Bioinformatics and Drug Design Group, Department of Pharmacy and Center of Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, 117543, Singapore*

[2]*College of Chemistry, Sichuan University, Chengdu, 610064, P.R. China*

Abstract: Machine learning methods have been explored as ligand-based virtual screening tools for facilitating drug lead discovery. These methods predict compounds of specific pharmacodynamic, pharmacokinetic or toxicological properties based on their structure-derived structural and physicochemical properties. Increasing attention has been directed at these methods because of their capability in predicting compounds of diverse structures and complex structure-activity relationships without requiring the knowledge of target 3D structure. This article reviews current progresses in using machine learning methods for virtual screening of pharmacodynamically active compounds from large compound libraries, and analyzes and compares the reported performances of machine learning tools with those of structure-based and other ligand-based (such as pharmacophore and clustering) virtual screening methods. The feasibility to improve the performance of machine learning methods in screening large libraries is discussed.

Keywords: Activator, adverse drug reaction, agonist, antagonist, compound, computer aided dug design, drug, drug discovery, inhibitor, molecule, pharmacodynamics, pharmacokinetics, statistical learning methods, toxicity, toxicology, virtual screening.

INTRODUCTION

Virtual screening (VS) has been extensively explored for facilitating drug lead discovery [1-4] and for identifying agents of desirable pharmacokinetic and toxicological properties [5, 6]. Both structure-based [1,7-18] and ligand-based [2,19-22] approaches have been used for developing VS tools. The former identifies active compounds by modeling and evaluating their interaction with a specific target based on binding site 3D structure. The latter searches active compounds by determining if their structure contains the framework or profile that matches those of the known active compounds. Machine learning (ML) methods have recently been explored for developing ligand-based VS tools [23-34] to complement or to be used in combination with structure-based [1,7-18] and other ligand-based [2,19-22] VS tools for improving the performance and speed of lead discovery.

ML methods utilize nonlinear supervised learning methods to develop statistical models that map physicochemical properties (molecular descriptors) with their activity classes, and thus are more capable of predicting a more diverse spectrum of compounds and more complex structure-activity relationships than structure-based VS methods and other ligand-based VS methods such as QSAR, pharmacophore, and clustering methods [2, 19-23, 32, 33]. This capability arises because ML methods are capable of generating complex nonlinear mappings from molecular descriptors to activity classes without restriction on structural frameworks, and without requiring prior knowledge of relevant molecular descriptors and functional form of structure-activity relationships [29, 35-38]. Moreover, performance of ML can be enhanced by active learning that mimics drug discovery cycle [39, 40] and by applicability domain analysis that focuses on compounds of highest chance of correct prediction [35, 41-43].

Therefore, ML methods have been explored as part of the efforts to overcome several problems that have impeded progress in the application of structure-based VS and other ligand-based VS tools [1, 36]. These problems include the vastness and sparse nature of chemical space needs to be searched, limited availability of target structures (only 15% of known proteins have known 3D structures), limited diversity biased by training molecules, complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects.

This article reviews current progress in using ML methods for virtual screening of pharmacodynamically active compounds from large compound libraries. The literature-reported performances of ML-based VS tools are analyzed and tentatively compared with those of structure-based and other ligand-based VS tools. The feasibility to improve the performance of ML methods in screening large libraries is also discussed. Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS algorithms and their parameters used. The reported studies are, in most cases, not detailed enough to address such questions as which hits are predicted by all or majority of the methods. The screened library in these studies is either part of MDDR (up to ~170,000 compounds) or specially assembled VS libraries from 15 commercially and freely available libraries (~2 million compounds) [44] plus in-house libraries of various sizes. The number of unique

*Address correspondence to this author at the Bioinformatics and Drug Design Group, Department of Pharmacy and Center of Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, 117543, Singapore; Tel: 65-6874-6877; Fax: 65-6774-6756; E-mail: phacyz@nus.edu.sg

1386-2073/09 $55.00+.00

compounds and scaffolds are ~160,000 and 3057 for the former and ≥729,000 and ≥18,775 for the later [45]. Therefore, there seems to be a correlation between the size and diversity of the VS libraries, and the structural diversity of the libraries used in most studies is sufficiently broad to provide a tentative comparison for crudely estimating the level of performance of ML methods with respect to those achieved by other types of VS tools.

COMPOUND REPRESENTATION IN MACHINE LEARNING METHODS

Structural and physicochemical features of compounds can be quantitatively represented by molecular descriptors, which have been extensively used for developing ML VS tools [46-53] as well as SAR [54, 55] and QSAR [56, 57]. The most popularly used computer programs for deriving molecular descriptors are DRAGON [58], Molconn-Z [59], JOELib [60], and Xue descriptor set [50]. Web-servers such as MODEL [61] (http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi) have also emerged. These methods can be used for deriving >3,000 molecular descriptors including constitutional descriptors, topological descriptors [62], RDF descriptors [63], molecular walk counts [64], 3D-MoRSE descriptors [65], BCUT descriptors [66], WHIM descriptors [67], Galvez topological charge indices and charge descriptors [68], GETAWAY descriptors [69], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices [70], Randic molecular profiles [71], electrotopological state descriptors [72], linear solvation energy relationship descriptors [73], and other empirical and molecular properties.

Not all of the available descriptors are needed for representing features of a particular class of compounds. Moreover, without properly selecting the appropriate set of descriptors, the performance of a developed ML VS tool may be affected to some degrees because of the noise arising from the high redundancy and overlapping of the available descriptors. Selection of appropriate set of molecular descriptors is also important for distinguishing similar and dissimilar compounds in the same and different activity class [74]. Testing molecules that are similar to a lead are highly useful when optimizing the lead but may become wasteful in searching for new leads against the known lead. It has been found that, if a molecular descriptor is to be a valid and useful measure of similarity in drug discovery, a plot of differences in its values *vs* differences in biological activities exhibits a characteristic trapezoidal distribution enhancement that reveals a neighborhood behavior for the descriptor [75].

Descriptors most appropriate for representing compounds of a particular property can be selected either by intuition as those used in QSAR and QSPR studies [23, 32, 33], or by using such feature selection methods as genetic algorithm-based approach [76], recursive feature eliminations (RFE) [77], and simulated annealing-based approach [78]. However, in many cases, it is difficult to uniquely select an optimal set of descriptors due to the high redundancy and overlapping of many descriptors [79]. Separate sets of descriptors containing different members of redundant descriptor classes have been found to give similar prediction accuracies [80]. Thus, interpretation of prediction results may need to be conducted at the descriptor class level where redundant and overlapping descriptors are grouped into classes [81].

It has been found that the physicochemical features of some compounds cannot be fully represented by the available descriptors [81-83]. These compounds typically contain one or more combinations of inflexible multi-rings, highly polar tetrazole rings, aromatic rings separated by a specific atom, complex two ring system with multiple heteroatoms, polycyclic aromatic structures, long flexible chains, hydrazine group, and multiple ionisable groups. Therefore, there is a need for deriving new descriptors to adequately represent features of these and other compounds.

COMMONLY USED MACHINE LEARNING METHODS

Several ML methods have been used for developing VS tools. These include logistic regression (LR), linear discriminant analysis (LDA), *k*-nearest neighbor (*k*NN), binary kernel discrimination (BKD), decision tree (DT), naive Bayesian classifier (NBC), artificial neural networks (ANN), probabilistic neural network (PNN), and support vector machines (SVM). Websites for the freely downloadable codes of some methods are given in Table **1**.

Artificial Neural Network (ANN)

An artificial neural network (ANN) is an information-processing paradigm that mimics the information-processing operations of the densely interconnected neurons [84, 85]. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network governed by simple mathematical models defined by a function $f(x)$: $x \rightarrow y$. The term artificial neural network arises because the function $f(x)$ is defined as a composition of other functions $g_i(x)$, which can further be defined as a composition of other functions. This can be conveniently represented as a network structure, with arrows depicting the dependencies between variables. A widely used type of composition is the nonlinear weighted sum $f(x) = K(\sum_i w_i g_i(x))$, where K is a predefined function.

Binary Kernel Discrimination (BKD)

BKD is a ML method initially used in chemoinformatics [24] and subsequently applied to the prediction of active compounds [30, 86]. BKD is a similarity-based approach for classifying chemical properties *via* kernel functions. The commonly used kernel function is $K_\lambda(i,j) = [\lambda^{n-d_{ij}}(1-\lambda)^{d_{ij}}]^{\beta/n}$ where λ is a smoothing parameter to be determined in specific cases, n is the length of the binary feature vectors with molecular descriptors as their components, d_{ij} is the Hamming distance between the feature vectors for compounds i and j, and β (β ≤ n) is a user-defined constant. Training set compounds are ranked using the scoring function $S_{BKD}(j) = \left(\sum_{i=active} k_\lambda(i,j)\right) / \left(\sum_{i=inactive} k_\lambda(i,j)\right)$ with the optimum value of λ being found from analysis of the training set. The optimum is obtained by computing scores for each training set compound based on the other training set compounds, for a number of different values of λ in the

Table 1. Websites that Contain Freely Downloadable Codes of Machine Learning Methods

BKD	
Binding Database	http://www.bindingdb.org/bind/vsOverview.jsp
Decision Trees	
PrecisionTree	http://www.palisade.com.au/precisiontree/
DecisionPro	http://www.vanguardsw.com/decisionpro/jdtree.htm
C4.5	http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
C5.0	http://www.rulequest.com/download.html
***k*NN**	
k-Nearest Neighbor	http://www.cs.cmu.edu/~zhuxj/courseproject/knndemo/KNN.html
PERL Module for *k*NN	http://aspn.activestate.com/ASPN/CodeDoc/AI-Categorize/AI/Categorize/kNN.html
Java class for *k*NN	http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/classify/old/KNN.html
LDA	
DTREG	http://www.dtreg.com/lda.htm
LR	
Paul Komarek's Logistic Regression Software	http://komarix.org/ac/lr/lrtrirls
Web-based logistic regression calculator	http://statpages.org/logistic.html
Neural Network	
BrainMaker	http://www.calsci.com/
Libneural	http://pcrochat.online.fr/webus/tutorial/BPN_tutorial7.html
fann	http://leenissen.dk/fann/
NeuralWorks Predict	http://www.neuralware.com/products.jsp
NeuroShell Predictor	http://www.mbaware.com/neurpred.html
SVM	
SVM light	http://svmlight.joachims.org/
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
mySVM	http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html
BSVM	http://www.csie.ntu.edu.tw/~cjlin/bsvm/
SVMTorch	http://www.idiap.ch/learning/SVMTorch.html
Multiple ML Tools	
Weka (Java package that include decision trees, linear regression, and SVM)	http://www.cs.waikato.ac.nz/ml/weka/

range 0.50-0.99. For each value of λ the sum of the ranks of the active compounds is computed. If this is plotted against λ a clear minimum should be observed indicating the optimum λ, *i.e.*, the value that minimizes the summed ranks of the actives in the training set. This optimum value is then used for scoring the compounds in the test set.

C4.5 Decision Tree (C4.5 DT)

C4.5 DT is a branch-test-based classifier [87]. A branch of the decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 decision tree uses recursive partitioning to examine every attribute of the data and to rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector $\boldsymbol{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector $\boldsymbol{x}$ is predicted to be that of the leaf. This process continues to allow the tree to grow to the full size, which is then pruned back to an appropriate size based on the evaluation of its overall prediction performance.

The estimation criterion in the decision tree algorithm is the selection of an attribute to test at each decision node in the tree. The goal is to select the attribute that is most useful for classifying examples. A good quantitative measure of the worth of an attribute is a statistical property called informa-

tion gain $Gain(S,A) = Entropy(s) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$ that measures how well a given attribute separates the training examples according to their target classification. Here $Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$, S is called entropy that characterizes the purity or impurity of an arbitrary collection of examples, p_i is the proportion of S belonging to class I, A is the set of all possible values for attribute A, and S_v is the subset of S for which attribute A has value v (*i.e.*, $S_v = \{s \hat{I} S \mid A(s) = v\}$).

k Nearest Neighbor (kNN)

In *k*NN, the Euclidean distance between an unclassified vector $\boldsymbol{x}$ and each individual vector $\boldsymbol{x}_i$ in the training set is measured [88, 89]. A total of k number of vectors nearest to the unclassified vector $\boldsymbol{x}$ are used to determine the class of that unclassified vector. The class of the majority of the k nearest neighbors is chosen as the predicted class of the unclassified vector $\boldsymbol{x}$.

Linear Discriminant Analysis (LDA)

LDA [90] separates two classes of vectors by constructing a hyperplane defined by the following linear discriminant function: $L = \sum_{i}^{k} w_i x_i$, where L is the resultant classification score and w_i is the weight associated with the corresponding descriptor x_i. A positive or negative L value indicates that a vector $\boldsymbol{x}$ belongs to the positive or negative class respectively. LDA has been extensive explored for developing QSAR models of active compounds [91-94].

Naive Bayesian Classifier (NBC)

A NBC [95] is a simple probabilistic classifier that applies the Bayes theorem under the strong (naive) independence assumptions. The task is to determine $P(Y|\mathbf{x})$, the posterior probability of Y conditioned on the vectors $\{\mathbf{x}\}$ based on more information (such as background knowledge) than the prior probability $P(Y)$ un-conditioned on $\mathbf{x}$. Similarly, $P(\mathbf{x}|Y)$ is posterior probability of $\mathbf{x}$ conditioned on Y. The Bayes theorem gives $P(Y|\mathbf{x}) = P(\mathbf{x}|Y)P(Y)/P(\mathbf{x})$, which is useful for calculating the posterior probability $P(Y|\mathbf{x})$ from $P(Y)$, $P(\mathbf{x})$, and $P(\mathbf{x}|Y)$.

Logistic Regression (LR)

LR [96] is based on the assumption that a logistic relationship exists between the probability of class membership and one or more descriptors. The probability $Y = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots + \beta k Xk)}}$, where $\mathbf{X}$={$X1$, …, Xk} is a feature vector, Xk mathematical notation is a descriptor, $\beta 0$ is the regression model constant, $\beta 1$ to βk is the coefficients corresponding to the descriptors $X1$ to Xk. $Y > 0.5$ or $Y < 0.5$ indicates that the vector $\mathbf{X}$ belongs to the positive or negative class respectively.

Probabilistic Neural Network (PNN)

PNN is a form of neural network that uses Bayes optimal decision rule $h_i c_i f_i(x) > h_j c_j f_j(x)$ for classification [97], Here, h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively. An unknown vector $\boldsymbol{x}$ is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a multivariate case can be estimated by using the Parzen's nonparametric estimator [98] $g(x) = \frac{1}{n} \sum_{i=1}^{n} \exp(-\sum_{j=1}^{p} \left[\frac{x_j - x_{ij}}{\sigma_j} \right]^2)$. Traditional neural networks such as feed-forward back-propagation neural network rely on multiple parameters and network architectures to be optimized. In contrast, PNN only has a single adjustable parameter, a smoothing factor σ for the radial basis function in the Parzen's nonparameteric estimator. Thus the training process of PNN is usually orders of magnitude faster than those of the traditional neural networks.

Support Vector Machine (SVM)

SVM is a supervised ML method based on the structural risk minimization principle for minimizing both training and generalization error [99]. There are linear and nonlinear SVMs with the later more extensively used in practical applications [100]. Details of SVM classification algorithms and their applications in chemistry can be found in the literature [100, 101]. Linear SVM constructs a hyperplane separating two different classes of feature vectors with a maximum margin. For linear SVM with classification errors, there are slack variables and the conditions are different from the ones listed here. Nonlinear SVM projects input vectors into a high dimensional feature space by using a kernel function such as $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left\|\mathbf{x}_j - \mathbf{x}_i\right\|^2 / 2\sigma^2}$. The linear SVM procedure is then applied to the feature vectors in this feature space. A given vector $\mathbf{x}$ can be classified by using $class(x_k) = sign(\sum_{i=1}^{m} \lambda_i y_i x_i \cdot x_k + b)$, a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class respectively.

VIRTUAL SCREENING PERFORMANCE MEASUREMENT

The performance of ML methods can be evaluated by the quantity of true positives *TP* (number of class +1 compounds classified as +1), true negatives *TN* (number of class -1 compounds classified as -1), false positives *FP* (number of class -1 compounds classified as +1), and false negatives *FN* (number of class +1 compounds classified as -1), sensitivity *SE=TP/(TP+FN*) (accuracy for class +1 compounds), specificity *SP=TN/(TN+FP)* (accuracy for class -1 compounds),

overall accuracy $Q = \frac{TP + TN}{TP + TN + FP + FN}$, Matthews correlation coefficient $C = \frac{TP * TN - FN * FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$, and AUROC (area under the ROC curve) respectively. The frequently used measures of VS performance are yield (ratio of predicted known hits and all known hits in the screened libraries), hit rate (fraction of known hits in the predicted hits) and enrichment factor (magnitude of hit-rate improvement over random selection) [1, 2, 7, 19, 24-26], which are expressed as *SE*, *TP/(TP+FP)* and *TP(TP+FP+TN+FN)/(TP+FP)(TP+FN)* respectively.

ASSESSMENT OF REPORTED VIRTUAL SCREENING PERFORMANCE OF MACHINE LEARNING METHODS

The reported performance of ML methods in screening pharmacodynamically active compounds from libraries of >25,000 compounds is summarized in Table **2**. The screening tasks of these reported studies are primarily focused on the prediction of compounds that inhibit, antagonize, block, agonize, or activate specific therapeutic target protein [102]. Examples of these tasks are the prediction of COX2 inhibitors for anti-inflammation [103] and anti-arthritis [104], thrombin inhibitors for anticoagulation [105] and the treatment of other diseases [106], chemokine receptor antagonists for pathophysiology and autoinflammatory diseases [107], HIV-1 protease inhibitors for anti-HIV [108], selective protein kinase inhibitors for modulating cancer and cardiovascular diseases [109, 110], PDE5 inhibitors for erectile dysfunction and other disorders [111], and classes of antidepressant agents that modulate neurotransmission (such as 5HT reuptake inhibitors, MAO inhibitors) [112].

For tentative comparison, the reported performances of structure-based VS methods and two classes of ligand-based VS methods, pharmacophore and clustering, are summarized in Tables **3-5** respectively. ML methods have been found to show generally good performances. In the majority of the reported studies, the yields, hit rates, and enrichment factors of ML methods are in the range of 50%~94%, 10%~98%, and 30~108 respectively. In comparison, the yields, hit rates, and enrichment factors of the majority of the reported studies by other methods shown in Table **3**, Tables **4** and **5** are in the range of 7%~95%, 1%~32%, and 5~1189 for structure-based, 11%~76%, ~0.33%, and 3~41 for pharmacophore, and 20%~63%, 2%~10%, and 6~54 for clustering methods respectively. Therefore, the general performance of ML appears to be comparable to or in some cases better than the reported performances of the VS studies by using structure-based, pharmacophore and clustering methods.

While exhibiting equally good yield, in screening extremely-large (≥1 million) and large (130,000~400,000) libraries, the currently developed ML VS tools appear to show lower hit-rates and, in some cases, lower enrichment factors than the best performing structure-based VS tools. For instance, in screening extremely-large libraries, the reported yields, hit-rates and enrichment factors of ML VS tools are in the range of 55%~81%, 0.2%~0.7% and 110~795 respectively [23, 32, 113], compared to those of 62%~95%, 0.65%~35% and 20~1,200 by structure-based VS tools [24, 25]. In screening libraries of ~98,000 compounds, the reported hit-rates of some ML VS tools are comparable to those of structure-based VS tools, but their enrichment factors are substantially smaller.

The lower hit rates at equally good yields in screening larger libraries likely arise from higher false positive rates. Because the vast majority of the compounds are expected to be true negatives, a model that predicts more compounds as active tends to show lower hit rate for larger libraries. Conversely, a model that classifies higher percentage of compounds as inactive tends to show higher hit rate for larger libraries. As the number of unique scaffolds generally increases with the size of VS libraries [45], the higher false positive rates suggest that the ML VS tools developed in these studies may not be trained by a sufficiently diverse spectrum of negatives. It is not uncommon for the pharmaceutical industry to screen >1 million compounds per high-throughput screening campaign [63]. Therefore, improvement of hit-rate and enrichment factor is highly desirable for developing practically useful ML VS tools.

Two approaches have been explored for minimizing false hits. One is the selection of top-ranked hits, which has been extensively used in ligand-based [25, 26, 30, 31, 86, 114] and structure-based [10, 12-14, 115, 116] VS tools. The other is the elimination of unlikely hits at the pre-screening stage by using such filters as Lipinski's rule of five [117] for drug-like compounds [11], identification of specific chemical groups or interaction patterns [10, 12, 16, 118], and pharmacophore recognition [13]. In addition to the application of the two approaches, the performance of ML VS tools in screening large libraries may be further improved by using training sets of more diverse spectrum of compounds to develop more optimally performing ML VS tools. These tools have been generated by using two-tier supervised classification ML methods [24-26, 28-31, 119], which require training sets of diverse spectrum of active and inactive compounds. The training inactive compounds in most of the reported studies have been collected from a few hundred known inactive compounds or/and putative inactive compounds from a few dozen biological target classes in MDDR database [24-26, 28-31, 119], which may not always be sufficient to fully represent inactive compounds in the vast chemical space, thereby making it difficult to optimally minimize false hit prediction rate of ML models.

In some cases, active compounds selected from the literature for training ML algorithms are from a single chemical series. As a result, the set of true positives tend to oversample small regions of chemical space distinct from that covered by the set of putative inactive compounds, which may lead to a biased ML model that performs good in the overly populated regions but performs significantly worse elsewhere.

ASSESSMENT OF THE FEASIBILITY FOR IMPROVING HIT RATE AND ENRICHMENT FACTOR OF MACHINE LEARNING METHODS

To test the feasibility of further improving the performance of ML in screening large libraries by using training-sets

Table 2. Performance of Machine Learning Methods in Virtual Screening Test for Identifying Inhibitors, Agonists and Substrates of Proteins of Pharmaceutical Relevance

Screening Task	Compounds Screened		Method and Reference of Reported Study	Molecular Descriptors	Compounds in Training Set (No of Positives / No of Negatives)	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included				No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
COX2 inhibitors	2.5M	22	SVM [31]	Molecular fingerprints	94/200K	2,500	0.1%	18	81%	0.7%	795
	25,300	25	SVM+ BKD [25]	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5
COX inhibitors	102,514	536	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	76	14.3%	1.4%	2.7
	98,435	536	CKD [29]	Pipeline pilot	100/4000	984	1%	232	43.4%	23.7%	43.1
				ECFP4	100/4000	984	1%	365	68.1%	37.2%	67.7
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	240	44.7%	24.4%	44.5
Thrombin inhibitors	2.5M	46	SVM [31]	Molecular fingerprints	188/200K	11,250	0.45%	25	55%	0.2%	108.7
	102,514	703	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	367	52,3%	7.1%	10.3
	98,435	703	CKD [29]	Pipeline pilot	100/4000	984	1%	435	61.9%	44.4%	61.7
				ECFP4	100/4000	984	1%	603	85.8%	61.5%	85.5
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	381	54.2%	38.9%	54.0
Protease inhibitors	171,726	118	SVM [26]	Extended connectivity fingerprints	228/4200	1717	1%	26	22%	1.5%	21.8
			LMNB [26]					19	16%	1%	14.5
Chemokine receptor antagonists	171,560	128	SVM [26]	Extended connectivity fingerprints	258/4199	1716	1%	70	55%	4.1%	54.9
			LMNB [28, 30]					68	53%	3.9%	52.3
5HT3 antagonists	102,514	652	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	236	36.3%	4.6%	7.2
	98,435	852	CKD [29]	Pipeline pilot	100/4000	984	1%	480	56.4%	49.0%	56.3
				ECFP4	100/4000	984	1%	680	79.8%	69.4%	79.8
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	529	62.1%	54.0%	62.1
5HT1A antagonists	102,514	727	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	224	30.9%	4.3%	6.1
	98,435	727	CKD [29]	Pipeline pilot	100/4000	984	1%	268	36.9%	27.3%	36.9
				ECFP4	100/4000	984	1%	426	58.6%	43.5%	58.7
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	319	43.9%	32.6%	44.0
5HT reuptake inhibitors	102,514	259	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	65	25%	1.2%	4.7
	98,435	259	CKD [29]	Pipeline pilot	100/4000	984	1%	131	50.7%	13.4%	51.5
				ECFP4	100/4000	984	1%	194	75.6%	19.7%	75.9
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	137	52.9%	14.0%	53.8

(Table 2) contd…..

Screening Task	Compounds Screened		Method and Reference of Reported Study	Molecular Descriptors	Compounds in Training Set (No of Positives / No of Negatives)	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included				No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
D2 antagonists	102,514	295	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	90	30.6%	1.7%	5.9
	98,435	295	CKD [29]	Pipeline pilot	100/4000	984	1%	132	44.7%	13.5%	44.9
				ECFP4	100/4000	984	1%	219	74.4%	22.4%	74.7
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	137	46.4%	14.0%	53.8
Rennin inhibitors	102,514	1030	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	972	94.4%	18.9%	18.9
	98,435	1030	CKD [29]	Pipeline pilot	100/4000	984	1%	842	81.8%	86.0%	81.9
				ECFP4	100/4000	984	1%	960	93.2%	98.0%	93.3
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	710	68.9%	72.4%	69.0
Angiotesin II AT1 antagonists	102,514	843	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	776	92.1%	15.1%	18.4
	98,435	843	CKD [29]	Pipeline pilot	100/4000	984	1%	393	46.6%	40.1%	46.6
				ECFP4	100/4000	984	1%	593	70.4%	60.6%	70.4
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	384	45.6%	39.2%	45.6
Substance P antagonists	102,514	1146	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	378	33%	7.3%	6.5
	98,435	1146	CKD [29]	Pipeline pilot	100/4000	984	1%	705	61.5%	71.9%	61.5
				ECFP4	100/4000	984	1%	942	82.2%	96.1%	82.2
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	509	44.4%	51.9%	44.4
HIV protease inhibitors	102,514	650	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	377	58%	7.3%	11.5
	98,435	650	CKD [29]	Pipeline pilot	100/4000	984	1%	436	67.1%	44.5%	67.4
				ECFP4	100/4000	984	1%	574	88.3%	58.6%	88.7
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	355	54.6%	36.2%	54.9
Protein kinase C inhibitors	102,514	353	BKD [30, 86]	Extended connectivity fingerprints	100/400	5125	5%	81	23.1%	1.5%	4.4
	98,435	353	CKD [29]	Pipeline pilot	100/4000	984	1%	238	67.3%	24.2%	67.3
				ECFP4	100/4000	984	1%	291	82.5%	29.7%	82.5
			SVM-RBF [29]	Pipeline pilot	100/4000	984	1%	206	58.3%	21.0%	58.3
MAO inhibitors	101,437	1166	BKD [24]	Atom pairs and topological torsions APTT descriptors	1166/3834	6000	5.9%	600	51.4%	10%	11.5
Muscarinic M1 agonists	98,435	748	CKD [29]	Pipeline pilot	100/4000	984	1%	467	62.4%	47.4%	62.4
				ECFP4	100/4000	984	1%	597	79.8%	60.7%	79.8

(Table 2) contd…..

Screening Task	Compounds Screened		Method and Reference of Reported Study	Molecular Descriptors	Compounds in Training Set (No of Positives / No of Negatives)	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included				No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
NMDA receptor antagonists	98,435	1211	CKD [29]	Pipeline pilot	100/4000	984	1%	604	49.9%	61.4%	49.9
				ECFP4	100/4000	984	1%	889	73.4%	90.3%	73.4
Nitric oxide synthase inhibitors	98,435	277	CKD [29]	Pipeline pilot	100/4000	984	1%	192	69.3%	19.5%	69.7
				ECFP4	100/4000	984	1%	244	88.2%	27.3%	97.6
Aldose reductase inhibitors	98,435	782	CKD [29]	Pipeline pilot	100/4000	984	1%	436	55.8%	44.3%	56.1
				ECFP4	100/4000	984	1%	665	85.0%	67.6%	85.5
Reverse transcriptase inhibitors	98,435	419	CKD [29]	Pipeline pilot	100/4000	984	1%	238	56.9%	24.2%	56.3
				ECFP4	100/4000	984	1%	337	80.4%	34.2%	79.6
Aromatase inhibitors	98,435	413	CKD [29]	Pipeline pilot	100/4000	984	1%	284	68.7%	28.8%	68.6
				ECFP4	100/4000	984	1%	389	94.1%	39.5%	94.0
Phospholipase A2 inhibitors	98,435	604	CKD [29]	Pipeline pilot	100/4000	984	1%	297	49.2%	30.2%	49.5
				ECFP4	100/4000	984	1%	447	74.0%	45.4%	74.5
CDK2 inhibitors	25,300	25	SVM+ BKD [25]	DRAGON descriptors	125/5035	506	2%	18	72%	3.5%	35.4
FXa inhibitors	25,300	25	SVM+ BKD [25]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	N/A
PDE5 inhibitors	50,000	19	RO5+ DS [136]	Pharmacophore and macroscopic descriptors	130/10K	1821	3.6%	11	57.8%	0.6%	15.8
	25,300	25	SVM+ BKD [25]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	41.5
Alpha1A AR antagonists	25,300	25	SVM+ BKD [25]	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5

The relevant literature references are given in the method column.

BKD – binary kernel discrimination; CKD – Continuous kernel discrimination; DS – decision tree; LMNB – laplacian modified naive Bayesian; SVM – support vector machine; DRAGON – (an application for the calculation of molecular descriptors); AR – androgen receptor; PDE 5 – phosphodiesterase type 5; FXa – factor Xa; CDK2 – cyclin-dependent kinase 2; MAO – mono amino oxidase; HIV – human immunodeficiency virus; COX – cycloocygenase.

of more diverse spectrum of inactive compounds, SVM [25-29, 31] VS tools were developed for identifying two classes of compounds, one is the active compounds of a single mechanism, dihydrofolate reductase (DHFR) inhibitors, and the other is the active compounds of multiple mechanisms, central nervous system (CNS) active agents. DHFR inhibitors are for anti-microbial [120], anti-cancer [121], and anti-parasitic diseases [122]. CNS active agents are of diverse activities that produce anxiolytic, antipsychotic, antidepressant, analgesic, anticonvulsant, antimigraine, antiischemic, antiparkinsonian, nootropic, neurologic, epileptic, neuroleptic, neurotropic, neuronal injury inhibiting, narcotics antagonizing, and CNS stimulating effects [123]. Although the selection of these two compound classes is somewhat arbitrary, due to their diverse therapeutic applications and structural frameworks, they nevertheless serve as part of useful benchmarks for testing the performance of ML VS tools in screening large compound libraries. While only SVM was tested here, other ML methods are expected to produce similar level of performances as demonstrated by several studies [26, 29].

These SVM VS tools were trained by using known active compounds and putative inactive compounds extracted from compound families that contain no known active compound. No filter was applied to the training set prior to the model construction. A total of 755 DHFR inhibitors were collected from a publication [124], and 16,182 CNS active agents were retrieved from MDDR database. Compound families can be generated by clustering distinct compounds from chemical databases into groups of similar structural and physicochemical properties [20]. There are 7,855 and 3,440 families that contain no known DHFR inhibitor and CNS active agent, respectively. Thus datasets of 44,856 putative non-DHFR inhibitors and 20,465 putative non-CNS active compounds were generated by random selection of 5~6 representative compounds from each of these families, respectively.

Table 3. Performance of Docking Methods in Virtual Screening Test for Identifying Inhibitors, Agonists and Substrates of Proteins of Pharmaceutical Relevance

Screening Task	Compounds Screened		Method and Reference of Reported Study	No of Pre-Docking Selected Compounds	Docking Cut-Off	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included				No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
Factor Xa inhibitors	2M	630	AUTODOCK + pre-docking RO5 and EA screen [11]	60,000	Binding energy < -10.5 kcal/mol	60,000	3%	392	62%	0.65%	20
COX2 inhibitors	1.2M	355	DOCK+ pre-docking chemical group screen [10]	13,711	DOCK scores < -35	959	0.08% for all; 7% for actually docked	337	95%	35.2%	1189.2 for all; 13.6 for actually docked
Human casein kinase II	400K	>4	DOCK4 + H-bond and hinge segment screen [16]	<400K	N/A	35	0.0087%	4	N/A	11.4%	N/A
Thyroid hormone receptor antagonists	250K	>14	ICM VLS module (Molsoft) [18] + pre-docking RO5	190K	Selected 75 from top-100 dock scores	75	0.03% for all; 0.039% for actually docked	14	N/A	18.7%	N/A
PTP1B inhibitors	235K	>127	DOCK3.5 + atom count (17~60) screen [17]	165,581	Top-500 + Top-500	889	0.38%	127	N/A	14.3%	N/A
	141K	10	GOLD + elements and chemical group screen [12]	<141K	Top-2%	<2820	<2.5%	8	80%	<0.28%	39.4
BCL-2 inhibitors	206,876	>1	DOCK3.5 + non-peptidic screen [137]	<206,876	Top-500	35	0.017%	1	N/A	2.9%	N/A
HIV-1 protease inhibitors	141K	5	GLIDE + elements and chemical group screen [12]	<141K	Top-5%	<7050	<5%	1	20%	<0.014%	4.6
HDM2 inhibitors	141K	14	DOCK + elements and chemical group screen [12]	<141K	Top-5%	<7050	<5%	4	28.6%	<0.056%	5.7
UPA inhibitors	141K	10	GOLD + elements and chemical group screen [12]	<141K	Top-2%	<2820	<2.5%	9	90%	<0.32%	45.1
Alpha 1A adrenergic receptor antagonists	141K	>38	GOLD on homology model + pharmacophore screen [13]	22,950	Top-300	300	0.21%	38	N/A	N/A	N/A
Thrombin inhibitors	141K	10	GLIDE + elements and chemical group screen [12]	<141K	Top-2%	<2820	<2.5%	3	30%	<0.11%	15.5
	133.8K	760	FlexX + Similarity [15]	<133.8K	Top-1%	1338	1%	231	29.3%	17.3%	30.5
DHFR inhibitors	135K	165	DOCK3.5.54 applied to holo form [14]	135K	Top-1% of 50k docked	1350	1%	47	25%	3.4%	27.8
			DOCK3.5.54 applied to appo form [14]	135K	Top-1% of 100k docked	1000	1%	16	9.7%	1.6%	13.1

(Table 3) contd…..

Screening Task	Compounds Screened		Method and Reference of Reported Study	No of Pre-Docking Selected Compounds	Docking Cut-Off	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included				No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
Neutral endopeptidase inhibitors	135K	356	DOCK3.5.54 [14]	135K	Top-1% of 125.5K docked	1255	0.74%	3	0.8%	0.24%	~1
Thrombin inhibitors	135K	788	DOCK3.5.54 [14]	135K	Top-1% of 121.5K docked	1215	0.9%	61	7.7%	5.0%	8.6
Thymidylate synthase inhibitors	135K	185	DOCK3.5.54 [14]	135K	Top-1% of 54K docked	540	0.4%	49	26.5%	9.1%	66.4
Phospholipase C inhibitors	135K	25	DOCK3.5.54 [14]	135K	Top-1% of 123K docked	1230	0.9%	5	20%	0.4%	21.6
Adenosine kinase inhibitors	135K	356	DOCK3.5.54 applied to holo form [14]	135K	Top-5% of database	4500	3.3%	10	2.8%	0.22%	~1
			DOCK3.5.54 applied to appo form [14]	135K	Top-5% of database	4500	3.3%	5	1.4%	0.11%	<1
	133.8K	59	FlexX + Similarity [15]	<133.8K	Top-1%	1338	1%	13	22%	0.97%	22.0
Acetylcholinesterase inhibitors	135K	637	DOCK3.5.54 applied to holo form [14]	135K	Top-1% of 77K docked	770	0.57%	49	7.7%	6.4%	13.6
			DOCK3.5.54 applied to appo form [14]	135K	Top-1% of 37.5K docked	375	0.28%	25	3.9%	6.7%	14.2
HMG-CoA reductase inhibitors	133.8K	1016	FlexX + Similarity [15]	<133.8K	Top-1%	1338	1%	35	3.4%	2.6%	3.4

The relevant literature references are given in the method column.

Table 4. Performance of Pharmacophore Methods in Virtual Screening Test for Identifying Inhibitors, Agonists and Substrates of Proteins of Pharmaceutical Relevance

Screening Task	Compounds Screened		Method and Reference of Reported Study	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included		No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
ACE inhibitors	3.8M	55	Pharmacophore [127]	1M	26%	39	70.1%	0.0039%	2.8
	3.8M	55	Structure-based pharmacophore [128]	91K	2.4%	6	10.9%	0.0066%	4.6
11β-hydroxysteroid dehydrogenase 1 inhibitors	1.77M	144	Pharmacophore [21]	20.3K	1.15%	17	11.8%	0.084%	10.3
Rhinovirus 3C protease inhibitors	380K	30	Pharmacophore [22]	6,917	1.82%	23	76.7%	0.33%	41.8

The relevant literature references are given in the method column.

Table 5. Performance of Clustering Methods in Virtual Screening Test for Identifying Inhibitors, Agonists and Substrates of Proteins of Pharmaceutical Relevance

Screening Task	Compounds Screened		Method and Reference of Reported Study	Compounds Selected		Known Hits Selected			
	No of Compounds	No of Known Hits Included		No of Compounds Selected	Percentage of Screened Compounds Selected	No of Hits Selected	Yield	Hit Rates	Enrichment Factor
ACE inhibitors	344.5K	490	Hierachical k-means [20]	5590	1.6%	246	50.2%	4.4%	31.2
			NIPALSTREE [20]	8174	2.4%	188	38.4%	2.3%	16.2
			Hierachical k-means + NIPALSTREE disjunction [20]	12240	3.6%	306	62.4%	2.5%	17.6
			Hierachical k-means + NIPALSTREE conjunction [20]	1662	0.48%	128	26.1%	7.7%	54
COX inhibitors	344.5K	1556	Hierachical k-means [20]	15322	4.4%	761	48.9%	5.0%	11
			NIPALSTREE [20]	22321	6.5%	625	40.2%	2.8%	6.16
			Hierachical k-means + NIPALSTREE disjunction [20]	33793	9.8%	980	63.0%	2.9%	6.42
			Hierachical k-means + NIPALSTREE conjunction [20]	3980	1.2%	406	26.1%	10.2%	22.6
Adrenoceptor ligand	344.5K	542	Hierachical k-means [20]	21285	6.2%	298	55.0%	1.4%	8.99
			NIPALSTREE [20]	28125	8.2%	270	49.8%	0.96%	6.14
			Hierachical k-means + NIPALSTREE disjunction [20]	42365	12.3%	394	72.7%	0.93%	5.93
			Hierachical k-means + NIPALSTREE conjunction [20]	6692	1.9%	174	32.1%	2.6%	16..3
Glucocorticoid receptor ligand	344.5K	91	Hierachical k-means [20]	3750	1.1%	27	29.7%	0.72%	27..3
			NIPALSTREE [20]	3469	1.0%	17	18.7%	0.49%	18.7
			Hierachical k-means + NIPALSTREE disjunction [20]	7317	2.1%	30	33.0%	0.41%	15.6
			Hierachical k-means + NIPALSTREE conjunction [20]	538	0.16%	14	15.4%	2.6%	98
GABA receptor ligand	344.5K	478	Hierachical k-means [20]	10000	2.9%	110	23%	1.1%	7.97
			NIPALSTREE [20]	17143	5.0%	84	17.6%	0.49%	3.51
			Hierachical k-means + NIPALSTREE disjunction [20]	24265	7.0%	165	34.5%	0.68%	4.86
			Hierachical k-means + NIPALSTREE conjunction [20]	2636	0.77%	29	6.1%	1.1%	7.77

The relevant literature references are given in the method column.

The developed SVM VS tools were tested in screening libraries of 2.986 million compounds from the PUBCHEM

database that are not in the training sets of these SVM VS tools. Although filters such as Lipinski's rule of five can substantially improve VS performance [11], no filter was used here so as to objectively evaluate the true performance of the developed SVM models. PUBCHEM was selected as the testing library because it is a public source (thus convenient for comparative studies by all scientists) that contains significantly higher number of unique compounds than other VS libraries [125]. Moreover, a substantial percentage of compounds contained in the other libraries and datasets are present in PUBCHEM [125]. Therefore, the levels of scaffold diversity and compound representation of PUBCHEM appear to be reasonably high for testing VS performance.

As shown in Table **6**, the developed SVM VS tools identify 52.4% and 66.6% of the known hits, which are comparable to the range of 62%~95% by structure-based VS tools [10, 11] and 55%~81% by published ML and other ligand-based VS tools [25, 28, 31] in screening libraries of ≥1 million compounds, and they are also comparable to the percentages in screening libraries of 98,400~344,500 compounds by other structure-based [7-9, 12-18] and ligand-based [20, 22, 26, 28-31] VS tools. Moreover, these SVM VS tools appear to show relatively lower "false" hit identification rate. Without the use of top-ranked cut-off or additional filter, they identified a total of 160 and 9,502 virtual hits, which are comparable to and in some cases smaller than those identified by structure-based [7-18] and other ligand-based [20, 25, 26, 28-30, 52, 81, 126] VS tools even though a substantially larger number of compounds (2.983M *vs* 98.4K~2.5M) were screened. By using Lipinski's rule of five [117] as a filter, the numbers of identified virtual hits were further reduced to 115 and 8,035, suggesting that introduction of such filters or combination with other VS methods may enable further reduction of the number of predicted hits.

The hit-rates of the developed SVM VS tools are 73.8% and 4.7% respectively, which are comparable to those of 0.65%~35% by structure-based VS tools [24, 25] and substantially improved against those of 0.2%~0.7% by other reported SVM VS tools [25, 28, 31] in screening extremely large libraries. These hit-rates are also greater than the majority of the hit-rates in screening large libraries of 98,400~344,500 compounds by structure-based [7-18] and other ligand-based [20-22, 26, 28-31, 127, 128] VS tools. The enrichment factors of the developed SVM VS tools are 10,543 and 214 for the four classes of compounds respectively, which are comparable to the best performing and better than most of the reported enrichment factors of 20~1,200 by structure-based [24, 25] and 110~795 by other reported SVM [25, 28, 31] VS tools in screening extremely large libraries. These results suggest it is feasible to improve the performance of ML VS tools using training-sets of more diverse spectrum of inactive compounds.

PERSPECTIVES

ML methods have shown promising capability in virtual screening of compounds of diverse ranges of structures for identifying compounds of a wide variety of pharmacodynamic and other properties. In virtual screening of large libraries, these methods have been found to be capable of achieving comparable performance as other structure-based and ligand-based VS methods. By using training sets of more diverse spectrum of inactive compounds, the hit-rates and enrichment factors of SVM VS tools can be substantially improved to the level comparable to and in some cases higher than those of the best performing structure-based and ligand-based VS tools reported in the literature.

A key to the successful application of VS tools is the identification and charting of the relevant chemical space, which practically depends on the overall coverage of chemical-space [129, 130] and adequate representation of individual scaffolds [131]. Efforts have been directed at the development and incorporation of chemical-space access and chemical-landscape analysis methods into the screening libraries and methods [132, 133]. For instance, geometric hashing method has been explored for enabling scalable partitioning and efficient exploration of huge chemical spaces

Table 6. Performance of Support Vector Machines Virtual Screening Tools Developed in this Work for Identifying HIV Protease Inhibitors, DHFR Inhibitors, Dopamine Antagonists, and CNS Active Agents in Screening 2.986 Million Compounds

Screening Task	Compounds Screened				Virtual Hits Selected by SVM					Known Hits Selected by SVM			
	No of Compounds	No of Known Hits Not in Training Sets of SVM-LBVS Tool	Percent of Known Hits	No of Families Covered by Known Hits	No of Selected Virtual Hits	Percent of Selected Virtual Hits Not in the Families Covered by Known Hits	Percent of Screened Compounds Selected as Virtual Hits	No of Selected Virtual Hits Passed Rule-of-Five	Percent of Selected Virtual Hits Passed Rule-of-Five and Not in the Families Covered by Known Hits	No of Known Hits Selected	Yield	Hit Rates	Enrichment Factor
DHFR inhibitors	2.986M	225	0.007%	60	160	71.3%	0.0054%	115	64.4%	118	52.4%	73.8%	10543
CNS active agents	2.986M	664	0.022%	519	9502	85.7%	0.32%	8035	84.1%	442	66.6%	4.7%	214

Compound families are generated by clustering compounds on the basis of their molecular descriptors using established clustering method [20].

by ML classification and regression methods at modest computing costs [134].

Because of their high computing speed and capability for covering highly diverse spectrum compounds, SVM and other ML methods can be potentially explored to develop useful VS tools to complement structure-based and other ligand-based VS tools or to be used as part of integrated VS tools in facilitating lead discovery [11, 15, 128]. Regression-based ML methods can be used for quantitative prediction of the activity levels if the activity data are available for a sufficient number of compounds with specific binding activity. Regression methods have the capacity for estimating the contribution of specific structural and physicochemical features of the compounds to a particular activity [135]. This capacity may be explored for probing the mechanism of action for a specific group of compounds that possess a particular property.

REFERENCES

[1] Shoichet, B.K. *Nature*, **2004**, *432*(7019), 862-865.
[2] Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. *Drug Discov. Today*, **2004**, *9*, 27-34.
[3] Davies, J.W.; Glick, M.; Jenkins, J.L. *Curr. Opin. Chem. Biol.*, **2006**, *10*, 343-351.
[4] Willett, P. *Drug Discov. Today*, **2006**, *11*, 1046-1053.
[5] Van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug. Discov.*, **2003**, *2*, 192-204.
[6] Matthew W. B. Trotter, S.B.H. *QSAR Combi. Sci.*, **2003**, *22*, 533-548.
[7] Ghosh, S.; Nie, A.; An, J.; Huang, Z. *Curr. Opin. Chem. Biol.*, **2006**, *10*, 194-202.
[8] Shoichet, B.K.; McGovern, S.L.; Wei, B.; Irwin, J.J. *Curr. Opin. Chem. Biol.*, **2002**, *6*, 439-446.
[9] Jansen, J.M.; Martin, E.J. *Curr. Opin. Chem. Biol.*, **2004**, *8*, 359-364.
[10] Mozziconacci, J.C.; Arnoult, E.; Bernard, P.; Do, Q.T.; Marot, C.; Morin-Allory, L. *J. Med. Chem.*, **2005**, *48*, 1055-1068.,
[11] Vidal, D.; Thormann, M.; Pons, M. *J. Chem. Inf. Model.*, **2006**, *46*, 836-843.
[12] Cummings, M.D.; DesJarlais, R.L.; Gibbs, A.C.; Mohan, V.; Jaeger, E.P. *J. Med. Chem.*, **2005**, *48*, 962-976.
[13] Evers, A.; Klabunde, T. *J. Med. Chem.*, **2005**, *48*, 1088-1097.
[14] Lorber, D.M.; Shoichet, B.K. *Curr. Top. Med. Chem.*, **2005**, *5*, 739-749.
[15] Stiefl, N.; Zaliani, A. *J. Chem. Inf. Model.*, **2006**, *46*, 587-596.
[16] Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. *J. Med. Chem.*, **2003**, *46*, 2656-2662.
[17] Doman, T.N.; McGovern, S.L.; Witherbee, B.J.; Kasten, T.P.; Kurumbail, R.; Stallings, W.C.; Connolly, D.T.; Shoichet, B.K. *J. Med. Chem.*, **2002**, *45*, 2213-2221.
[18] Enyedy, I.J.; Ling, Y.; Nacro, K.; Tomita, Y.; Wu, X.; Cao, Y.; Guo, R.; Li, B.; Zhu, X.; Huang, Y.; Long, Y.Q.; Roller, P.P.; Yang, D.; Wang, S. *J. Med. Chem.*, **2001**, *44*, 4313-4324.
[19] Oprea, T.I.; Matter, H. *Curr. Opin. Chem. Biol.*, **2004**, *8*, 349-358.
[20] Bocker, A.; Schneider, G.; Teckentrup, A. *J. Chem. Inf. Model.*, **2006**, *46*, 2220-2229.
[21] Schuster, D.; Maurer, E.M.; Laggner, C.; Nashev, L.G.; Wilckens, T.; Langer, T.; Odermatt, A. *J. Med. Chem.*, **2006**, *49*, 3454-3466.
[22] Steindl, T.; Laggner, C.; Langer, T. *J. Chem. Inf. Model.*, **2005**, *45*, 716-724.
[23] Manallack, D.T.; Livingstone, D.J. *Eur. J. Med. Chem.*, **1999**, *34*, 195-208.
[24] Harper, G.; Bradshaw, J.; Gittins, J.C.; Green, D.V.; Leach, A.R. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 1295-1300.
[25] Jorissen, R.N.; Gilson, M.K. *J. Chem. Inf. Model.*, **2005**, *45*, 549-561.
[26] Glick, M.; Jenkins, J.L.; Nettles, J.H.; Hitchings, H.; Davies, J.W. *J. Chem. Inf. Model.*, **2006**, *46*, 193-200.
[27] Li, H.; Ung, C.Y.; Yap, C.W.; Xue, Y.; Li, Z.R.; Chen, Y.Z. *J. Mol. Graph. Model.*, **2006**, *25*, 313-323.
[28] Lepp, Z.; Kinoshita, T.; Chuman, H. *J. Chem. Inf. Model.*, **2006**, *46*, 158-167.
[29] Chen, B.; Harrison, R.F.; Papadatos, G.; Willett, P.; Wood, D.J.; Lewell, X.Q.; Greenidge, P.; Stiefl, N. *J. Comput. Aided. Mol. Des.*, **2007**, *21*, 53-62.
[30] Hert, J.; Willett, P.; Wilton, D.J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. *J. Chem. Inf. Model.*, **2006**, *46*, 462-470.
[31] Franke, L.; Byvatov, E.; Werz, O.; Steinhilber, D.; Schneider, P.; Schneider, G. *J. Med. Chem.*, **2005**, *48*, 6997-7004.
[32] Trotter, M.W.B.; Holden, S.B. *QSAR Combin. Sci.*, **2003**, *22*, 533-548.
[33] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.*, **2001**, *26*, 5-14.
[34] Plewczynski, D.; Spieser, S.A.; Koch, U. *J. Chem. Inf. Model.*, **2006**, *46*, 1098-1106.
[35] Schroeter, T.; Schwaighofer, A.; Mika, S.; Laak, A.T.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K.R. *Mol. Pharm.*, **2007**, *4*, 524-538.
[36] Li, H.; Yap, C.W.; Ung, C.Y.; Xue, Y.; Li, Z.R.; Han, L.Y.; Lin, H.H.; Chen, Y.Z. *J. Pharm. Sci.*, **2007**, *96*, 2838-2860.
[37] Fox, T.; Kriegl, J.M. *Curr. Topics Med. Chem.*, **2006**, *6*, 1579-1591.
[38] Duch, W.; Swaminathan, K.; Meller, J. *Curr. Pharm. Des.*, **2007**, *13*, 1497-1508.
[39] Warmuth, M.K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 667-673.
[40] Asogawa, M.; Osoda, T.; Fujiwara, Y.; Yamashita, Y. *NEC Tech. J.*, **2003**, *56*, 28-32.
[41] Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J.P.; Kohn, H.; Tropsha, A. *J. Med. Chem.*, **2004**, *47*, 2356-2364.
[42] Oloff, S.; Mailman, R.B.; Tropsha, A. *J. Med. Chem.*, **2005**, *48*, 7322-7332.
[43] Roberts, D.W.; Aptula, A.O.; Patlewicz, G. *Chem. Res. Toxicol.*, **2007**, *20*, 44-60.
[44] Mozziconacci, J.C.; Arnoult, E.; Baurin, N.; Marot, C.; Morin-Allory, L. In *9th Electronic Computational Chemistry Conference (ECCC9)* **2007**.
[45] Krier, M.; Bret, G.; Rognan, D. *J. Chem. Inf. Model.*, **2006**, *46*, 512-524.
[46] Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1882-1889.
[47] Doniger, S.; Hofman, T.; Yeh, J. *J. Comput. Biol.*, **2002**, *9*, 849-864.
[48] He, L.; Jurs, P.C.; Custer, L.L.; Durham, S.K.; Pearl, G.M. *Chem. Res. Toxicol.*, **2003**, *16*, 1567-1580.
[49] Snyder, R.D.; Pearl, G.S.; Mandakas, G.; Choy, W.N.; Goodsaid, F.; Rosenblum, I.Y. *Environ. Mol. Mutagen.*, **2004**, *43*, 143-158.
[50] Xue, Y.; Li, Z.R.; Yap, C.W.; Sun, L.Z.; Chen, X.; Chen, Y.Z. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1630-1638.
[51] Yap, C.W.; Cai, C.Z.; Xue, Y.; Chen, Y.Z. *Toxicol. Sci.*, **2004**, *79*, 170-177.
[52] Yap, C.W.; Chen, Y.Z. *J. Pharm. Sci.*, **2005**, *94*, 153-168.
[53] Zernov, V.V.; Balakin, K.V.; Ivaschenko, A.A.; Savchuk, N.P.; Pletnev, I.V. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 2048-2056.
[54] Fang, H.; Tong, W.; Shi, L.M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B.S.; Xie, Q.; Dial, S.L.; Moland, C.L.; Sheehan, D.M. *Chem. Res. Toxicol.*, **2001**, *14*, 280-294.
[55] Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. *Environ. Health. Perspect.*, **2004**, *112*, 1249-1254.
[56] Jacobs, M.N. *Toxicology*, **2004**, *205*, 43-53.
[57] Hu, J.Y.; Aizawa, T. *Water. Res.*, **2003**, *37*, 1213-1222.
[58] Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON*; Version 5.3. Talete SRL: Milan, IT, **2005**.
[59] Hall, L.H.; Kellogg, G.E.; Haney, D.N. *Molconn-Z*, Version 4.05+. EduSoft, L.C, **2002**.
[60] Wegner, J. K. *JOELib/JOELib2*; Department of Computer Science, University of Tübingen: Germany, **2005**.
[61] Li, Z.R.; Han, L.Y.; Xue, Y.; Yap, C.W.; Li, H.; Jiang, L.; Chen, Y.Z. *Biotechnol. Bioeng.*, **2007**, *97*, 389-396.
[62] Gasteiger.J.T., E. *Handbook of Chemoinformatics*; Weinheim: Wiley-VCH, **2003**.
[63] Hemmer, M.C.; Steinhauer, V.; Gasteiger, J. *Vib. Spectr.*, **1999**, *19*, 151-164.
[64] Rücker, G.; Rücker, C. *J. Chem. Inf. Comput. Sci.*,**1993**, *33*, 683-695.
[65] Schuur, J.H.; Setzer, P.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.*, **1996**, *36*, 334-344.

[66] Pearlman, R.S.; Smith, K.M. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 28-35.
[67] Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. *J. Comput. Aided. Mol. Des*., **1997**, *11*, 79-92.
[68] Galvez, J.; Garcia, R.; Salabert, M.T.; Soler, R. *J. Chem. Inf. Comput. Sci*., **1994**, *34*, 520-525.
[69] Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci*., **2002**, *42*, 682-692.
[70] Randic, M. *Tetrahedron* **1975**, *31*, 1477-1481.
[71] Randic, M. *N. J. Chem*., **1995**, *19*, 781-791.
[72] Kier, L.B.; Hall, L.H. *Molecular structure description: The electrotopological state*; San Diego: Academic Press; **1999**.
[73] Platts, J.A.; Butina, D.; Abraham, M.H.; Hersey, A. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 835-845.
[74] Sheridan, R.P.; Kearsley, S.K. *Drug Discov. Today*, **2002**, *7*, 903-911.
[75] Patterson, D.E.; Cramer, R.D.; Ferguson, A.M.; Clark, R.D.; Weinberger, L.E. *J. Med. Chem*., **1996**, *39*, 3049-3059.
[76] Lucasius, C.B.; Kateman, G. *Chemom. Intel. Lab. Syst*., **1993**, *19*, 1-33.
[77] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. *Mach. Learn*., **2002**, *46*, 389-422.
[78] Sutter, J.M.; H., K.J. *Microchem. J*., **1993**, *47*, 60-66.
[79] Gramatica, P.; Pilutti, P.; Papa, E. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1794-1802.
[80] Izrailev, S.; Agrafiotis, D.K. *J. Mol. Graph. Model*., **2004**, *22*, 275-284.
[81] Yap, C.W.; Chen, Y.Z. *J. Chem. Inf. Model*., **2005**, *45*, 982-992.
[82] Xue, Y.; Yap, C.W.; Sun, L.Z.; Cao, Z.W.; Wang, J.F.; Chen, Y.Z. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1497-1505.
[83] Li, H.; Xue, Y.; Ung, C.Y.; Yap, C.W.; Li, Z.R.; Chen, Y.Z. *Chem. Res. Toxicol*., **2005**, *18*, 1071-1080.
[84] Lippmann, R.P. *IEEE Acoust. Speech Signal Process. Mag.* **1987**, *4*, 4-22.
[85] Zupan, J.; Gasteiger, J., *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, **1999**.
[86] Chen, B.; Harrison, R.F.; Pasupa, K.; Willett, P.; Wilton, D.J.; Wood, D.J.; Lewell, X.Q. *J. Chem. Inf. Model*., **2006**, *46*, 478-486.
[87] Quinlan, J.R. *C4.5: programs for machine learning*; San Mateo, CA: Morgan Kaufmann, **1993**.
[88] Johnson, R.A.; Wichern, D.W. *Applied multivariate statistical analysis*; Englewood Cliffs, NJ: Prentice Hall, **1982**.
[89] Fix, E.; Hodges, J.L. *Discriminatory analysis: Non-parametric discrimination: Consistency properties*; Texas: USAF School of Aviation Medicine, **1951**.
[90] Huberty, C.J. *Applied discriminant analysis*. New York: John Wiley & Sons, **1994**.
[91] Basak, S.C.; Gute, B.D.; Ghatak, S. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 255-260.
[92] Patankar, S.J.; Jurs, P.C. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 885-899.
[93] Mahmoudi, N.; de Julian-Ortiz, J.V.; Ciceron, L.; Galvez, J.; Mazier, D.; Danis, M.; Derouin, F.; Garcia-Domenech, R. *J. Antimicrob. Chemother*., **2006**, *57*, 489-497.
[94] Prieto, J.J.; Talevi, A.; Bruno-Blanch, L.E. *Mol. Div*., **2006**, *10*, 361-375.
[95] Pedro, D.; Pazzani, M. *Mach. Learn*., **1997**, *29*, 103-107.
[96] Hosmer, D.W.; Lemeshow, S. *Applied logistic regression*; New York: Wiley, **1989**.
[97] Susnow, R.G.; Dixon, S.L. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1308-1315.
[98] Fujishima, S.; Takahashi, Y. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 1006-1009.
[99] Potter, T.; Matter, H. *J. Med. Chem*., **1998**, *41*, 478-488.
[100] Ivanciuc, O. Applications of Support Vector Machines in Chemistry. In: *Reviews in Computational Chemistry*. Lipkowitz, K.B.; Cundari, T.R., Eds.; Wiley-VCH: Weinheim **2007**, Vol. *3*, pp. 291-400.
[101] Vapnik, V.N. *The nature of statistical learning theory*, New York: Springer, **1995**.
[102] Zheng, C.J.; Han, L.Y.; Yap, C.W.; Ji, Z.L.; Cao, Z.W.; Chen, Y.Z. *Pharmacol. Rev*., **2006**, *58*, 259-279.
[103] Stichtenoth, D.O.; Frolich, J.C. *Drugs*, **2003**, *63*, 33-45.
[104] Hochberg, M.C. *Curr. Topics Med. Chem*., **2005**, *5*, 443-448.
[105] Linkins, L.A.; Weitz, J.I. *Curr. Pharmaceut. Des*., **2005**, *11*, 3877-3884.
[106] Francis, C.W. *Curr. Pharmaceut. Des*., **2005**, *11*, 3931-3941.
[107] Ribeiro, S.; Horuk, R. *Pharmacol. Therapeut*., **2005**, *107*, 44-58.
[108] Spaltenstein, A.; Kazmierski, W.M.; Miller, J.F.; Samano, V. *Curr. Topics Med. Chem*., **2005**, *5*, 1589-1607.
[109] Fabbro, D.; Ruetz, S.; Buchdunger, E.; Cowan-Jacob, S.W.; Fendrich, G.; Liebetanz, J.; Mestan, J.; O'Reilly, T.; Traxler, P.; Chaudhuri, B.; Fretz, H.; Zimmermann, J.; Meyer, T.; Caravatti, G.; Furet, P.; Manley, PW. *Pharmacol. Therapeut*., **2002**, *93*, 79-98.
[110] Kumar, R.; Singh, V.P.; Baker, K.M. *J. Mol. Cell. Cardiol*., **2007**, *42*, 1-11.
[111] Rotella, D.P. *Nat. Rev. Drug. Discov*., **2002**, *1*, 674-682.
[112] Pacher, P.; Kecskemeti, V. *Curr. Med. Chem*., **2004**, *11*, 925-943.
[113] Ji, Z.L.; Han, L.Y.; Yap, C.W.; Sun, L.Z.; Chen, X.; Chen, Y.Z. *Drug. Saf*., **2003**, *26*, 685-690.
[114] Wilton, D.J.; Harrison, R.F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G. *J. Chem. Inf. Model*., **2006**, *46*, 471-477.
[115] Alvarez, J.C. *Curr. Opin. Chem. Biol*., **2004**, *8*, 365-370.
[116] Schapira, M.; Raaka, B.M.; Das, S.; Fan, L.; Totrov, M.; Zhou, Z.; Wilson, S.R.; Abagyan, R.; Samuels, H.H. *Proc. Natl. Acad. Sci USA*, **2003**, *100*, 7354-7359.
[117] Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. *Adv. Drug. Deliv. Rev*., **2001**, *46*, 3-26.
[118] Perola, E. *Proteins*, **2006**, *64*, 422-435.
[119] Cui, J.; Han, L.Y.; Lin, H.H.; Zhang, H.L.; Tang, Z.Q.; Zheng, C.J.; Cao, Z.W.; Chen, Y.Z. *Mol. Immunol*., **2007**, *44*, 866-877.
[120] Then, R.L. *J. Chemother*., **2004**, *16*, 3-12.
[121] McGuire, J.J. *Curr. Pharm. Des*., **2003**, *9*, 2593-2613.
[122] Linares, G.E.; Ravaschino, E.L.; Rodriguez, J.B. *Curr. Med. Chem*., **2006**, *13*, 335-360.
[123] H.P. Rang, M.; J.M. Ritter. *Pharmacology*, 4th ed; Churchill Livingstone, **2001**.
[124] Sutherland, J.J.; O'Brien, L.A.; Weaver, D.F. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1906-1915.
[125] Southan, C.; Várkonyi, P.; Muresan, S. *Curr. Topics Med. Chem*., **2007**, *7*, 1502-1508.
[126] Grover, I.I.; Singh, I.I.; Bakshi, I.I. *Pharm. Sci. Technol. Today*, **2000**, *3*, 50-57.
[127] Pirard, B.; Brendel, J.; Peukert, S. *J. Chem. Inf. Model*., **2005**, *45*, 477-485.
[128] Rella, M.; Rushworth, C.A.; Guy, J.L.; Turner, A.J.; Langer, T.; Jackson, R.M. *J. Chem. Inf. Model*., **2006**, *46*, 708-716.
[129] Snyder, R.D.; Smith, M.D. *Drug Discov. Today*, **2005**, *10*, 1119-1124.
[130] Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. *J. Nat. Prod*., **2007**, *70*, 789-794.
[131] Nilakantan, R.; Nunn, D.S. *Drug Discov. Today*, **2003**, *8*, 668-672.
[132] Fitzgerald, S.H.; Sabat, M.; Geysen, H.M. *J. Comb. Chem*., **2007**, *9*, 724-734.
[133] Mauser, H.; Stahl, M. *J. Chem. Inf. Model*., **2007**, *47*, 318-324.
[134] Dutta, D.; Guha, R.; Jurs, P.C.; Chen, T. *J. Chem. Inf. Model*., **2006**, *46*, 321-333.
[135] Stanton, D.T. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 1423-1433.
[136] Yamazaki, K.; Kusunose, N.; Fujita, K.; Sato, H.; Asano, S.; Dan, A.; Kanaoka, M. *Bioorg. Med. Chem. Lett*., **2006**, *16*, 1371-1379.
[137] Wang, J.L.; Liu, D.; Zhang, Z.J.; Shan, S.; Han, X.; Srinivasula, S.M.; Croce, C.M.; Alnemri, E.S.; Huang, Z. *Proc. Natl. Acad. Sci USA*, **2000**, *97*, 7124-7129.

Received: April 29, 2008 **Revised: June 27, 2008** **Accepted: August 29, 2008**