Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds

X. H. Ma,[†,||] R. Wang,[†] S. Y. Yang,[‡] Z. R. Li,[‡,§] Y. Xue,[‡,§] Y. C. Wei,[‡] B. C. Low,[||] and Y. Z. Chen*,[†,‡]

Bioinformatics and Drug Design Group, Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610064, P. R. China, College of Chemistry, Sichuan University, Chengdu, 610064, P. R. China, and Department of Biological Science, National University of Singapore, Blk S2, Level 5, Science Drive 4, Singapore 117543

Received January 18, 2008

Virtual screening performance of support vector machines (SVM) depends on the diversity of training active and inactive compounds. While diverse inactive compounds can be routinely generated, the number and diversity of known actives are typically low. We evaluated the performance of SVM trained by sparsely distributed actives in six MDDR biological target classes composed of a high number of known actives (983−1645) of high, intermediate, and low structural diversity (muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors). SVM trained by regularly sparse data sets of 100 actives show improved yields at substantially reduced false-hit rates compared to those of published studies and those of Tanimoto-based similarity searching method based on the same data sets and molecular descriptors. SVM trained by very sparse data sets of 40 actives (2.4%−4.1% of the known actives) predicted 17.5−39.5%, 23.0−48.1%, and 70.2−92.4% of the remaining 943−1605 actives in the high, intermediate, and low diversity classes, respectively, 13.8−68.7% of which are outside the training compound families. SVM predicted 99.97% and 97.1% of the 9.997 M PUBCHEM and 167K remaining MDDR compounds as inactive and 2.6%−8.3% of the 19,495−38,483 MDDR compounds similar to the known actives as active. These suggest that SVM has substantial capability in identifying novel active compounds from sparse active data sets at low false-hit rates.

INTRODUCTION

As part of the efforts in further developing virtual screening (VS) methods for facilitating lead discovery,[1–4] support vector machines (SVM) [5–11] have recently been explored as ligand-based VS (LBVS) tools to complement or to be used in combination with structure-based VS (SBVS) [1,12–23] and other LBVS [2,6–11,24–29] tools. A particular objective for exploring these approaches is to overcome several problems that have impeded progress in more extensive applications of VS.[1,12,30] These problems include the vastness and sparse nature of chemical space to be searched, limited hit diversity due to the bias of training molecules, limited availability of target structures (only 15% of known proteins have experimentally determined 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects.

SVM is of particular interest because it classifies active compounds based on the differentiating physicochemical profiles between active and inactive compounds rather than structural similarity to active compounds per se. Moreover, SVM does not require the knowledge of target structure and activity-related molecular descriptors and the computation of binding affinity and solvation effects. Its fast speed enables efficient search of vast chemical space. Some of these advantages have been exhibited by the good VS performance in screening large compound libraries.[5,6,9,28] Nontheless, as in the cases of all statistical learning methods, the performance of SVM is significantly influenced by the levels of the training active and inactive compounds in representing the physicochemical profiles of the remaining compounds in the chemical space.

Active compounds (actives) typically occupy small pockets of the chemical space. It may be possible to construct a training active data set to substantially represent the properties of the remaining actives by using a relatively small number of known actives. However, inactive compounds (inactives) generally occupy larger portions of the chemical space. A large number of training inactives is needed to reach a sufficient level of diversity for representing the remaining inactives in the chemical space. SVM constructs a hyperplane in a higher dimensional molecular descriptor space to separate actives from inactives based on whether or not the molecular descriptor vector of a compound is distributed on the known active side of the hyperplane. As illustrated in Figure 1, the position and orientation of the SVM hyperplane, which extends to far regions of the chemical space, can in

* Corresponding author phone: 65-6874-6877; fax: 65-6774-6756; e-mail: phacyz@nus.edu.sg.
[†] Centre for Computational Science and Engineering, National University of Singapore.
[||] Department of Biological Science, National University of Singapore.
[‡] State Key Laboratory of Biotherapy, Sichuan University.
[§] College of Chemistry, Sichuan University.

10.1021/ci800022e CCC: $40.75
Published on Web 06/06/2008

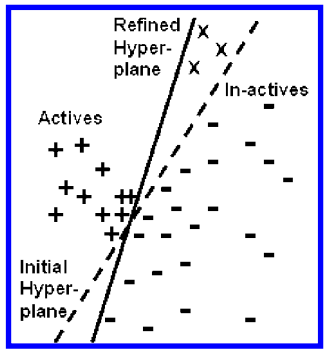

Figure 1. Illustration of the influence of the inactive compounds distributed remotely from the active compounds on the position and orientation of the hyperplane of support vector machines that separates active and inactive compounds. +: active compounds, −: inactive compounds used for constructing the initial hyperplane (dashed line), x: additional inactive compounds used for constructing the more-refined hyperplane (solid line).

many cases be influenced by inactives distributed remotely from the known actives as well as those closely resembling known actives. The level of influence tends to be stronger for sparsely distributed known actives and inactives as there is more room in the local space for altering the position and orientation of the hyperplane. Therefore, highly diverse inactive data sets are typically needed for constructing SVM VS models.[5,30]

Highly diverse inactive training data sets can be routinely generated by large-scale sampling of active compounds of other biological target classes[6,7,9–11,28,29,31] and by using representative compounds from compound families that contain no known active.[5] In contrast, the diversity and the level of representation of active training data sets are often constrained by the small number of known actives sparsely distributed in the active regions of chemical space (active regions are defined as regions of chemical space covered by discovered and yet-to-be-discovered actives). There is a need to evaluate the VS performance of SVM trained by sparse active data sets to determine its capability in identifying novel actives from sparsely distributed known actives.

In this work, we examined the VS performance of SVM trained by sparse active data sets generated from available active data sets of a sufficiently high number of known actives and varying degrees of structural diversity. The high number of actives in the studied data sets makes it possible to generate sufficiently sparse training active data sets, and varying degrees of diversity enables objective evaluation of the VS performance of SVM on different classes of actives. To facilitate comprehensive analysis and further comparative studies, six of the well-studied MDDR biological target classes[28] of a high number of actives (983−1645) of both high, intermediate, and low structural diversity were used for this study. These classes include muscarinic M1 receptor agonists and NMDA receptor antagonists representing high-diversity, thrombin inhibitors and HIV protease inhibitors representing intermediate-diversity, and cephalosporins and renin inhibitors representing low-diversity classes, respectively.

Muscarinic M1 receptor agonists are useful for the treatment of Alzheimer's disease by improving the performance in cognitive tests in Alzheimer's patients.[32] NMDA receptor antagonists have been explored for neuroprotection[33] and the treatment of postoperative pain.[34] Thrombin inhibitors produce anticoagulant effects and have been used as antithrombotic agents.[35] HIV protease inhibitors form an important class of anti-HIV agents some of which have been successfully used clinically.[36] Cephalosporins are in clinical development as broad-spectrum antibacterial agents.[37] Renin inhibitors have shown effectiveness in cardiovascular pharmacotherapy.[38] Because of their diverse therapeutic applications and structural frameworks, these compounds are highly useful for testing the performance of SVM as well as other methods.[28]

For each of the biological target class, two training data sets were generated. A regularly sparse active data set, which contains the same number of actives as those in earlier sparse data set studies,[28,30] was generated by extracting 100 actives (representing 6.1%−10.2% of the known actives) scattered in the known active regions of chemical space. A very sparse active data set was generated by extracting 40 active compounds (representing 2.4%−4.1% of the known actives) scattered in the known active regions of chemical space. To generate a data set of N number of actives from a larger number actives, all actives were clustered into N clusters followed by the extraction of one compound from each of these clusters. Putative inactive data sets were generated by extracting representative compounds from all compound families that contain no known active compound.[5] Compound families can be generated by clustering distinct compounds of chemical databases into groups of similar structural and physicochemical properties.[25,39]

The regularly sparse active data sets were used for facilitating crude estimation of the level of performance of our SVM VS tools with respect to those of other VS tools such as the data fusion method[28] and other methods[30] that have been frequently developed by using ∼100 active compounds. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the data sets and molecular descriptors used. To further evaluate whether the performance of our SVM VS tools are attributed to the SVM classification models or the molecular descriptors used, a study was conducted to compare the performance our SVM VS tools with that of the Tanimoto-based similarity searching method[40] using the same data sets and the same molecular descriptors.

The yields (percent of testing actives identified as active) of our SVM VS tools were estimated by using the remaining 89.7%−97.4% of the known actives. The false-hit rates (percent of inactives identified as active) of our SVM VS tools were estimated by using the remaining 167K MDDR compounds outside the training data sets and by using the 9.997 M PUBCHEM compounds that exclude the known actives. To further evaluate whether our SVM VS tools predict active and inactive compounds rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

VS performance may be overestimated by training data sets that contain higher percentages of inactives significantly different from the known actives, because the easily distinguishable features may make VS enrichments appearing artificially good.[41] Therefore, VS performance may be more strictly tested by using subsets of inactives that resemble the physicochemical properties of the known actives so that enrichment is not simply a separation of trivial physico-

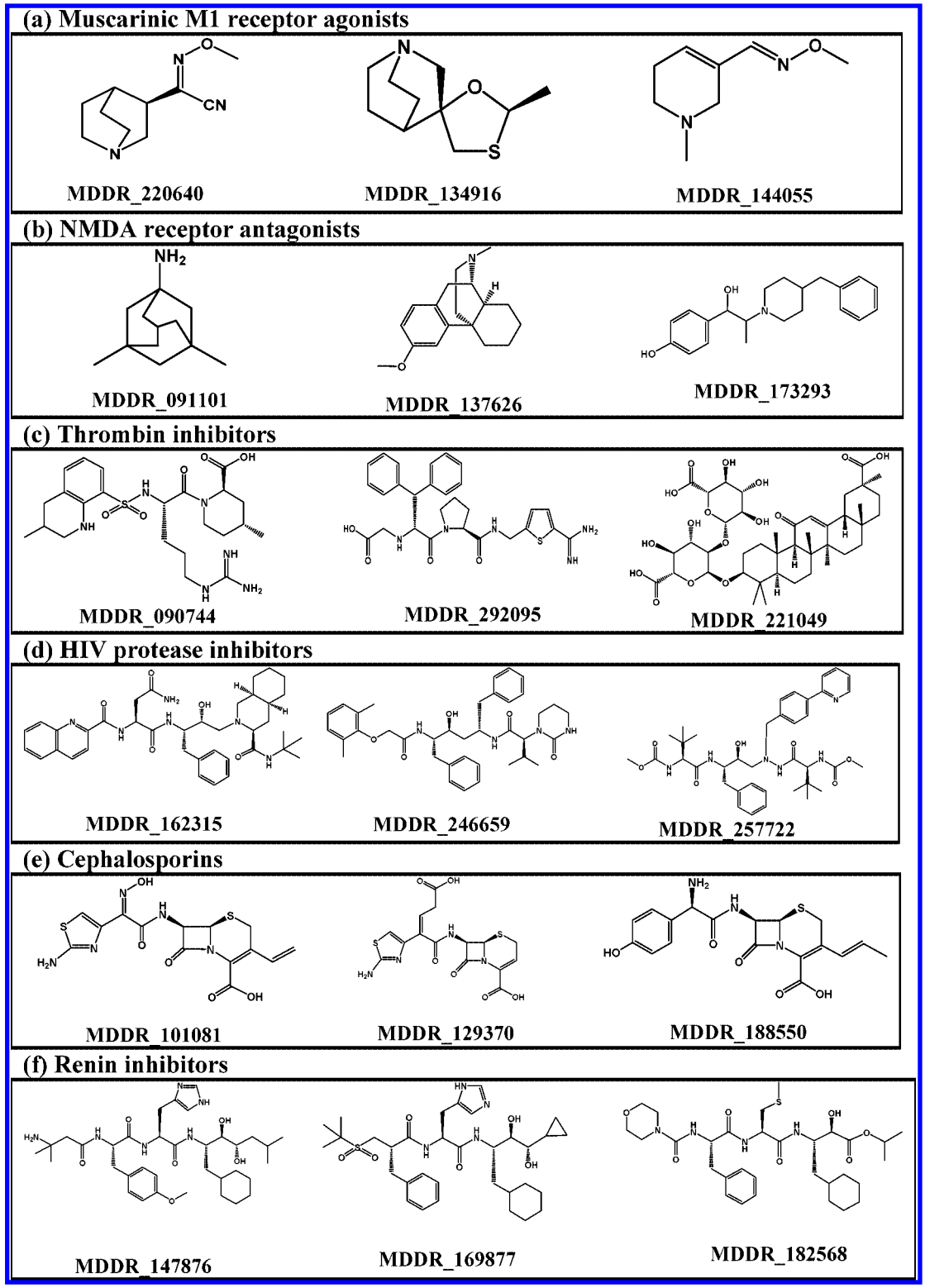

Figure 2. Structure of the selected muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors. The PUBCHEM accession number of these compounds is given.

chemical features.[42] In this work, the performance of our SVM VS tools was further evaluated by the subsets of MDDR compounds that are similar in physicochemical properties to those of the known actives.

METHODS

Construction of Active Training and Testing Data Sets. All actives of the six biological target classes are from MDDR, from which we obtained 983 muscarinic M1 receptor agonists, 1510 NMDA receptor antagonists, 1252 thrombin inhibitors, 1054 HIV protease inhibitors, 1645 cephalosporins, and 1241 renin inhibitors. The structure of representative compounds of these six classes is shown in Figure 2. To generate the popular-sized sparse and highly sparse active training data sets and the corresponding testing data sets, all known actives of each of these classes were clustered into 100 and 40 clusters, respectively, by using a K-means method[25,39] and molecular descriptors computed

from our own software.[43] For each class, the regularly sparse and very sparse active training data sets of 100 and 40 active compounds were generated by extracting one compound from each of the 100 and 40 active clusters, respectively. The remaining actives were used as the corresponding active testing set.

Generation of Putative Inactive Training and Testing Data Sets. Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds,[6,7,9–11,28,29,31] a new approach extensively used for generating inactive proteins in SVM classification of various functional classes of proteins[44–46] has recently been applied for generating putative inactive compounds.[5] An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the "inactive" chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. A drawback of this approach is the possible inclusion of some yet-to-be-discovered active compounds in the "inactive" class, which may affect the capability of SVM for identifying novel active compounds. As has been demonstrated in an earlier study[5] and in this work, such an adverse effect is expected to be relatively small for many biological target classes.

In applying this approach to proteins, all known proteins are clustered into ~8933 protein domain families based on the clustering of their amino acid sequences,[47] and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. Undiscovered active proteins of a specific functional class typically cover no more than a few hundred families, which gives a maximum possible "wrong" family representation rate of <10.2% even when all of the undiscovered active proteins are misplaced into the inactive class.[48] Importantly, inclusion of the representative of a "wrong" family into the inactive class does not preclude other active family members from being classified as active. Statistically, a substantial percentage of active members can be classified by ML methods as active even if its family representative is in the inactive class.[5,48] Therefore, in principle, a reasonably good SVM classification model can be derived from these putative inactive samples, which has been confirmed by a number of studies of proteins.[44–46,48]

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors.[25,39] As SVM predict compound activities based on their molecular descriptors, in developing SVM VS tools, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a K-means method[25,39] and molecular descriptors computed from our own software,[43] we generated 8993 compound families from the 9.974 M compounds in the PUBCHEM and MDDR databases that we were able to compute the molecular descriptors, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms,[49] and the 2851 clusters for 171,045 natural products.[50]

Analogue groups such as steroids and catecholamines are distributed in a few families. The classes of muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors are distributed in 203, 538, 161, 281, 95, and 138 families, respectively. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered "active" families in PUBCHEM database is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered "active" families (hundreds) and the families that contain no known active compound (~8993 based on the current versions of PUBCHEM and MDDR) for these and possibly many other target classes is expected to be <15%. Therefore, putative inactive training data sets can be generated by extracting a few representative compounds of those families that contain no known active compound in the active training set, with a maximum possible "wrong" family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class and with the expectation that a substantial percentage of active members in the putative "inactive" families can be classified as active despite their family representatives being placed into the inactive training sets. As has been shown in a recent study of SVM VS tools, a substantial percentage of identified virtual hits are from these "inactive" families.[5]

There are 8790, 8455, 8832, 8712, 8898, and 8855 compound families that contain no known muscarinic M1 receptor agonist, NMDA receptor antagonist, thrombin inhibitor, HIV protease inhibitor, cephalosporin, and renin inhibitor, respectively. Thus the inactive training data set corresponding to each sparse or biased active training data set was generated by random selection of 5−6 representative compounds from each of these "inactive" families and those active families with none of their members in the active training set. The remaining compounds of the "inactive" families in PUBCHEM and MDDR can be used as putative inactive testing sets. It is noted that 9.6%−68.7% of the active containing families are not covered in the active training set, and their representative compounds were deliberately placed into the inactive training set as they are not supposed to be known in our study. As shown in an earlier study[5] and in this work, a substantial percentage of the active compounds in these misplaced active containing families were predicted as active by our SVM models. Moreover, a small percentage of the compounds in these putative inactive data sets are expected to be unreported and undiscovered actives for each of the six biological target classes, and their presence in these data sets is not expected to significantly affect the estimated false positive rate of the developed SVM VS tools.

Molecular Descriptors. Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure−activity relationships,[51,52] quantitative structure−activity relationships,[53,54] and ML prediction models for pharmaceutical agents.[55–62] A total of 98 1D and 2D descriptors derived by using our software[43] were used in this work. These descriptors and the relevant references are given in Table 1, which include 18

Table 1. Molecular Descriptors Used in This Work

descriptor class	no. of descriptors in class	descriptors
simple molecular properties[81]	18	no. of C,N,O,P,S, no. of atoms, no. of rings, no. of bonds, no. of non-H bonds, molecular weight, no. of rotatable bonds, no. of H-bond donors, no. of H-bond acceptors, no. of 5-member aromatic rings, no. of 6-member aromatic rings, no. of N heterocyclic rings, no. of O heterocyclic rings, no. of S heterocyclic rings
chemical properties[82]	3	Sanderson electronegativity, molecular polarizability, aLogp
molecular connectivity and shape [81,83]	35	Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, gravitational topological index, molecular path count of length 1−6, total path count, Balaban index J, 0−2th valence connectivity index, 0−2th order delta chi index, Pogliani index, 0−2th solvation connectivity index, 1−3th order Kier shape index, 1−3th order kappa alpha shape index, Kier molecular flexibility index, topological radius, graph-theoretical shape coefficient, eccentricity, centralization, Logp from connectivity
electrotopological state[81,84]	42	sum of Estate of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; sum of Estate of all heavy atoms, all C atoms, all heteroatoms; sum of Estate of H-bond acceptors; sum of H Estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin; sum of H Estate of H-bond donors

descriptors in the class of simple molecular properties, 3 descriptors in the class of chemical properties, 35 descriptors in the class of molecular connectivity and shape, and 42 descriptors in the class of electrotopological state.

Support Vector Machines Method. The process of training a SVM prediction model and using it for predicting active compounds of a compound class from their molecular descriptors is schematically illustrated in Figure 3. The theory of SVM has been extensively described in the literature.[63,64] Thus only a brief description is given here. SVM is based on the structural risk minimization principle from statistical learning theory,[63] which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for overfitting.[65,66] In linearly separable cases, SVM constructs a hyperplane to separate active and inactive classes of compounds with a maximum margin. A compound is represented by a vector $\boldsymbol{x}_i$ composed of its molecular descriptors. The hyperplane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \quad \text{Class 1 (active)} \tag{1}$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \quad \text{Class 2 (inactive)} \tag{2}$$

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\boldsymbol{x}$ can be classified by

$$f(\boldsymbol{x}) = \text{sign}[(\mathbf{w}\cdot\mathbf{x}) + b] \tag{3}$$

A positive or negative $f(\mathbf{x})$ value indicates that the vector $\mathbf{x}$ belongs to the active or inactive class, respectively.

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures,[6,7,9,28,31,61,67,68] SVM maps the input vectors into a higher dimensional feature space by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. The kernel function used in this study is the RBF kernel, which has been extensively used and consistently shown better performance than other kernel functions.[69–71]

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-||\mathbf{x}_j - \mathbf{x}_i||^2/2\sigma^2} \tag{4}$$

Linear SVM can then applied to this feature space based on the following decision function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b\right) \tag{5}$$

where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

under the following conditions:

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{7}$$

A positive or negative $f(\boldsymbol{x})$ value indicates that the vector $\mathbf{x}$ belongs to the active or inactive class, respectively.

In developing our SVM VS tools, a hard margin c = 100,000 was used, and the σ values were found to be in the range of 1−1.5, 0.5−1, 0.5−1.5, 1−2, 3−4, and 1−2.5 for the muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors, respectively. In terms of true positives TP (true actives), true negatives TN (true inactives), false positives FP (false actives), and false negatives FN (false inactives), the yields and false-hit rate are given by TP/(TP+FN) and FP/(TP+FP), respectively.

Tanimoto-Based Similarity Searching Method. Compounds similar to at least one active in an active data set can be identified by using the Tanimoto coefficient sim(i,j)[40]

$$\mathrm{sim}(i,j) = \frac{\sum_{d=l}^{l} x_{di}x_{dj}}{\sum_{d=1}^{1} (x_{di})^2 + \sum_{d=1}^{1} (x_{dj})^2 - \sum_{d=l}^{l} x_{di}x_{dj}} \quad (8)$$

where l is the number of molecular descriptors. A compound i is considered to be similar to a known active j in the active data set if the corresponding sim(i,j) value is greater than a cutoff value. In this work, the similarity search was conducted for MDDR compounds. Therefore, in computing sim(i,j), the molecular descriptor vectors $\mathbf{x}_i$s were scaled with respect to all of the MDDR compounds. The cutoff values for similarity compounds are typically in the range of 0.8−0.9.[42,72] A stricter cutoff value of 0.9 was used in this study.

RESULTS AND DISCUSSION

Comparative Analysis of Virtual Screening Performance of SVM Trained by Regularly Sparse Active Data Sets. It is of interest to evaluate the performance of SVM trained from regularly sparse active data sets by comparison with literature reported VS performance based on similar data set construction/testing procedures and the same data sources. As discussed in the Introduction section, the comparison of these results should be viewed as providing very crude pictures about the level of performance of SVM. In this work, we specifically compared the performance of SVM VS tools with those of a standard similarity-based method, the data fusion method.[28,73] The data fusion method is based on Tanimoto-based similarity searching using multiple reference compounds, which have shown good performances for a number of active compound groups by using only a small number of ∼100 training active compounds,[28,73] which serves as a good reference method for evaluating the performance of SVM. To further evaluate whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of the Tanimoto-based similarity searching method based on the same training and testing data sets and molecular descriptors.

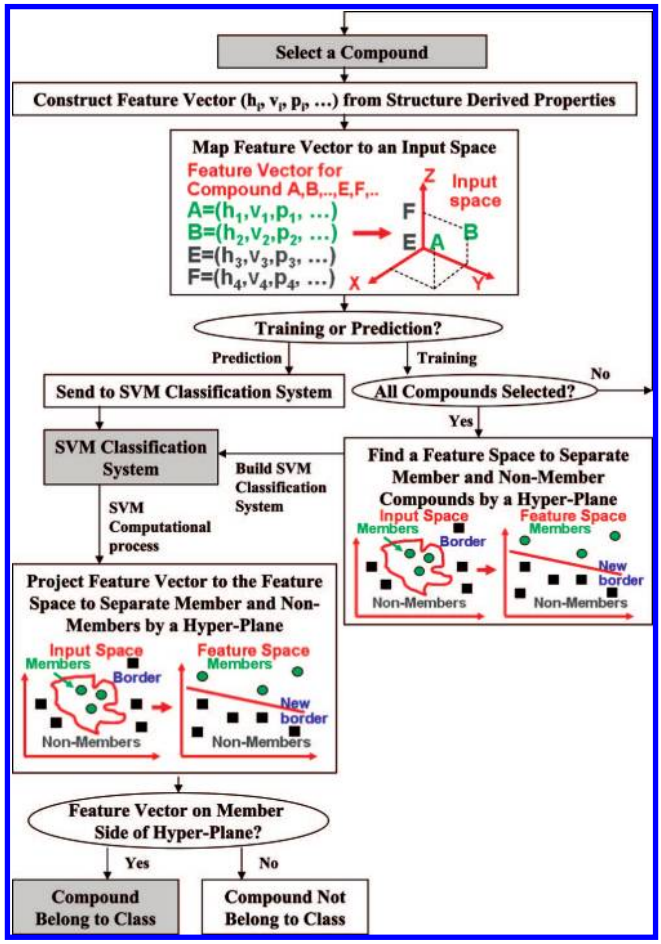

Figure 3. Schematic diagram illustrating the process of the training of a prediction model and using it for predicting active compounds of a compound class from their structurally derived properties (molecular descriptors) by using support vector machines. A, B, E, F, and (h_j, p_j, vj, . . .) represent such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

The statistics of the regularly sparse active data sets, the performance of our method, the reported performance of the data fusion method, and the results of the Tanimoto-based similarity searching method for the six biological classes are given in Table 2. As shown in Table 2, the percentage of known actives in these data sets is in the range of 6.1%−10.2%. The percentage of "active" families (defined as the families that include at least one known active compound) covered by these data sets is in the range of 15.4%−67.2% with five of the sets below 31.5%. Therefore, these data sets are reasonably sparse.

By using the same testing procedure of the data fusion method, the performance of the six developed SVM VS tools were evaluated by using the remaining 883−1545 actives and ∼167K MDDR compounds of other biological target classes. The yields of our SVM VS tools are 26.7%−49.5% for the high, 60.0%−67.3% for the intermediate, and 82.1%−91.9% for the low diversity classes, respectively. The reported yields of the data fusion method are 15.7%−46.6% for the high, 44.5%−58.0% for the intermediate, and 90.4%−94.7% for the low diversity classes, respectively.[28,73] The false-hit rates (estimated from the percentage of the ∼167K MDDR compounds of other biological target classes identified as active) of our SVM VS tools are in the range of 1.0%−2.9%. The false-hit rates of data fusion method can be deduced as 4% based on the reported top 5% hit selection criterion from ∼150K compounds of other MDDR biological target classes.[28,73] Compared with those of the data fusion method, the yields of our SVM VS tools are slightly improved for the high and intermediate classes, and the false-hit rates of our SVM VS tools are substantially reduced for all three classes. These results suggest that, by using the equally small number of active compounds as training data, SVM is capable of producing equally good or slightly better yields and generalization capability at substantially reduced false-hit rates than those of the data fusion method.

As shown in Table 2, the yields of the Tanimoto-based similarity searching method are 9.4%−24.2% for the high, 19.0%−27.8% for the intermediate, and 38.4%−39.3% for the low diversity classes, respectively. The false-hit rates are in the range of 3.3%−4.4%. Compared to these results, the yields of SVM are significantly improved, and the false-hit rates of SVM are substantially reduced. This suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used.

Table 2. Data Set Statistics and the Virtual Screening Performance of Support Vector Machines Developed by using Regularly Sparse Data Sets of 100 Active Compounds for screening MDDR Database[a]

compd diversity category defined in ref 28	compd biological target class (no. of compds) [av mean pairwise similarity value computed in ref 28]	active compds in training set: no. and % of active compds	active compds in training set: no. and % of known "active" chemical families covered by active compds	active compds in testing set: no. and % of active compds	active compds in testing set: no. and % of known "active" chemical families covered by active compds	SVM virtual screening performance (this work): yields (%)	SVM virtual screening performance (this work): false hit rates (%)	virtual screening performance of similarity searching methods reported in ref 28: yields (%)	virtual screening performance of similarity searching methods reported in ref 28: false hit rates (%)	virtual screening performance of Tanimoto-based similarity searching method (this work): yields (%)	virtual screening performance of Tanimoto-based similarity searching method (this work): false hit rates (%)
high	Muscarinic M1 receptor agonists (983) [0.206]	100 (10.2%)	64 (31.5%)	883 (89.8%)	171 (84.2%)	49.5	1.7	27.4−46.6%	4	24.2	3.9
	NMDA receptor antagonists (1510) [0.199]	100 (6.6%)	83 (15.4%)	1410 (93.4%)	503 (93.5%)	26.7	2.8	15.7−20.7%	4	9.4	4.4
intermediate	Thrombin inhibitors (1252) [0.321]	100 (8.0%)	46 (28.6%)	1152 (92.0%)	227 (91.7%)	60.0	2.9	44.5−52.3%	4	19.0	4.3
	HIV protease inhibitors (1054) [0.313]	100 (9.5%)	74 (26.3%)	954 (90.5%)	248 (88.3%)	67.3	2.9	51.6−58.0%	4	27.8	4.4
low	Cephalosporins (1645) [0.501]	100 (6.1%)	43 (67.2%)	1545 (93.9%)	78 (82.5%)	82.1	1.0	NA	NA	39.3	3.7
	Rennin inhibitors (1241) [0.459]	100 (8.1%)	51 (37.0%)	1141 (91.9%)	121 (87.7%)	90.9	1.8	90.4−94.7%	4	38.4	3.3

[a] The results are compared with that of the Tanimoto similarity searching method using the same data set and molecular descriptors and with the reported performance of similarity search methods trained by using ∼100 active compounds (ref 28) for identifying muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors. Known "active" chemical families refer to chemical families that contain at least one known active compound. Yields and false hit rates are the percent of testing active compounds identified as active.

Table 3. Data Set Statistics and Virtual Screening Performance of Support Vector Machines Developed by using Very Sparse Active Data Sets of 40 Active Compounds for Identifying Muscarinic M1 Receptor Agonists, NMDA Receptor Antagonists, Thrombin Inhibitors, HIV Protease Inhibitors, Cephalosporins, and Rennin Inhibitors from PUBCHEM and MDDR Databases[a]

compd diversity category defined in ref 28	compd biological target class (no of compds) [average mean pairwise similarity value computed in ref 28]	active compds in training set: no. and % of active compds	active compds in training set: no. and % of known "active" chemical families covered by active compds	active compds in testing set: no. and % of active compds	active compds in testing set: no. and % of known "active" chemical families covered by active compds	virtual screening performance: yields (%)	virtual screening performance: no. and % of identified testing active compds outside training chemical families	virtual screening performance: no. and % of 9.997 M PUBCHEM compds identified as active	virtual screening performance: no. and % of the remaining 167K MDDR compds as active
high	Muscarinic M1 receptor agonists (983) [0.206]	40 (4.1%)	34 (16.7%)	943 (95.9%)	191 (94.1%)	39.5	149 (40.1%)	1,130 (0.01%)	1,618 (1.0%)
	NMDA receptor antagonists (1510) [0.199]	40 (2.7%)	36 (6.7%)	1470 (97.3%)	524 (97.4%)	17.5	165 (64.2%)	2,336 (0.02%)	2,001 (1.2%)
intermediate	Thrombin & inhibitors (1252) [0.321]	40 (3.2%)	25 (15.5%)	1212 (96.8%)	237 (96.0%)	23.0	102 (57.0%)	529 (0.005%)	1,198 (0.7%)
	HIV protease inhibitors (1054) [0.313]	40 (3.8%)	36 (12.8%)	1014 (96.2%)	269 (95.7%)	48.1	301 (68.7%)	530 (0.005%)	2,658 (1.6%)
low	Cephalosporins (1645) [0.501]	40 (2.4%)	27 (42.2%)	1605 (97.6%)	86 (89.7%)	92.4	205 (13.8%)	770 (0.007%)	791 (0.5%)
	Rennin inhibitors (1241) [0.459]	40 (3.2%)	31 (22.5%)	1201 (96.8%)	130 (94.2%)	70.2	410 (48.6%)	398 (0.004%)	2,220 (1.3%)

[a] Known "active" chemical families refer to chemical families that contain at least one known active compound.

Table 4. Evaluation of Support Vector Machines Virtual Screening Tools for Identifying Muscarinic M1 Receptor Agonists, NMDA Receptor Antagonists, Thrombin Inhibitors, HIV Protease Inhibitors, Cephalosporins, and Renin Inhibitors against the Subset of Inactive MDDR Compounds that Are Similar to at Least One Known Active Compound in Each Respective Active Compound Class[a]

				SVM virtual screening performance	
compd diversity category defined in ref 28	compd biological target class (no. of compds)	no. and % of active compds in training set	no. of inactive compds similar to an active compd (testing set)	no. of inactive compds predicted as active	false hit rate (%)
high	Muscarinic M1 receptor agonists (983)	40 (4.1%)	19,495	531	4.4
		100 (10.2%)		1068	7.8
	NMDA receptor antagonists (1510)	40 (2.7%)	38,436	729	2.6
		100(6.6%)		1349	4.6
intermediate	Thrombin inhibitors (1252)	40 (3.2%)	32,037	1535	5.7
		100 (8.0%)		1267	6.4
	HIV protease inhibitors (1054)	40 (3.8%)	29,990	603	3.3
		100 (9.5%)		1398	6.4
low	Cephalosporins (1645)	40 (2.4%)	29,127	181	5.8
		100 (6.1%)		612	7.6
	Rennin inhibitors (1241)	40 (3.2%)	24,166	637	6.3
		100 (8.1%)		887	8.3

[a] Similarity is defined by Tanimoto coefficient ≥ 0.9 to any known active in a full active data set, which is computed by using molecular descriptors. The yields are given in Tables 2 and 3, respectively.

Virtual Screening Performance of SVM Trained by Very Sparse Active Data Sets. The level of sparseness of the very sparse active data sets for the six biological target classes can be measured by the percentage of known actives in these training sets and the percentage of "active" families they occupy. As shown in Table 3, the percentage of known actives in the sparse active training sets is in the range of 2.4%−4.7%. The percentage of "active" families covered by the sparse active training sets is in the range of 6.7%−42.2% with five of these below 22.5%. Therefore, the level of sparseness of the very sparse active data sets is significantly higher than those of the regularly sparse active data sets.

The SVM VS tools developed by using the very sparse active data sets for identifying active compounds of the six biological target classes were tested by using three testing sets for each compound class. These testing sets are the active testing set for each class, 9.98 million distinct compounds from the PUBCHEM, and the remaining 167K MDDR compounds outside the training sets of our developed SVM models. The performance of these SVM VS tools is given in Table 3. In spite of the use of very sparse active training sets of <4.7% of the actives covering 6.7%−42.2% of the "active" families, our SVM VS tools were able to achieve yields of 17.5%−35.5% for the high, 23.0%−48.1% for the intermediate, and 70.2%−92.4% for the low diversity classes. Therefore, our method appears to have some level of generalization capability in identifying a substantial amount of novel active compounds outside the known active chemical families from a very sparse active training data set.

In addition to the exhibition of effective hit selection performance, our SVM models appear to show substantially lower false-hit rates. In screening 9.997 M PUBCHEM compounds that exclude the known actives, without using top-ranked cutoff or additional filter, our SVM VS tools identified 398−2336 compounds as active, representing 0.004%−0.01% of the 9.997 M PUBCHEM compounds. The estimated false-hit rates in screening 167K MDDR compounds of the other biological classes are in the range of 0.5%−1.6%. Even though a substantially larger number of compounds (9.997 M vs 98.4K−2.5M) were screened, these false-hit rates are comparable and in many cases better than those of 0.08%−3% by SBVS tools,[12–23] 0.1%−5% by other reported ML models,[7,9–11,28] 0.16%−82.% by clustering methods,[25] and 1.15%−26% by pharmacophore models.[26,27,74,75]

Evaluation of False-Hit Rates of SVM against Inactives of Similar Molecular Descriptors to the Known Actives. The subsets of MDDR compounds that are similar in molecular descriptors to at least one known active of the six biological target classes were selected by using the condition that the Tanimoto coefficient sim(i,j) is ≥ 0.9 with respect to at least one known active of each of these classes. A total of 19,495, 38,436, 32,037, 29,990, 29,127, and 24,166 inactives of similar molecular descriptors were collected for the muscarinic M1 receptor agonist, NMDA receptor antagonist, thrombin inhibitor, HIV protease inhibitor, cephalosporin, and renin inhibitor classes, respectively. Each of these six sets of inactives were used as the testing sets for evaluating the false-hit rates of our developed SVM VS tools against similarity compounds.

As shown in Table 4, against these similarity data sets, the false-hit rates of our SVM VS tools developed by using regularly sparse and very sparse active data sets are in the range of 4.6%−8.3% and 2.6%−6.3%, respectively. Compared to the ranges of hit rates of 1.0%−2.9% and 0.5%−1.6% against the full set of the ~167K MDDR compounds of other biological target classes, our developed SVM VS tools appear to show a fairly good performance in distinguishing the actives from the inactives that resemble the physicochemical properties of the known actives.

Evaluation of SVM Identified False Hits. Some of the false hits are known inhibitors that share structural frameworks with those of the studied biological target class. For instance, a number of SVM identified "false" hits of HIV protease inhibitors are known renin inhibitors. Some of the HIV protease inhibitors have been designed based on the transition state analogues of renin inhibitors.[76] Many

of the SVM identified false hits of thrombin inhibitors are known peptidomimetic inhibitors of renin, HIV protease, farnesyltransferase, and trypsin. Peptidomimetic inhibitors arising from similar structural frameworks have been designed for renin, thrombin, HIV protease, Ras farnesyltransferase, and various other proteases.[77] Therefore, some of the false hits may partly arise from the misidentification of compounds of similar structural frameworks. It cannot be ruled out that some of them may exhibit weak inhibitory activities due to the similar structural frameworks and thus were "correctly" identified by our SVM VS tools.

Examination of the false hits identified by SVM and other machine learning methods consistently suggests that the currently used molecular descriptors are insufficient to adequately represent some of the compounds that contain complex structural or chemical configurations.[59,61,78] Examples of these agents are those with a large rigid structure combined with a short flexible hydrophilic tail, compounds that contain multirings with various heteroatoms such as nitrogen, oxygen, sulfur, fluorine, and chlorine. Due to the limited coverage of the number of bond links in a heteroatom loop, the currently available topological descriptors are not yet capable of describing the special features of a complex multiring structure that contains multiple heteroatoms. It appears that none of the currently available descriptors are capable of fully representing molecules containing a long flexible chain. Therefore, it might be helpful to explore different combinations of descriptors and to select more optimal set of descriptors by using more refined feature selection algorithms and parameters.[59,79] However, the indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each other, may introduce noise as well as extending the coverage of some of the aspects of these special features. Thus, it may be necessary to introduce new descriptors for more appropriately representing these and other special features.

Does SVM Select Active Compounds or Membership of Compound Families? To further evaluate whether our SVM VS tools identify active compounds rather than membership of certain compound families, compound family distributions of the identified actives and inactives for the six biological target classes were analyzed. As shown in Table 3, 40.1%, 64.2%, 57.0%, 68.7%, 13.8%, and 48.6% of the identified muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and renin inhibitors belong to the families that contain no known active. For those families that contain at least one known active, >70% of the compounds (>90% in majority cases) in each of these families were predicted as inactive by our SVM VS tools. These results suggest that our SVM VS tools identify active compounds rather than membership to certain compound families. Some of the identified actives not in the family of known active compounds may serve as potential "novel" active compounds. Therefore, as in the case shown by earlier studies,[5,80] SVM has a certain capacity for identifying novel active compounds from sparse as well as regular-sized active data sets.

CONCLUDING REMARKS

SVM VS tools developed by using highly sparse active data sets show some level of capability in identifying novel active compounds at comparable and in many cases substantially lower false-hit rates than those of typical SBVS and LBVS tools reported in the literature. The performance of SVM is significantly better than that of the Tanimoto-based similarity search method based on the same data sets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of the integrated VS tools in facilitating lead discovery from sparse active data sets.[16,20,75]

REFERENCES AND NOTES

(1) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862–865.
(2) Lengauer, T.; Lemmen, C. et al. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9* (1), 27–34.
(3) Davies, J. W.; Glick, M. et al. Streamlining lead discovery by aligning in silico and high-throughput screening. *Curr. Opin. Chem. Biol.* **2006**, *10* (4), 343–351.
(4) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23−24), 1046–1053.
(5) Han, L. Y.; Ma, X. H. et al. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor J. Mol. Graphics Modell. 2007, accepted for publication.
(6) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45* (3), 549–561.
(7) Glick, M.; Jenkins, J. L. et al. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46* (1), 193–200.
(8) Li, H.; Ung, C. Y. et al. Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. *J. Mol. Graphics Modell.* **2006**, *25* (3), 313–323.
(9) Lepp, Z.; Kinoshita, T. et al. Screening for new antidepressant leads of multiple activities by support vector machines. *J. Chem. Inf. Model.* **2006**, *46* (1), 158–167.
(10) Chen, B.; Harrison, R. F. et al. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21* (1−3), 53–62.
(11) Franke, L.; Byvatov, E. et al. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48* (22), 6997–7004.
(12) Ghosh, S.; Nie, A. et al. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.* **2006**, *10* (3), 194–202.
(13) Shoichet, B. K.; McGovern, S. L. et al. Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **2002**, *6* (4), 439–446.
(14) Jansen, J. M.; Martin, E. J. Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 359–364.
(15) Mozziconacci, J. C.; Arnoult, E. et al. Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. *J. Med. Chem.* **2005**, *48* (4), 1055–1068.
(16) Vidal, D.; Thormann, M. et al. A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model.* **2006**, *46* (2), 836–843.
(17) Cummings, M. D.; DesJarlais, R. L. et al. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48* (4), 962–976.
(18) Evers, A.; Klabunde, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J. Med. Chem.* **2005**, *48* (4), 1088–1097.

(19) Lorber, D. M.; Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Curr. Top. Med. Chem.* **2005**, *5* (8), 739–749.
(20) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (2), 587–596.
(21) Vangrevelinghe, E.; Zimmermann, K. et al. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46* (13), 2656–2662.
(22) Doman, T. N.; McGovern, S. L. et al. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **2002**, *45* (11), 2213–2221.
(23) Enyedy, I. J.; Ling, Y. et al. Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J. Med. Chem.* **2001**, *44* (25), 4313–4324.
(24) Oprea, T. I.; Matter, H. Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.* **2004**, *8* (4), 349–358.
(25) Bocker, A.; Schneider, G. et al. NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2220–2229.
(26) Schuster, D.; Maurer, E. M. et al. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J. Med. Chem.* **2006**, *49* (12), 3454–3466.
(27) Steindl, T.; Laggner, C. et al. Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening. *J. Chem. Inf. Model.* **2005**, *45* (3), 716–724.
(28) Hert, J.; Willett, P. et al. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46* (2), 462–470.
(29) Harper, G.; Bradshaw, J. et al. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1295–1300.
(30) Li, H.; Yap, C. W. et al. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J. Pharm. Sci.* **2007**, *96* (11), 2838–2860.
(31) Cui, J.; Han, L. Y.; Lin, H. H.; Zhang, H. L.; Tang, Z. Q.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties. *Mol. Immunol.* **2007**, *44*, 866–877.
(32) Clader, J. W.; Wang, Y. Muscarinic receptor agonists and antagonists in the treatment of Alzheimer's disease. *Curr. Pharm. Des.* **2005**, *11* (26), 3353–3361.
(33) Lipton, S. A. Pathologically-activated therapeutics for neuroprotection: mechanism of NMDA receptor block by memantine and S-nitrosylation. *Curr. Drug. Targets* **2007**, *8* (5), 621–632.
(34) De Kock, M. F.; Lavand'homme, P. M. The clinical role of NMDA receptor antagonists for the treatment of postoperative pain. *Best Pract. Res. Clin. Anaesthesiol.* **2007**, *21* (1), 85–98.
(35) Lepor, N. E. Anticoagulation for acute coronary syndromes: from heparin to direct thrombin inhibitors. *Rev. Cardiovasc. Med.* **2007**, *8* (3), S9–17.
(36) Spaltenstein, A.; Kazmierski, W. M. et al. Discovery of next generation inhibitors of HIV protease. *Curr. Top. Med. Chem.* **2005**, *5* (16), 1589–1607.
(37) Page, M. G. Emerging cephalosporins. *Expert Opin. Emerging Drugs* **2007**, *12* (4), 511–524.
(38) Sepehrdad, R.; Frishman, W. H. et al. Direct inhibition of renin as a cardiovascular pharmacotherapy: focus on aliskiren. *Cardiol. Rev.* **2007**, *15* (5), 242–256.
(39) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157–166.
(40) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(41) Verdonk, M. L.; Berdini, V. et al. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793–806.
(42) Huang, N.; Shoichet, B. K. et al. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.
(43) Xue, Y.; Yap, C. W. et al. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497–1505.
(44) Cai, C. Z.; Han, L. Y. et al. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003**, *31* (13), 3692–3697.
(45) Han, L. Y.; Cai, C. Z. et al. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* **2004**, *32* (21), 6437–6444.
(46) Lin, H. H.; Han, L. Y. et al. Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* **2006**, *62* (1), 218–231.
(47) Wegner, J. K. *JOELib/JOELib2*; Department of Computer Science, University of Tübingen: Germany, 2005.
(48) Han, L. Y.; Zheng, C. J. et al. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discovery Today* **2007**, *12* (7−8), 304–313.
(49) Fink, T.; Reymond, J. L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery J. Chem. Inf. Model. 2007 (published on Web 01/30/2007).
(50) Koch, M. A.; Schuffenhauer, A. et al. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (48), 17272–17277.
(51) Fang, H.; Tong, W. et al. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **2001**, *14*, 280–294.
(52) Tong, W.; Xie, Q. et al. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112* (12), 1249–1254.
(53) Jacobs, M. N. In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* **2004**, *205* (1−2), 43–53.
(54) Hu, J. Y.; Aizawa, T. Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. *Water Res.* **2003**, *37* (6), 1213–1222.
(55) Byvatov, E.; Fechner, U. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1882–1889.
(56) Doniger, S.; Hofman, T. et al. Predicting CNS Permeability of Drug Molecules:Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.* **2002**, *9* (6), 849–864.
(57) He, L.; Jurs, P. C. et al. Predicting the Genotoxicity of Polycyclic Aromatic Compounds from Molecular Structure with Different Classifiers. *Chem. Res. Toxicol.* **2003**, *16* (12), 1567–1580.
(58) Snyder, R. D.; Pearl, G. S. et al. Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen.* **2004**, *43* (3), 143–158.
(59) Xue, Y.; Li, Z. R. et al. Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1630–1638.
(60) Yap, C. W.; Cai, C. Z. et al. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci.* **2004**, *79* (1), 170–177.
(61) Yap, C. W.; Chen, Y. Z. Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci.* **2005**, *94* (1), 153–168.
(62) Zernov, V. V.; Balakin, K. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2048–2056.
(63) Vapnik, V. N. The nature of statistical learning theory In *The nature of statistical learning theory*; Springer: New York, 1995.
(64) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery* **1998**, *2* (2), 127–167.
(65) Pochet, N.; De Smet, F. et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **2004**, *20*, 3185–3195.
(66) Li, F.; Yang, Y. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **2005**, *21*, 3741–3747.
(67) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45* (4), 982–992.
(68) Grover, I. I.; Singh, I. I. et al. Quantitative structure-property relationships in pharmaceutical research - Part 2. *Pharm. Sci. Technol. Today* **2000**, *3* (2), 50–57.
(69) Trotter, M. W. B.; Buxton, B. F. et al. Support vector machines in combinatorial chemistry. *Meas. Control.* **2001**, *34* (8), 235–239.
(70) Burbidge, R.; Trotter, M. et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26* (1), 5–14.
(71) Czerminski, R.; Yasri, A. et al. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20* (3), 227–240.
(72) Bostrom, J.; Hogner, A. et al. Do structurally similar ligands bind in a similar fashion. *J. Med. Chem.* **2006**, *49* (23), 6716–6725.
(73) Whittle, M.; Gillet, V. J. et al. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *46* (6), 2206–2219.

(74) Pirard, B.; Brendel, J. et al. The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. *J. Chem. Inf. Model.* **2005**, *45* (2), 477–485.
(75) Rella, M.; Rushworth, C. A. et al. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model.* **2006**, *46* (2), 708–716.
(76) Eder, J.; Hommel, U. et al. Aspartic proteases in drug discovery. *Curr. Pharm. Des.* **2007**, *13* (3), 271–285.
(77) Ripka, A. S.; Rich, D. H. Peptidomimetic design. *Curr. Opin. Chem. Biol.* **1998**, *2* (4), 441–452.
(78) Li, H.; Ung, C. et al. Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem. Res. Toxicol.* **2005**, *18* (6), 1071–1080.
(79) Li, H.; Yap, C. W. et al. Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.* **2005**, *45* (5), 1376–1384.
(80) Ung, C. Y.; Li, H. et al. In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol. Pharmacol.* **2007**, *71* (1), 158–168.
(81) Todeschini, R.; Consonni, V. Handbook of Molecular DescriptorsIn *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
(82) Miller, K. J. Additive Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.
(83) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.
(84) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.

CI800022E