SVM Model for Virtual Screening of Lck Inhibitors

Chin Y. Liew,[†] Xiao H. Ma,[‡] Xianghui Liu,[‡] and Chun W. Yap*,[†]

Pharmaceutical Data Exploration Laboratory, Department of Pharmacy, National University of Singapore, and Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore

Received October 20, 2008

Lymphocyte-specific protein tyrosine kinase (Lck) inhibitors have treatment potential for autoimmune diseases and transplant rejection. A support vector machine (SVM) model trained with 820 positive compounds (Lck inhibitors) and 70 negative compounds (Lck noninhibitors) combined with 65 142 generated putative negatives was developed for predicting compounds with a Lck inhibitory activity of $IC_{50} \leq 10$ μM. The SVM model, with an estimated sensitivity of greater than 83% and specificity of greater than 99%, was used to screen 168 014 compounds in the MDDR and was found to have a yield of 45.8% and a false positive rate of 0.52%. The model was also able to identify novel Lck inhibitors and distinguish inhibitors from structurally similar noninhibitors at a false positive rate of 0.27%. To the best of our knowledge, the SVM model developed in this work is the first model with a broad applicability domain and low false positive rate, which makes it very suitable for the virtual screening of chemical libraries for Lck inhibitors.

INTRODUCTION

T-cell-mediated immune response has been suggested to be involved in the pathogenesis of many immunological diseases such as type I diabetes, asthma, rheumatoid arthritis, multiple sclerosis, inflammatory bowel disease, psoriasis, systemic lupus erythematosus, and transplant rejection. Lymphocyte-specific protein tyrosine kinase (Lck), a member of the Src family of nonreceptor tyrosine kinases, is mainly expressed in T cells[1] and natural killer cells.[2] It is implicated in T-cell antigen receptor (TCR) linked signal transduction pathways that control the activation and differentiation of T cells.[3,4] T-cell activation proceeds with the engagement of major histocompatibility complex antigen to TCR. Lck is then recruited to the TCR complex via its association with CD4 and CD8 co-receptors and later phosphorylates tyrosine residues within immunoreceptor tyrosine-based activation motifs located in the ζ chains of the TCR complex. This allows the binding of ζ-chain-associated protein kinase 70 (ZAP-70) to TCR. The downstream event in signal transduction is further triggered when ZAP-70 is phosphorylated by Lck.[5−7] Consequently, the inhibition of T-cell activation has been explored with synthetic Lck inhibitors that have potential as a treatment for autoimmune diseases and transplant rejection.[8]

This work will focus on the development of a computational Lck inhibitor model for the identification of potential Lck inhibitors. The use of computational models to perform virtual screening for drug candidates is routinely conducted during the drug discovery process and has been used for drug discoveries in signal transduction.[9,10] It is a favorable alternative to high-throughput screening (HTS) and combinatorial chemistry because virtual screening can identify drug candidates in a fast and cheap manner.[11] Currently, Lck inhibitor identification has been investigated using ligand-based screening,[12−18] pharmacophore-based screening,[19] and protein structure-based modeling.[20] These studies have been useful for the prediction of the Lck inhibitory potential of compounds in congeneric series and the identification of common molecular features in Lck inhibitors. However, the number of compounds used in these studies is frequently less than 200, and studies have shown that models developed using a limited number of compounds tend to have a limited applicability domain,[21,22] which may result in a large number of false positives when deployed for the virtual screening of large chemical libraries.[23−25]

In this study, 66 032 compounds from 8423 chemical families were used to develop a support vector machine (SVM) model for the identification of Lck inhibitors, which is a significantly larger number than the typical hundreds of compounds used in earlier studies. This will increase the applicability domain of the current model compared to earlier models. SVM is a machine learning method based on statistical learning theory.[26] Compared to traditional artificial neural networks, SVM performed more efficiently in studies with large data sets and a large number of descriptors, and it could also generalize to new data better.[27] SVM has also shown promising classification results in the area of drug design; examples of the use of SVM include the prediction of drug metabolism, p-glycoprotein substrates, blood-brain barrier penetration, pregnane X receptor activators, and torsade de pointes, and various toxicological end points.[27] SVM has consistently shown good prediction ability for compounds of varied structures in these studies. Unlike most of the non-machine learning methods, SVM classifies compounds on the basis of the discriminative properties between active and inactive compounds rather than structural similarity to active compounds.[24] Therefore, it is useful for the classification of systems where there is limited knowledge on the mechanism or specific association between the activities and molecular properties.[28] SVM has also recently

* Corresponding author phone: 065-65165971; fax: 065-67791554; e-mail: phayapc@nus.edu.sg.

[†] Pharmaceutical Data Exploration Laboratory.

[‡] Bioinformatics and Drug Design Group.

10.1021/ci800387z CCC: $40.75
Published on Web 03/06/2009

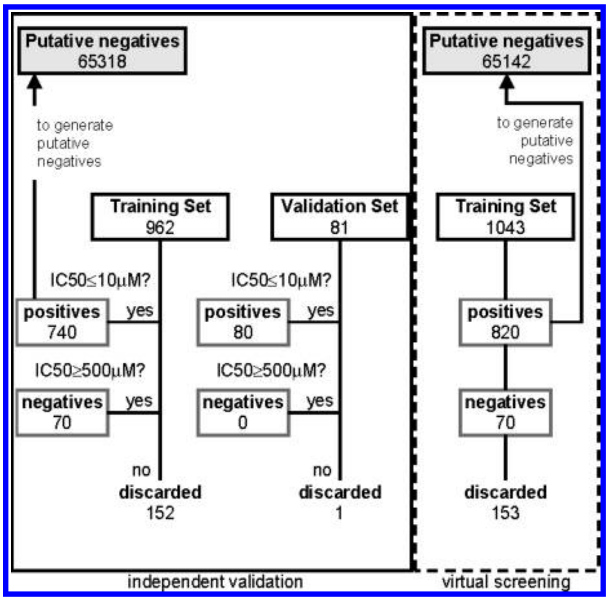

Figure 1. Flowchart for selection of compounds for training and independent validation sets. The positive compounds were used as references to generate putative negatives.

been used to develop ligand-based screening tools to improve the coverage, performance, and speed of virtual screening.[23] Hence, it is expected that a computational model developed by using SVM and a large number of compounds will be useful for the virtual screening of potential Lck inhibitors from large chemical libraries.

MATERIALS AND METHODS

Training Set. A total of 962 compounds with Lck inhibitory activity were gathered from published studies within the 1991−2008 period (Supporting Information available). The compounds were then categorized into positive (Lck inhibitors) and negative (Lck noninhibitors) compounds using cutoff values of $IC_{50} \leq 10\ \mu M$ and $IC_{50} \geq 500\ \mu M$, respectively. Compounds with IC50s between these two criteria were discarded from the training set, as shown in Figure 1. This resulted in the selection of 740 positive and 70 negative compounds for the training set.

A common problem in ligand-based screening studies is the lack of negative compounds, resulting in an imbalance between positive and negative compounds in the data set.[29−31] This frequently leads to a problem of a high false positive rate for the computational model. Thus, a new approach used for generating putative inactive compounds[23,24] in SVM classification has been adopted in this study to augment the negative set. This method can generate putative negatives without requiring the knowledge of actual inactive compounds. Studies had shown that SVM classification models derived from these putative negatives can perform reasonably well in virtual screening.[23,24]

The putative negatives generation process was initiated by creating compound families where known compounds are clustered in the chemical space defined by their molecular descriptors.[32,33] By employing K-means clustering and molecular descriptors calculated from MODEL,[34] 8423 compound families were produced from approximately 13.7 million compounds with computable molecular descriptors from the PUBCHEM database and MDDR. The scale of compound families obtained was consistent with the 12 800-compound-occupying neurons for 26.4 million compounds of up to 11 atoms of C, N, O, and F[35] and the 2851 clusters for 171 045 natural products[36] reported in two studies.

On the basis of the 8423 compound families, the families for the training set were analyzed and matched. Matching has produced a data set of 65 318 putative negatives that were generated by randomly selecting eight compounds from each of the families that do not contain any of the 740 positive compounds in the training set. For families with less than eight compounds, all of their members were selected. The set of putative negatives was then added to the training set.

Molecular Descriptors. Molecular descriptors are quantitative representations of structural and physicochemical features of molecules. The 2D structures and 3D coordinates of the collected compounds were drawn and generated by using ChemDraw[37] and Corina,[38] respectively. A total of 100 molecular descriptors, which are listed in Table 1, were computed by MODEL[34] in this study. These include 13 simple molecular properties, 13 charge descriptors, 34 molecular connectivity and shape descriptors, and 40 electrotopological state indices. The descriptors were selected from more than 1000 descriptors described in the literature by discarding those that are redundant and nonapplicable to pharmaceutical agents.[34] Details of the descriptors can be found in the reference manual for MODEL.[39]

Determination of Structural Diversity. Structural diversity of a collection of compounds can be evaluated by using the diversity index (DI), which is the average value of the similarity between pairs of compounds in a data set:[40]

$$\mathrm{DI} = \frac{\sum_{i,j \in D \wedge i \neq j} \mathrm{sim}(i,j)}{|D|(|D| - 1)} \tag{1}$$

where sim(i,j) is a measure of similarity between compounds i and j, D is the data set, and $|D|$ is set cardinality. The data set is more diverse when DI approaches 0. Tanimoto coefficients[41] were used to compute sim(i,j) in this study:

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{k} x_{d_i} x_{d_j}}{\sum_{d=1}^{k} (x_{d_i})^2 + \sum_{d=1}^{k} (x_{dj})^2 - \sum_{d=1}^{k} x_{d_i} x_{d_j}} \tag{2}$$

where k is the number of descriptors calculated for the compounds in the data set.

Support Vector Machine (SVM). SVM, a classifier based on the structural risk minimization principle, is less affected by duplicated data and has a lower risk of model overfitting.[42] The theory for SVM had been thoroughly described in the literature.[26] Hence, only a brief description is given here.

In linearly separable data, SVM builds a maximal margin hyperplane to separate positive compounds from negative compounds, as shown in Figure 2. The hyperplane is built by searching for a vector w and a parameter b that minimizes $||w||^2$ such that the following criteria are satisfied:

$$w{\cdot}x + b \geq 1 \text{ if } y = 1 \text{ (positive compounds)} \tag{3}$$

$$w{\cdot}x + b \leq 1 \text{ if } y = -1 \text{ (negative compounds)} \tag{4}$$

where compounds are denoted by x, class index by y, and the normal vector to the hyperplane by w. The perpendicular

Table 1. Descriptors[39] Used in This Study

descriptor class	no. of descriptors	descriptors
simple molecular properties	13	molecular weight; Sanderson electronegativity sum; no. of atoms, bonds, rings; H-bond donor/acceptor; rotatable bonds; N or O heterocyclic rings; no. of C, N, O atoms
charge descriptors	10	relative positive/negative charge, 0th−2nd electronic-topological descriptors, electron charge density connectivity index, total absolute atomic charge, charge polarization, topological electronic index, local dipole index
molecular connectivity and shape descriptors	37	1st−3rd order Kier shape index, Schultz/Gutman molecular topological index, total path count, 1−6 molecular path count, Kier molecular flexibility, Balaban/Pogliani/Wiener/Harary index, 0th edge connectivity, edge connectivity, extended edge connectivity, 0th−2nd valence connectivity, 0th−2nd order $\delta-\chi$ index, 0th−2nd solvation connectivity, 1st−3rd order κ α shape, topological radius, centralization, graph-theoretical shape coefficient, eccentricity, gravitational topological index
electrotopological state indices	40	sum of E-state of atom types sCH_3, dCH_2, $ssCH_2$, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH_3, sNH_2, $ssNH_2$, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH, H-bond acceptors, all heavy/C/hetero atoms; sum of H E-state of atom types HsOH, HdNH, HsSH, $HsNH_2$, HssNH, HaaNH, HtCH, $HdCH_2$, HdsCH, HaaCH, HCsats, H-bond donors

distance from the hyperplane to the origin is $|-b|/||w||$, and $||w||^2$ is the Euclidean norm of w. On the basis of the derived w and b, a new compound x can be classified as a positive or negative when sign$[(w \cdot x) + b]$ is positive or negative, respectively.

Nonlinear SVM is useful for classifying compounds of diverse structures which are usually not linearly separable. SVM maps the input vectors into a higher dimensional feature space by using a kernel function, as illustrated in Figure 3. The radial basis function kernel, which has been widely used and had consistently shown better performance,[27] was used in this study:

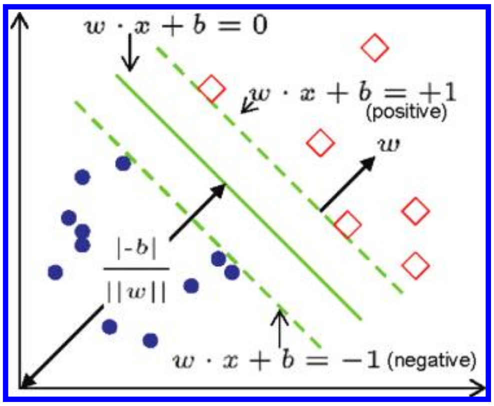

Figure 2. Decision boundary and margin separating positive class and negative class in SVM.

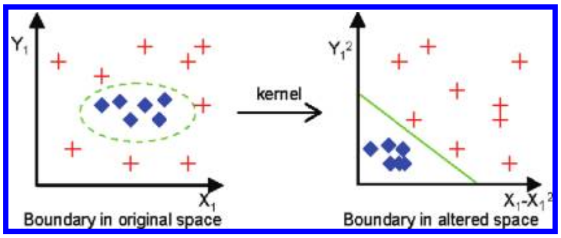

Figure 3. Example of nonlinearly separable vectors transformed into an altered space.

$$K(x_i, x_j) = e^{-||x_i - x_j||^2/2\sigma^2} \tag{5}$$

For the SVM model in this study, hard margin SVM was used, and σ was found to be 1.1 for the best performing model.

Applicability Domain. The applicability domain of the SVM model was calculated on the basis of the range of the individual molecular descriptors.[43] The minimum and maximum values of each molecular descriptor were obtained by considering all of the compounds in the training data set. Figure 4 is a visualization of the use of ranges to define the applicability domain for a model consisting of three descriptors. The applicability domain is the box (or a hyper-rectangle for models with more than three descriptors) that is defined by the extremes of the data. Classification of compounds that fall outside the hyper-rectangle is considered unreliable. Hence, in this work, compounds in both the external validation set and the MDDR data were checked for their suitability for classification by the SVM model with the hyper-rectangle. A compound was considered unsuitable if it violated one or more of the 100 molecular descriptor ranges and was excluded from the prediction process.

Model Validation. In order to fully assess the suitability of the SVM model for the virtual screening of chemical libraries for Lck inhibitors, the model was validated using a number of methods.

First, the SVM model, $SVM_{Tr+PutNeg}$ (subscript indicates the set of compounds that was used to train the model; Tr, collected training set; PutNeg, putative negative compounds; Val, independent validation set), which was developed using the training set of 810 compounds and 65 318 putative negative compounds, was internally validated using 5-fold cross-validation. In 5-fold cross-validations, the training set is divided into five groups of approximately equal size through stratified sampling. SVM was trained with four subsets of data, after which the performance of the model was tested with the fifth subset. This process was repeated five times, resulting in five combinations, so that every subset was used as the testing set once.

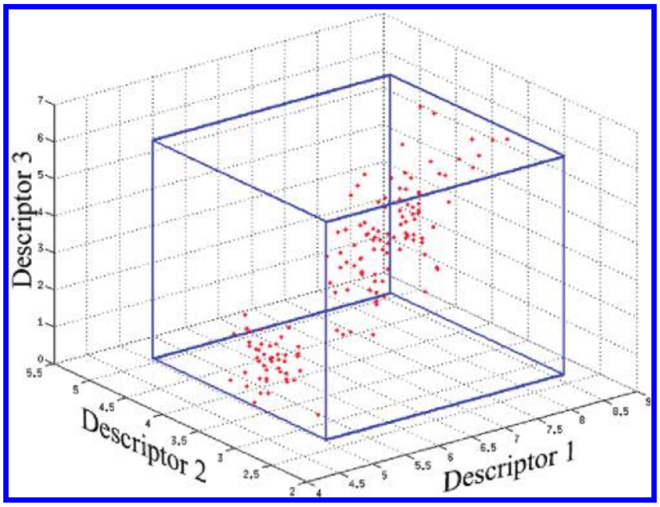

Figure 4. The box that encloses the data is the applicability domain of a model built from a data set with three descriptors.

$SVM_{Tr+PutNeg}$ was also externally validated using an independent validation set. A total of 81 compounds were obtained from three most recent studies (Supporting Information available), and these were subjected to the same preparations and filters as those compounds in the training set. In the end, all except one compound were selected for the validation set, as shown in Figure 1.

The performance results of $SVM_{Tr+PutNeg}$ from the 5-fold cross-validation and from the independent validation set were also compared. Concordance between the two sets of results would suggest that the risk of overfitting was low.

In order to further evaluate the suitability of a SVM model for identifying Lck inhibitors from large chemical libraries, compounds in the MDDR were screened with the SVM model. As the independent validation set was subsequently found to contain a substantial number of compound families that were not represented in the original training set, the entire validation set was added to the training set, and a set of 65 142 putative negative compounds was regenerated to match the new profile of the training set. A new SVM model ($SVM_{Tr+PutNeg+Val}$) was then developed from the new training set and used for screening MDDR compounds.

Before screening, the MDDR compounds were characterized in terms of their Lck inhibitory activity and structural similarity for ease of measuring performances. It was found that the MDDR contained 24 compounds with a Lck inhibitory activity of $IC_{50} \leq 10\ \mu M$, and these were labeled as "known inhibitors". It also had another 30 compounds that were labeled as "suspected inhibitors", and these include compounds with a Lck inhibitory activity which did not fulfill the $IC_{50} \leq 10\ \mu M$ cutoff or without an IC_{50} value. A third set of compounds, "structurally similar noninhibitors", was obtained by including those compounds in the MDDR (excluding compounds in the first two sets) that had a Tanimoto coefficient of ≥ 0.9 with at least one of the 24 known inhibitors. It is to be noted that compounds in these three sets were not present in the training set.

The effect of adding putative negative compounds to the training set was determined by developing a SVM model (SVM_{Tr+Val}) using the training set plus the independent validation set. The performance of this model was assessed using the MDDR compounds, and the results were compared to those from $SVM_{Tr+PutNeg+Val}$.

Table 2. Diversity Index (DI) of Several Compounds Classes in Descending Order of Structural Diversity

chemical class	no. of compounds	DI
satellite structures[49]	9	0.250
National Cancer Institute diversity set[49]	1990	0.452
FDA approved drugs[49]	1183	0.452
estrogen receptor ligands[49]	1009	0.511
benzodiazepine receptor ligands[49]	405	0.686
dihydrofolate reductase inhibitors[49]	756	0.727
Lck inhibitors in training set (this study)	740	0.734
penicillins[49]	59	0.790
fluoroquinolones[49]	39	0.791
cephalosporins[49]	73	0.812
cyclooxygenase 2 inhibitors[49]	467	0.840

Finally, logistic regression (LR), which is a classical statistical method and is less complex than SVM, was used to develop two models, LR_{Tr+Val} and $LR_{Tr+PutNeg+Val}$. The performances of LR_{Tr+Val} and $LR_{Tr+PutNeg+Val}$ were determined using MDDR compounds. The purpose of using a classical statistical method is to determine whether the use of SVM will result in a model that is more complex than is necessary for the virtual screening of chemical libraries for Lck inhibitors.

Evaluation of Prediction Performance. In the case of classification methods, the performance of machine learning methods can be assessed by the quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).[44] The prediction accuracy for positive compounds (Lck inhibitors) and negative compounds (Lck noninhibitors) are sensitivity, SE = TP/(TP + FN), and specificity, SP = TN/(TN + FP), respectively. The overall prediction performance can be calculated by the overall prediction accuracy (Q):

$$Q = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$

and Matthew's correlation coefficient[45] (MCC):

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FN} \times \mathrm{FP}}{\sqrt{(\mathrm{TP} + \mathrm{FN})(\mathrm{TP} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})}} \tag{7}$$

The area under the ROC curve (AUC), which has been widely used for classification performance in many fields,[46] was also computed.

For the performance of the SVM model in virtual screening, the yield (percentage of predicted compounds in known inhibitors), hit rate (HR = percentage of known inhibitors in predicted compounds), false positive rate (FPR = percentage of predicted compounds in noninhibitors), and enrichment factor (EF = ratio of hit rate to the percentage of known inhibitors in MDDR), which shows the magnitude of hit-rate improvement over random selection, were evaluated.

RESULTS

Data set Diversity and Distribution. Table 2 shows that the 740 Lck inhibitors have an intermediate DI of 0.734, which is comparable to that of known dihydrofolate reductase inhibitors. A three-dimensional visualization of the collected compounds using the first three principle components after principle component analysis (PCA) is shown in Figure 5.

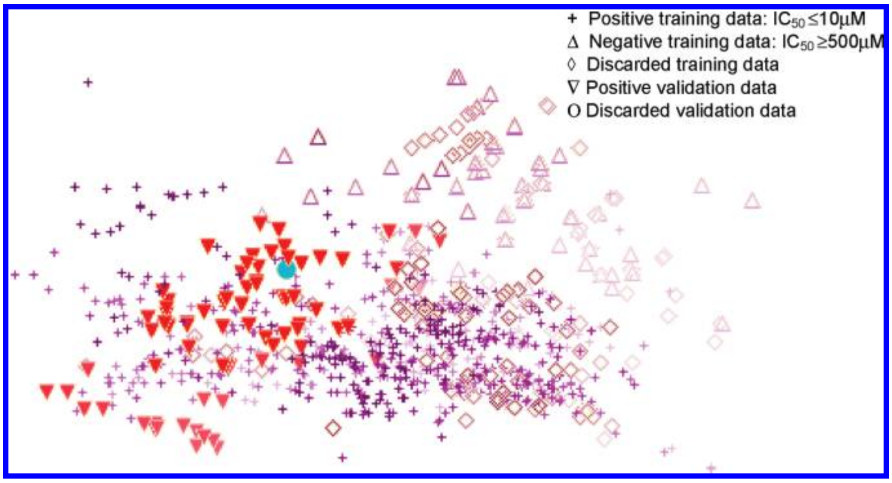

Figure 5. Visualization of the chemical space for the training and independent validation sets using the first three principle components from PCA.

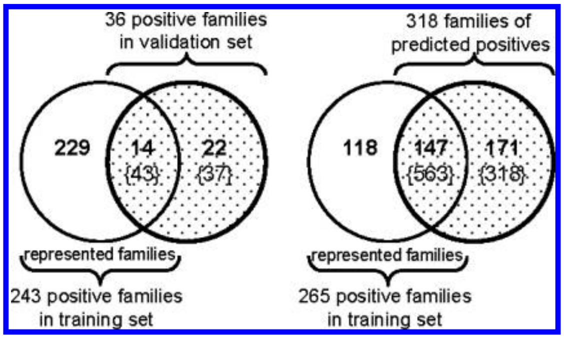
36 positive families in validation set
318 families of predicted positives
229
14 {43}
22 {37}
118
147 {563}
171 {318}
represented families
represented families
243 positive families in training set
265 positive families in training set

Figure 6. Distribution of families for the 80 positive compounds in validation set and 881 virtual screening predicted positives (the number of compounds is given in curly brackets). Families in the shaded region but not in the intersection are those families which are not represented in the training set.

The results showed that, in general, the compounds were well distributed in the chemical space, and there was no clear boundary between the positive and negative compounds. Although there were a few compounds which were isolated from the majority of the compounds, there was no evidence to indicate that these compounds were outliers, and thus they were left in the training set.

Figure 6 shows the distribution of Lck inhibitors in terms of compound families. The analysis found that the 740 inhibitors in the training set and 80 inhibitors in validation set belonged to 243 and 36 families, respectively (total of 265 (3.1%) unique families from the total 8423 families). The analysis also showed that the characteristics of the independent validation set were different from those of the positive training data set, as only 38.9% of the families in the validation set were represented in the training set. This suggests that the independent validation set was not only useful for evaluating the performance of the models on similar compounds but also on novel compounds.

Applicability Domain. A total of 168 016 MDDR compounds were checked, and all except two MDDR compounds with long chains were found to be within the applicability domain for $\mathrm{SVM}_{\mathrm{Tr+PutNeg+Val}}$ and $\mathrm{LR}_{\mathrm{Tr+PutNeg+Val}}$. For $\mathrm{SVM}_{\mathrm{Tr+Val}}$ and $\mathrm{LR}_{\mathrm{Tr+Val}}$, 79 793 compounds from the MDDR were within the applicability domain. Among the 79 793 compounds, 19 are known Lck inhibitors.

Model Performances. Table 3 gives the performance of $\mathrm{SVM}_{\mathrm{Tr+PutNeg}}$ for predicting Lck inhibitors and noninhibitors by means of 5-fold cross-validation and an independent validation set. The models for the 5-fold cross-validation had performed consistently well in predicting positive compounds (average SE = 87.8%) and also in predicting negative compounds (average SP = 99.9%), with an overall accuracy of 99.7%, a MCC of 0.788, and an AUC of 0.997. When tested on the independent validation set, $\mathrm{SVM}_{\mathrm{Tr+PutNeg}}$ performed with an overall sensitivity of 83.8%, which is comparable to the results in 5-fold cross-validation.

A total of 168 014 compounds in the MDDR were screened with $\mathrm{SVM}_{\mathrm{Tr+PutNeg+Val}}$. The results are given in Table 4. $\mathrm{SVM}_{\mathrm{Tr+PutNeg+Val}}$ had predicted 881 compounds to have Lck inhibitory activity. Analysis of the compound families of these 881 compounds has shown that they belong to 318 families, and only 46.2% of these are represented in the training set. A total of 121 936 compounds in the MDDR were also found to be similar in structure to at least one of the known inhibitors, that is, the Tanimoto coefficient ≥ 0.90. A total of 334 of these "structurally similar noninhibitors" were predicted as positives, resulting in a false positive rate of 0.27% for this group of compounds.

For the 79 793 MDDR compounds that were screened by $\mathrm{SVM}_{\mathrm{Tr+Val}}$ and $\mathrm{LR}_{\mathrm{Tr+Val}}$, 48 823 and 64 727 compounds were predicted to have Lck inhibitory activity, respectively, with yields of 89.5% (17 known inhibitors out of 19) for both models. However, the false positive rates were high at 61.2% and 81.1% for $\mathrm{SVM}_{\mathrm{Tr+Val}}$ and $\mathrm{LR}_{\mathrm{Tr+Val}}$, respectively.

DISCUSSIONS

Cutoff Value for Lck Inhibitory Activity. It is common in the development of classification models to use a single cutoff value to separate compounds into positive and negative compounds. However, in this study, it is inaccurate to use a single cutoff value. This is because a single cutoff value of 10 μM would cause some compounds which exhibit weak Lck inhibitory activity to be classified into the negative group. This is undesirable because some novel drug leads may initially exhibit weak activity but can be further modified into potent drugs. Thus, using a single cutoff value of 10 μM may result in potential Lck inhibitors being included in the negative group, which may affect the performance of the model in identifying potentially useful novel Lck inhibitors. It is also not desirable to use a single cutoff value

Table 3. Classification Performance of SVM in Predicting Lck Inhibitory Activity

		no. of compounds											
test		total	pos	neg	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)	MCC	AUC
5 fold cross validation	fold 1	13226	148	13078	139	9	93.9	13065	13	99.9	99.8	0.857	0.993
	fold 2	13226	148	13078	127	21	85.8	13065	13	99.9	99.7	0.776	0.997
	fold 3	13226	148	13078	128	20	86.5	13062	16	99.9	99.7	0.766	0.999
	fold 4	13225	148	13077	129	19	87.2	13060	17	99.9	99.7	0.768	0.998
	fold 5	13225	148	13077	127	21	85.8	13063	14	99.9	99.7	0.771	0.999
	average	13226	148	13078	130	18	87.8	13063	15	99.9	99.7	0.788	0.997
independent validation set		80	80		67	13	83.8						

Table 4. Performance of SVM Model in Virtual Screening of 168 014 MDDR Compounds for Lck Inhibitors

compound types	no. (%) in MDDR	total no. of unique families	no of families represented in training set	predicted positives	hits[a]	yield (%)	hit rate (%)	false positive rate (%)	enrichment factor
known inhibitors[b]	24 (0.014)	24	14 (58.3%)	881	11	45.8	1.25	0.52	87
suspected inhibitors[c]	30 (0.018)	29	10 (34.5%)	881	6	20.0	0.68	0.52	38
overall	54 (0.032)	52	23 (44.2%)	881	17	31.5	1.93	0.51	60

[a] Hits: Predicted positive compounds that are known/suspected inhibitors in MDDR. [b] Known inhibitors: compounds in MDDR identified to have Lck inhibitory activity $IC_{50} \leq 10\ \mu M$. [c] Suspected inhibitors: compounds in MDDR that were reported to have IC_{50} between 10 and 500 μM or without IC_{50} value.

of 500 μM, as it may result in an unacceptably large number of false positives when screening a chemical library.

Hence, in this study, two cutoff values were used to separate compounds into positive ($IC_{50} \leq 10\ \mu M$) and negative ($IC_{50} \geq 500\ \mu M$) compounds. This will minimize the risk of including potential Lck inhibitors in the negative group and reduce the number of false positives during virtual screening. The wide margin between the two cutoff values was to account for variances in biological assays which may arise because of differences in laboratories and equipment. A possible drawback of this method is that too many compounds may have IC_{50} values between the two cutoff values and thus could be excluded from the training and validation sets. However, this does not pose much of a problem for the current study, as only approximately 16% and 1% of the compounds were excluded from the training and validation sets, respectively.

Putative Negative Compounds. In this study, the novel method developed by Han et al.[23] for enriching true negative compounds with putative negative compounds was used to increase the quantity and diversity of negative compounds for training a SVM model. The performance of this method was evaluated by validating $SVM_{Tr+PutNeg}$ internally and externally using 5-fold cross-validation and an independent validation set, respectively. The usefulness of adding putative negative compounds was also assessed by comparing the prediction results for $SVM_{Tr+PutNeg+Val}$ and SVM_{Tr+Val} from MDDR compounds.

The high sensitivity value of $SVM_{Tr+PutNeg}$ determined by using 5-fold cross-validation was consistent with the corresponding value determined by using the independent validation set. Unfortunately, the independent validation set, which was compiled from the three most recent publications, did not contain any negative compounds. Thus, it was not possible to determine the actual specificity of the model. However, the actual specificity of $SVM_{Tr+PutNeg}$ would be expected to be close to the specificity value determined by 5-fold cross-validation since the sensitivity values determined by 5-fold cross-validation and the independent validation set were similar. The concordance between the results from 5-fold cross-validation and the independent validation set suggests that the risk of overfitting was low.

The high false positive rate of 61.2% for SVM_{Tr+Val} compared to the low false positive rate of 0.52% for $SVM_{Tr+PutNeg+Val}$ suggests that the addition of putative negative compounds was useful for reducing the false positive rate of SVM models.

It had also been found that the applicability domains of models developed from training sets that included putative negative compounds were larger than those developed from training sets without the putative negative compounds. While this result is not surprising, since having more compounds in the training set would usually translate into greater ranges in descriptor values and hence a larger applicability domain, the enlarged applicability domain would enable the Lck inhibitory potential of more compounds in chemical libraries to be reliably predicted by the SVM models. Hence, these results together with the high sensitivity and specificity and low false positive rate suggests that SVM models are potentially useful for classifying compounds in large chemical libraries into Lck inhibitors and noninhibitors.

A potential disadvantage of the putative negative compounds generation method is the probable inclusion of undiscovered inhibitors in the negative set, resulting in a SVM model that cannot identify an active compound that has a structure similar to that of putative negative compounds. Nevertheless, the results of this work and two other studies[23,24] have shown that, for many biological target classes, such an undesirable effect is expected to be relatively small, and it is still possible for a substantial percentage of positive compounds to be classified correctly despite their membership in negative families. Furthermore, due to the large effort in searching for the known positives of Lck inhibitors, the number of undiscovered positive compound families is expected to be relatively small, likely no more than several hundred. Consequently, the ratio of positive families (that contain Lck inhibitors whether known or undiscovered) to total families is expected to be < 10% for

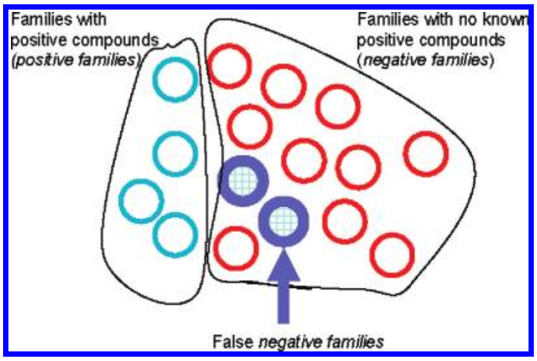
Families with positive compounds (positive families)
Families with no known positive compounds (negative families)
False negative families

Figure 7. Illustrating the inclusion of undiscovered positive families into families with no known positive compounds.

this study. As a result, putative negative compounds can be generated by extracting a few compounds from the negative families with somewhat little risk. This is because a maximum "false negative families" rate of < 10% is expected even for the worst case scenario, where all undiscovered positive compound families are misidentified as negative families, as illustrated in Figure 7.

Predicting Positive Compounds Unrepresented in Training Set. Figure 5 shows that a substantial number of the positive compounds in the validation set were clustered away from the positive compounds in the training set, and Figure 6 shows that 61.1% of the positive compound families in the validation set were not represented in the training set. Despite the apparent dissimilarity and lack of representation, the classification performance of $\text{SVM}_{\text{Tr+PutNeg}}$ for the independent validation set has a sensitivity of 83.8%. Further analysis showed that the sensitivity for compounds whose compound families are represented and not represented in the training set was 95.3% and 70.3%, respectively.

These results suggest that SVM, like most machine learning methods, requires knowledge of compound families for optimum performance. This might be attributed to the reduction of false negative families risk, as shown in Figure 7, by having knowledge of more positive families. However, given that $\text{SVM}_{\text{Tr+PutNeg}}$ was also able to predict novel compounds unseen in the training set with reasonably good accuracy, it was likely that the predictions were based on the compounds' characteristics and not by their mere membership in the represented family. This suggests that SVM models may be suitable for screening large chemical libraries where compounds are usually distributed in many compound families and thus are not well represented in the training set.

Evaluation of SVM Model Using MDDR. The performance of $\text{SVM}_{\text{Tr+PutNeg+Val}}$ on the 24 known and 30 suspected Lck inhibitors present in the MDDR (Table 4) was comparable to the yields of 22−55% and 44−69%, HRs of 1.5−4.1% and 14%−72%, and EFs of 22−55 and 44−69 that were obtained in a previous study on SVM models for the virtual screening of 98 400- and 100 000−103 000-compound libraries, respectively.[23] This suggests that the SVM model is potentially useful for screening large chemical libraries without requiring any prescreening filtering methods such as Lipinski's rule of five[47] and lead-likeness.[48]

Results from Table 4 also suggested a positive correlation between the family representation in the training set and the yield of the predictions. This is not surprising and is consistent with the earlier results obtained from the independent validation set. Thus, it is important to constantly refine the SVM model by introducing newly discovered positive and negative families from the drug discovery process into the training set so that a more refined hyperplane, which will improve the screening performance, can be obtained.

A previous study had tested SVM models trained by sparsely distributed actives on structurally similar nonactives in the range of 19 495−38 436 compounds. The SVM models had false positives rate of 2.6−7.8% (highly diverse data), 3.3−6.4% (moderately diverse data), and 5.8−8.3% (sparsely diverse data).[24] A similar experiment was done in this study, and $\text{SVM}_{\text{Tr+PutNeg+Val}}$ appears to perform fairly well in terms of the false hit rate (0.27%). This result is consistent with the high specificity value for $\text{SVM}_{\text{Tr+PutNeg}}$ obtained from 5-fold cross-validation and also with the results from the independent validation set, which suggested that the SVM models do not base their predictions merely on membership in the represented family but rather on compounds' characteristics.

Comparison of SVM Model with Logistic Regression Model. The prediction performances of $\text{SVM}_{\text{Tr+Val}}$ and $\text{LR}_{\text{Tr+Val}}$ on MDDR compounds were similar. However, when putative negative compounds were included in the training set, only the SVM model ($\text{SVM}_{\text{Tr+PutNeg+Val}}$) had a low false positive rate. The LR model ($\text{LR}_{\text{Tr+PutNeg+Val}}$) performed worse with the addition of putative negative compounds. This suggests that LR may not be suitable for data with a large class imbalance, and the use of complex methods like SVM is appropriate for developing models for predicting Lck inhibitors.

Application of SVM Model for Novel Lck Inhibitor Design. Analysis of the three most recent publications of Lck inhibitor synthesis and evaluation shows that the calculated Tanimoto coefficient (T) of one compound to another within the same publication can range from fairly dissimilar ($T = 0.105$, average $T = 0.369$) to closely resembling each other (highest $T = 0.996$, average $T = 0.994$). In this work, the Tanimoto coefficient for the 864 predicted positive MDDR compounds (excluding known and suspected inhibitors) calculated against the 820 positive training compounds ranged from 2.66×10^{-6} to 0.999. Thus, the SVM model developed in this work, $\text{SVM}_{\text{Tr+PutNeg+Val}}$, is able to identify novel compounds that are potential Lck inhibitors. This is useful because compounds with great dissimilarity from currently known compounds may be explored as new starting points for drug design, which may have been difficult to discover through the traditional synthesis process.

Figure 8 shows two novel potential Lck inhibitors that were identified by $\text{SVM}_{\text{Tr+PutNeg+Val}}$. These two are the compounds most dissimilar from the 820 positive training compounds (min. $T = 1.403 \times 10^{-5}$, max. $T = 0.316$). These compounds have some of the important pharmacophores discussed in other studies. They contain an aliphatic chain and a potential hydrogen-bond formation end[19] (bromophenyl or amide) and also the quinazoline N, which is another potential site for hydrogen bonding with Met319 in the Lck catalytic domain, as reported by Chen et al.[17] Hence, the SVM model developed in this

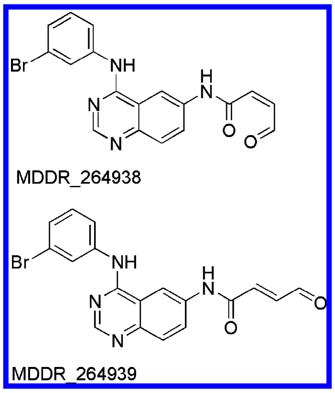

Figure 8. Two compounds in MDDR predicted to have Lck inhibitory activity by the SVM model which have the greatest dissimilarity from the collected compounds.

work, $\mathrm{SVM}_{\mathrm{Tr+PutNeg+Val}}$, is potentially useful as a tool to screen for novel Lck inhibitors early in the drug discovery stages.

CONCLUSION

In this work, a SVM model capable of identifying novel Lck inhibitors from large chemical libraries, with a low false positive rate of 0.52%, was developed from a large training set of Lck inhibitors and noninhibitors. The model was validated in a number of ways: internal validation over 5-fold cross-validation, external validation with compounds from the most recent published papers, screening of MDDR, comparison of models developed with and without putative negative compounds, and checking for overfitting by comparison with a LR model. The use of putative negative compounds was found to be useful for increasing the applicability domain and decreasing the false positive rate of the resultant computational model. In addition, the use of SVM, which does not depend solely on family memberships, enabled the model to distinguish noninhibitors that are structurally similar to the known inhibitor. Thus, the SVM model presented in this work is potentially useful as a complement to HTS for screening large libraries for novel Lck inhibitors with potent activity.

Supporting Information Available: Table 1 shows inhibitory activity IC_{50}(nM), SMILES, and references (in PubMed Unique Identifier or Digital Object Identifier) of all collected compounds used for building and validation of the SVM and logistic regression models. This material is available free of charge via the Internet at http://pubs.acs.org.

ACKNOWLEDGMENT

Our appreciation to Professor Chen Yu Zong (Bioinformatics and Drug Design Group, National University of Singapore) for his valuable discussions.

REFERENCES AND NOTES

(1) Veillette, A.; Abraham, N.; Caron, L.; Davidson, D. The lymphocyte-specific tyrosine protein kinase p56lck. *Semin. Immunol.* **1991**, *3* (3), 143–52.

(2) Biondi, A.; Paganin, C.; Rossi, V.; Benvestito, S.; Perlmutter, R. M.; Mantovani, A.; Allavena, P. Expression of lineage-restricted protein tyrosine kinase genes in human natural killer cells. *Eur. J. Immunol.* **1991**, *21* (3), 843–6.

(3) Weiss, A.; Littman, D. R. Signal transduction by lymphocyte antigen receptors. *Cell* **1994**, *76* (2), 263–74.

(4) Isakov, N.; Wange, R. L.; Samelson, L. E. The role of tyrosine kinases and phosphotyrosine-containing recognition motifs in regulation of the T cell-antigen receptor-mediated signal transduction pathway. *J. Leukocyte Biol.* **1994**, *55* (2), 265–71.

(5) Shaw, A. S.; Amrein, K. E.; Hammond, C.; Stern, D. F.; Sefton, B. M.; Rose, J. K. The lck tyrosine protein kinase interacts with the cytoplasmic tail of the CD4 glycoprotein through its unique amino-terminal domain. *Cell* **1989**, *59* (4), 627–36.

(6) Trevillyan, J. M.; Chiou, X. G.; Ballaron, S. J.; Tang, Q. M.; Buko, A.; Sheets, M. P.; Smith, M. L.; Putman, C. B.; Wiedeman, P.; Tu, N.; Madar, D.; Smith, H. T.; Gubbins, E. J.; Warrior, U. P.; Chen, Y. W.; Mollison, K. W.; Faltynek, C. R.; Djuric, S. W. Inhibition of p56(lck) tyrosine kinase by isothiazolones. *Arch. Biochem. Biophys.* **1999**, *364* (1), 19–29.

(7) Palacios, E. H.; Weiss, A. Function of the Src-family kinases, Lck and Fyn, in T-cell development and activation. *Oncogene* **2004**, *23* (48), 7990–8000.

(8) Kamens, J. S.; Ratnofsky, S. E.; Hirst, G. C. Lck inhibitors as a therapeutic approach to autoimmune disease and transplant rejection. *Curr. Opin. Invest. Drugs* **2001**, *2* (9), 1213–9.

(9) Fischer, P. M. Computational chemistry approaches to drug discovery in signal transduction. *Biotechnol. J.* **2008**, *3* (4), 452–70.

(10) Seifert, M. H.; Lang, M. Essential factors for successful virtual screening. *Mini-Rev. Med. Chem.* **2008**, *8* (1), 63–72.

(11) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11* (13−14), 580–94.

(12) Novic, M.; Nikolovska-Coleska, Z.; Solmajer, T. Quantitative Structure-Activity Relationship of Flavonoid p56lck Protein Tyrosine Kinase Inhibitors. A Neural Network Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (6), 990–998.

(13) Nikolovska-Coleska, Z.; Suturkova, L.; Dorevski, K.; Krbavcic, A.; Solmajer, T. Quantitative structure-activity relationship of flavonoid inhibitors of p56(lck) protein tyrosine kinase: A Classical/Quantum chemical approach. *Quant. Struct.-Act. Relat.* **1998**, *17* (1), 7–13.

(14) Zupan, J.; Novic, M. Optimisation of structure representation for QSAR studies. *Anal. Chim. Acta* **1999**, *388* (3), 243–250.

(15) Oblak, M.; Randic, M.; Solmajer, T. Quantitative structure-activity relationship of flavonoid analogues. 3. Inhibition of p56lck protein tyrosine kinase. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 994–1001.

(16) Thakur, A.; Vishwakarma, S.; Thakur, M. QSAR study of flavonoid derivatives as p56lck tyrosinkinase inhibitors. *Bioorg. Med. Chem.* **2004**, *12* (5), 1209–14.

(17) Chen, P.; Doweyko, A. M.; Norris, D.; Gu, H. H.; Spergel, S. H.; Das, J.; Moquin, R. V.; Lin, J.; Wityak, J.; Iwanowicz, E. J.; McIntyre, K. W.; Shuster, D. J.; Behnia, K.; Chong, S.; de Fex, H.; Pang, S.; Pitt, S.; Shen, D. R.; Thrall, S.; Stanley, P.; Kocy, O. R.; Witmer, M. R.; Kanner, S. B.; Schieven, G. L.; Barrish, J. C. Imidazoquinoxaline Src-family kinase p56Lck inhibitors: SAR, QSAR, and the discovery of (S)-N-(2-chloro-6-methylphenyl)-2-(3-methyl-1-piperazinyl)imidazo- [1,5-a]pyrido[3,2-e]pyrazin-6-amine (BMS-279700) as a potent and orally active inhibitor with excellent in vivo antiinflammatory activity. *J. Med. Chem.* **2004**, *47* (18), 4517–29.

(18) Badiger, A. M.; Noolvi, M. N.; Nayak, P. V. QSAR study of benzothiazole derivatives as p56lck inhibitors. *Lett. Drug Des. Discovery* **2006**, *3* (8), 550–560.

(19) Bharatham, N.; Bharatham, K.; Lee, K. W. P56 LCK inhibitor identification by pharmacophore modelling and molecular docking. *Bull. Korean Chem. Soc.* **2007**, *28* (2), 200–206.

(20) Tominaga, Y.; Jorgensen, W. L. General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *J. Med. Chem.* **2004**, *47* (10), 2534–49.

(21) Gramatica, P. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **2007**, *26* (5), 694–701.

(22) Parker, C. N.; Bajorath, J. Towards Unified Compound Screening Strategies: A Critical Evaluation of Error Sources in Experimental and Virtual High-Throughput Screening. *QSAR Comb. Sci.* **2006**, *25* (12), 1153–1161.

(23) Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z. 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.* **2008**, *26* (8), 1276–1286.

(24) Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z. Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds. *J. Chem. Inf. Model.* **2008**, *48* (6), 1227–1237.

(25) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Li, Z. R.; Han, L. Y.; Lin, H. H.; Chen, Y. Z. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J. Pharm. Sci.* **2007**, *96* (11), 2838–2860.

(26) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995; p xv, 188.

(27) Doucet, J.-P.; Barbault, F.; Xia, H.; Panaye, A.; Fan, B. Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design. *Curr. Comput.-Aided Drug Des.* **2007**, *3*, 263–289.

(28) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497–1505.

(29) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. 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.

(30) Lepp, Z.; Kinoshita, T.; Chuman, H. Screening for new antidepressant leads of multiple activities by support vector machines. *J. Chem. Inf. Model.* **2006**, *46* (1), 158–167.

(31) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21* (1−3), 53–62.

(32) Oprea, T. I.; Gottfries, J. Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157–166.

(33) Bocker, A.; Schneider, G.; Teckentrup, A. 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.

(34) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. 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.

(35) Fink, T.; Raymond, 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**, *47* (2), 342–353.

(36) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Weldmann, H. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (48), 17272–17277.

(37) CambridgeSoft Desktop Software - ChemDraw (Windows/Mac). http://www.cambridgesoft.com/ (accessed Dec 29, 2008).

(38) CORINA: Generation of 3D coordinates. http://www.molecular-networks.com/software/corina/index.html (accessed Dec 29, 2008).

(39) Li, Z. R.; Han, L. Y.; Chen, Y. Z. MODEL Reference Manual. http://jing.cz3.nus.edu.sg/model/ (accessed Dec 29, 2008).

(40) Perez, J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34* (2), 143–52.

(41) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983–996.

(42) Tan, P.-N.; Steinbach, M.; Kumar, V., Classification: Alternative Techniques. In *Introduction to Data Mining*; Addison-Wesley: 2005; pp 207−315.

(43) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *ATLA Altern. Lab. Anim.* **2005**, *33* (5), 445–59.

(44) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16* (5), 412–424.

(45) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442–51.

(46) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 239–55.

(47) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(48) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem.* **1999**, *38* (24), 3743–3748.

(49) Yap, C. W.; Xue, Y.; Li, H.; Li, Z. R.; Ung, C. Y.; Han, L. Y.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. Prediction of compounds with specific pharmacodynamic, pharmacokinetic or toxicological property by statistical learning methods. *Mini-Rev. Med. Chem.* **2006**, *6* (4), 449–59.

CI800387Z