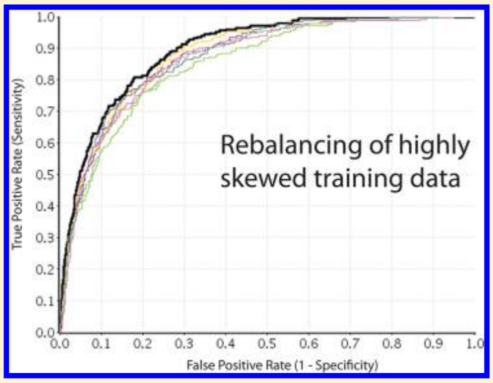
Journal of
CHEMICAL INFORMATION
AND MODELING

Structure Based Model for the Prediction of Phospholipidosis Induction Potential of Small Molecules

Hongmao Sun,*[†] Sampada Shahane,[†] Menghang Xia,[†] Christopher P. Austin,[†] and Ruili Huang*[†]

[†]National Institutes of Health (NIH) Chemical Genomics Center, NIH, Bethesda, Maryland 20892, United States

S Supporting Information



ABSTRACT: Drug-induced phospholipidosis (PLD), characterized by an intracellular accumulation of phospholipids and formation of concentric lamellar bodies, has raised concerns in the drug discovery community, due to its potential adverse effects. To evaluate the PLD induction potential, 4,161 nonredundant drug-like molecules from the National Institutes of Health Chemical Genomics Center (NCGC) Pharmaceutical Collection (NPC), the Library of Pharmacologically Active Compounds (LOPAC), and the Tocris Biosciences collection were screened in a quantitative high-throughput screening (qHTS) format. The potential of drug-lipid complex formation can be linked directly to the structures of drug molecules, and many PLD inducing drugs were found to share common structural features. Support vector machine (SVM) models were constructed by using customized atom types or Molecular Operating Environment (MOE) 2D descriptors as structural descriptors. Either the compounds from LOPAC or randomly selected from the entire data set were used as the training set. The impact of training data with biased structural features and the impact of molecule descriptors emphasizing whole-molecule properties or detailed functional groups at the atom level on model performance were analyzed and discussed. Rebalancing strategies were applied to improve the predictive power of the SVM models. Using the undersampling method, the consensus model using one-third of the compounds randomly selected from the data set as the training set achieved high accuracy of 0.90 in predicting the remaining two-thirds of the compounds constituting the test set, as measured by the area under the receiver operator characteristic curve (AUC-ROC).

■ INTRODUCTION

Phospholipids are essential and dynamic components of the plasma and intracellular membranes in normal cells.[1] Phospholipidosis (PLD) is a type of phospholipid storage disorder characterized by excessive accumulation of intracellular phospholipids in a variety of tissue types, including lung, liver, brain, kidney, heart, circulating lymphocytes, etc.[2] Many drugs can accumulate in these tissues to a remarkable degree by forming complexes with the polar phospholipids in the lysosome, giving rise to the so-called drug-induced PLD.[3] Although there is no direct link of PLD with any toxic implications, this phenomenon may delay or halt the development of pharmaceuticals.[4] Identification and exclusion of PLD inducing compounds from the drug discovery pipeline at an early stage could help in minimizing the attrition rate at the expensive development stage. Many PLD inducing drugs share some common structural features − typically a hydrophobic region with one or more aromatic rings and a hydrophilic side chain with one or more positively charged centers at physiological pH.[5] These compounds are also known as cationic amphiphilic drugs (CADs).[3] Therefore, the combination of a couple of key physicochemical parameters, such as pK_a and calculated logP (clogP),[5] or net charges (NC) and clogP,[6] can result in reasonably accurate predictions of PLD inducing potentials on small sets of compounds. Utilizing machine learning methods, such as Bayesian statistics, artificial immune systems, and support vector machines (SVM), has been shown to significantly improve the predictive accuracy.[7,8] However, these PLD models were mostly based on *in vivo* data collected from literature on various species, including rat, mouse, dog, rabbit, hamster, monkey, and human, or across different tissue types, such as lung, macrophage, liver, kidney, nerve, eye, heart, blood, muscle, etc.[8] Not only is the quality of these training data in terms of data integrity questionable but also the data sizes are relatively small, ranging from tens to a few hundreds of compounds. In some cases, compounds were categorized as PLD negative not based on the experimental evidence but solely on the absence of positive reports.[4,8] In this study, over four thousand compounds from the National Institutes of Health Chemical Genomics Center (NCGC) Pharmaceutical Collection (NPC) of approved and investigational drugs,[9] Sigma's Library of Pharmacologically Active Compounds (LOPAC), and the Tocris Biosciences bioactive compound collection were screened for PLD induction in HepG2 cells using an automated imaging-based assay system in a quantitative high throughput screening (qHTS) format.[10] The structure-based models using this qHTS data set may

Received: April 16, 2012
Published: June 23, 2012

provide useful information for predicting compound-induced PLD.

MATERIALS AND METHODS

Data Set. The 1280 LOPAC compounds, 2816 NPC compounds, and 1395 Tocris compounds were screened for the induction of phospholipidosis in HepG2 cells. Amiodarone, a known phospholipidosis inducer,[11] was used as a positive control in the screening. Briefly, HepG2 cells were plated in 1536-well plates coated with Collagen-I and treated with compounds in the presence of LipidTox dye for 24 h at 37 °C. After cells were fixed with 3.2% formaldehyde and Hoechst (1:1000) solution at room temperature for 30 min, the plates were washed once with DPBS using the Kalypsis washer-dispenser, and then sealed and stored at 4 °C before imaging. Fluorescence intensity of the assay plates was measured using Image Xpress Micro (Molecular Devices, U.S.A.) with DAPI and TRITC filters and their proprietary program. After the primary qHTS, a concentration−response curve (CRC) was generated for every compound with concentrations ranging from 2.45 nM to 38 μM. Analysis of compound CRC was performed as previously described.[10] Concentration−response data for each compound were fitted to a four-parameter Hill equation, yielding concentrations of half-maximal activity (AC_{50}) and maximal response (efficacy) values. Compounds were designated as Class 1−4 according to the type of CRC observed.[10] Curve classes are heuristic measures of data confidence, classifying concentration−responses on the basis of efficacy, the number of data points observed above background activity, and the quality of fit. Compounds with class 1.1, 1.2, 2.1, or 2.2 with >50% efficacy were defined as active. Compounds with class 4 curves were defined as inactive, and compounds with other curve classes were considered inconclusive. The compounds were processed through a Pipeline Pilot[12] protocol to remove salts, redundant, and heavy metal containing compounds. Originally, active, inconclusive, and inactive compounds were assigned a score of 2, 1, and 0, respectively. An average score was computed for replicated compounds. The compound was labeled active if the average score was above 1.5 or inactive if the score was below 0.5. The compounds with a score between 0.5 and 1.5 were considered inconclusive and removed from the data set. The final data set contained 4,161 nonredundant compounds with 382 (9.2%) actives.

Training and Testing Sets. The qHTS of PLD inducing potential was first performed on the LOPAC compounds, thus the initial QSAR models were trained using the LOPAC collection containing 1,128 nonredundant and noninconclusive compounds, where 136 compounds (12.1%) were PLD active. The compounds in the NPC and Tocris collections (N&T data set), which were assayed later-on, served as the external test set. The assay results indicated a reduced PLD active rate of 8.1% for the N&T data set. In an attempt to enhance the chemical space coverage of the training set, one-third of the compounds were randomly selected from the combined set of 4,161 compounds including LOPAC, NPC, and Tocris, with the remaining two-thirds of compounds labeled as the test set. In this case, the training set comprised 1,395 compounds with 9.4% actives, and the test set had 9.1% actives out of 2,766 compounds. In order to investigate the impact of the size of a training set on the performance of a QSAR model, 10%, 25%, and 67% of the compounds were randomly selected from the combined data set to train the QSAR models and to predict the PLD activities of the test sets containing the remaining 90%, 75%, and 33% compounds, respectively.

Molecular Descriptors. Customized atom types[13] and MOE 2D descriptors[14] are employed in this study as structural descriptors for model constructions. Atom types are assigned in the context of their local structural environment to capture such features as aromaticity, ring membership, and neighboring functional groups. A number of correction factors are added to depict the global properties, including polar surface area (PSA), ratio of ring atoms, count of rotatable bonds, etc. The atom type descriptors contain 218 atom types and 36 correction factors.[13] The 186 MOE 2D descriptors largely consist of physicochemical properties, subdivided surface areas, connectivity and shape indices, and atom and bond counts, featuring whole-molecule properties.

Support Vector Machines (SVM). SVM, a kernel-based algorithm, is one of the few statistical algorithms that take into consideration the generalization problem.[15−17] SVM minimizes generalization errors by pursuing the maximum separation of two classes. The algorithm has been widely applied in chemoinformatics to tackle difficult quantitative structure−activity relationship (QSAR) problems and has been proven in many cases to outperform other algorithms.[7,18,19] In this study, LIBSVM, a software implementation of SVM developed by Chang and Lin, was used in model construction.[20] The kernel function used throughout this study is the Gaussian Radial Basis Function (RBF)

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \| \mathbf{x}_i - \mathbf{x}_j \|^2)$$

The tunable parameters, C, measuring the penalty against misclassified compounds,[20] and γ, indicating the nonlinearity of the kernel, were optimized using an exhaustive searching method. A Python-driven grid-based method was applied to maximize the prediction accuracy in a 7-fold cross-validation (CV) of the training data. To evaluate the performance of the predictive models, receiver operating characteristic (ROC) curves were employed, where sensitivity is plotted against 1-specificity. The area under the ROC curve (AUC-ROC) represents the relative trade-offs between benefits and costs of a binary classifier, which is commonly used for evaluating performance of classification models, because of its simplicity and insensitivity to class skewness.[21]

RESULTS AND DISCUSSION

Data Quality. qHTS provided high quality and rich data sets because each compound was tested at multiple concentrations, and the CRCs are fitted through an automatic procedure to produce curve classes.[10,22] In this study, 133 compounds were replicated in these three library collections, LOPAC, NPC, and Tocris, and thus tested three to five times. Among 133 compounds, 120 compounds (90%) exhibited identical curve classes in all replicates. There were another 504 compounds that were duplicated and tested twice, and 475 of them, or 94%, showed the same activity in both copies. The high consistency of assay results is a good indicator of high data quality.

It is of interest to compare the cell-based PLD activities with *in vivo* data. Among the 385 compounds with structures in Kruhlak's data set,[23] 164 compounds were tested in our cell-based PLD assay, and 136 of them, or 83%, exhibited the similar activities *in vivo*. Given the differences between cell-based and *in vivo* assays, such as metabolism[24] and tissue specificity, that may influence compound activities, the consistency level between *in vitro* and *in vivo* activities observed here is considerably high. Higher consistency (91%) was observed for PLD inactive compounds than the PLD actives (56%). Figure 1 depicts six compounds that showed different activities in the two assays, where the results of the cell-based assay appeared to be more reasonable, on the basis of the previously summarized pharmacophore of PLD inducing molecules.[3]

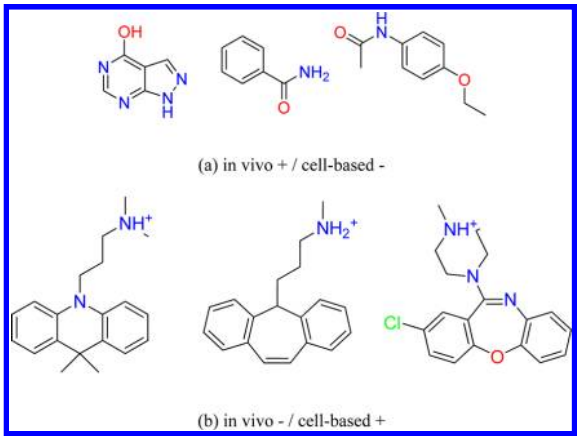


Figure 1. Example structures with mismatched assays results in cell-based and *in vivo* assays: (a) PLD active in in vivo assay but inactive in cell-based assay and (b) PLD inactive in in vivo but active in cell-based assay.

Simple Physicochemical Properties Are Insufficient To Separate PLD Active from Inactive. Because of distinct structural features shared by PLD inducing drugs, calculated physicochemical properties have been commonly used to predict PLD inducing potential.[5,6,8] Plotting ClogP against calculated pK_a demonstrated a certain level of discerning power, while the more sophisticated "rule of thumb", i.e. a compound was more likely to be PLD active if $(\text{ClogP})^2 + (\text{p}K_a)^2 > 90$, given that ClogP > 1 and p*Ka* > 8,[5] provided better predictions. A higher specificity level can be achieved by modifying the rule — a compound is PLD active if $(\text{ClogP})^2 + (\text{p}K_a)^2 > 50$, given that ClogP ≥ 2 and p*Ka* ≥ 6.[8] Replacing calculated pK_a with a negative charge (NC), the ClogP − NC plot exhibited a high accuracy of 98% in predicting the PLD inducing potential of 63 drugs, when a simple rule of "ClogP > 1 and NC ≥ 1" was applied to assign PLD active compounds.[6] In this study, the NC of a compound was computed using the molecular modeling software MOE,[14,14] and logP was calculated with an in-house model.[25] The results clearly showed that the above-mentioned simple rule produced many false positives and false negatives when applied to either the training LOPAC set or the N&T test set (Figure 2). In the training set, 40 of 136 PLD active compounds and 228 of 992 PLD inactive compounds were misclassified by the simple rule, providing a low sensitivity and specificity of 70.6% and 77.0%, while the sensitivity and specificity reached 67.9% and 81.9%, respectively, in the test set. In summary, using a few calculated physicochemical properties does not appear to have sufficient power to differentiate PLD actives from inactives when a large data set is involved.

Effect of Training Data and Molecular Descriptors on Model Performance. SVM models were built using two different training sets and two sets of different molecular descriptors. Using LOPAC as the training data (12.1% PLD actives), MOE 2D descriptors outperformed atom types in predicting the PLD activities of the test set compounds (8.1% PLD actives), as shown in Table 1. However, the difference

Table 1. Performance of Predictive Models Using Different Training Data and Molecular Descriptors, As Measured by the AUC-ROC

	training data	
descriptor	LOPAC	random
atom type	0.83	0.87
MOE 2D	0.87	0.87

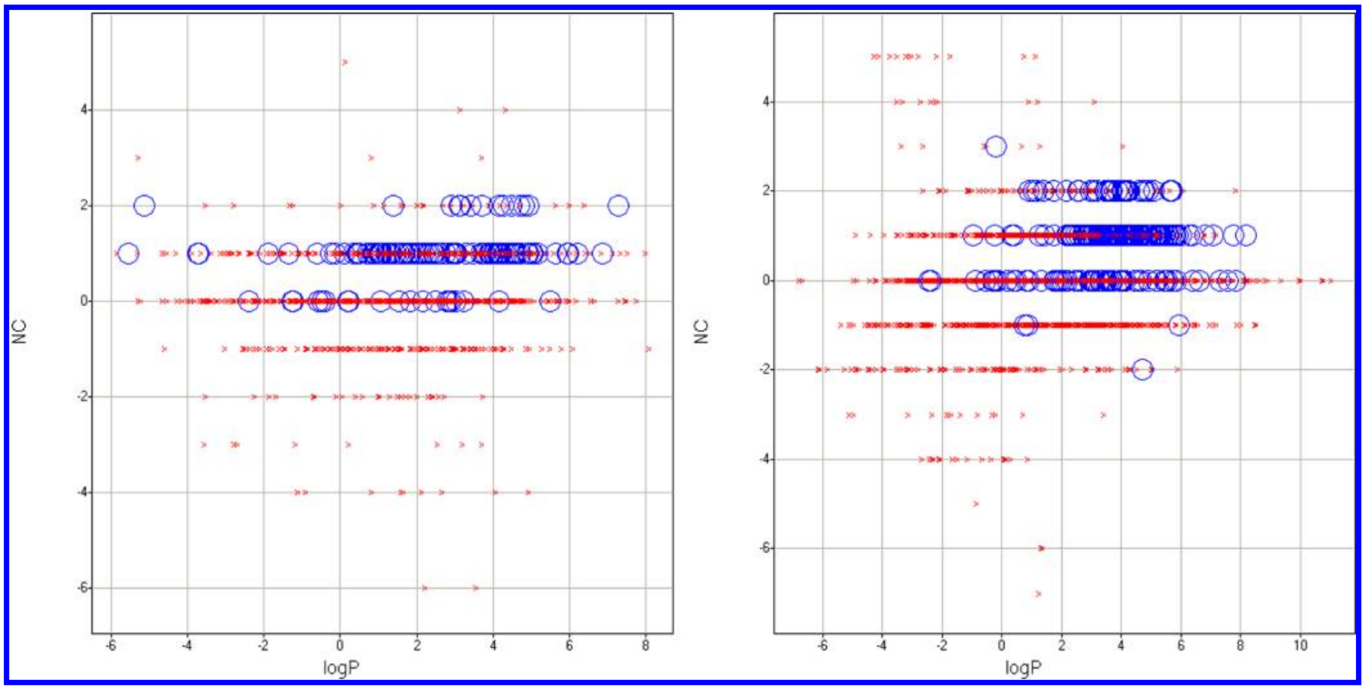


Figure 2. Scatter plots of calculated logP vs net charge (NC) for (a) the LOPAC training set and (b) the N&T test set, where PLD active compounds are depicted as blue open circles and PLD inactive compounds are red crosses.

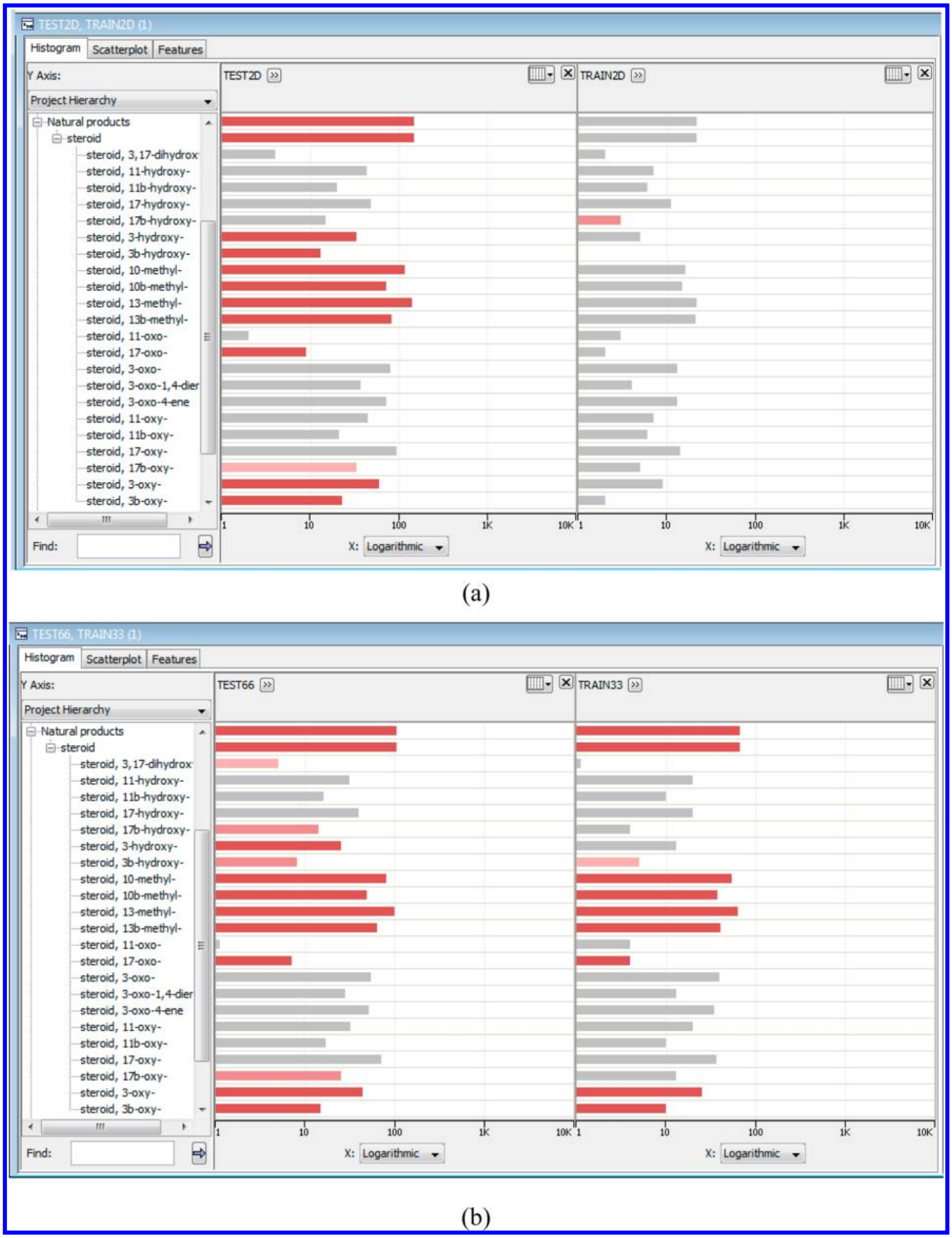


Figure 3. Side-by-side comparison of structural features in both training and test sets for (a) LOPAC training data and (b) random training data. The histogram bars are color-coded based on the feature Z-scores, while their lengths depict the frequencies of structural classes on a log scale.

disappeared when one-third of the compounds randomly selected from the combined data set were employed as training data. Adding new features to the atom type descriptor set, including calculated logP, net charge, and the top 10 MOE 2D descriptors, did not improve the model performance in terms of the AUC-ROC values. Combination of atom types and all MOE 2D descriptors yielded a model with an AUC-ROC of 0.86, slightly inferior to the model using MOE 2D alone. The results implied that atom type descriptors carried some specific structural information which was different in the LOPAC training and N&T test sets, and the difference might have biased the model toward the training set and rendered it less applicable to compounds in the test set. When the training and test compounds were separated according to chemical features using the software Leadscope,[26] it became clear that both the feature distribution (the relative lengths of the bars in Figure 3) and the Z-score (the colors of the bars) were quite different for the LOPAC and the N&T compounds, as illustrated by the distribution of the steroids, for example, in both data sets (Figure 3a). Indeed, there are nearly 60% of LOPAC compounds that target neurotransmitter signaling, while the N&T compounds have more diverse effects on the targets.[9] On the contrary, the training compounds randomly selected from the three collections well represented the test compounds, as shown in Figure 3b. The percentages of PLD active compounds were 9.4% and 9.1% for the training and test sets, which were more balanced than the LOPAC/N&T data sets as training and test sets. Principal component analysis (PCA) of the atom type descriptors supplied further support to the observation that the randomly selected training data had better coverage of the chemical space of the test set (Figure S1). Different from the atom types, MOE 2D descriptors cover mostly whole molecule properties, which are much less sensitive to the detailed structural features carried by the atom types, therefore, reselecting the training data did not make any difference in the performance of the corresponding models (Table 1). In this particular case, whole molecule properties, such as the logP and subdivided surface areas, rather than specific structural features, seemed to be more suitable for describing the nonspecific interactions between small molecules and phospholipids that cause PLD.

Using 33% of the compounds randomly selected from the entire data set as training data, the models built from atom types and MOE 2D descriptors shared the same AUC-ROC value of 0.87 (Table 1) in predicting the test set but with slightly different sensitivity and specificity. About 95% of PLD active compounds were included in the top 50% of compounds ranked using atom type model (green line in Figure 4) and the number dropped to 90% for the MOE 2D model (blue line in Figure 4). Combination of atom types and MOE 2D descriptors (combined model) increased the AUC-ROC value to 0.89, while the consensus probability calculated by averaging the probability of both models from atom types and MOE 2D descriptors offered a slightly better AUC-ROC value of 0.90. The combined model (red line in Figure 4) can identify all of the PLD active compounds by screening <50% the whole test set, while the consensus model can identify more PLD actives by screening the first 10% and 20% of the test set (Figure 4).

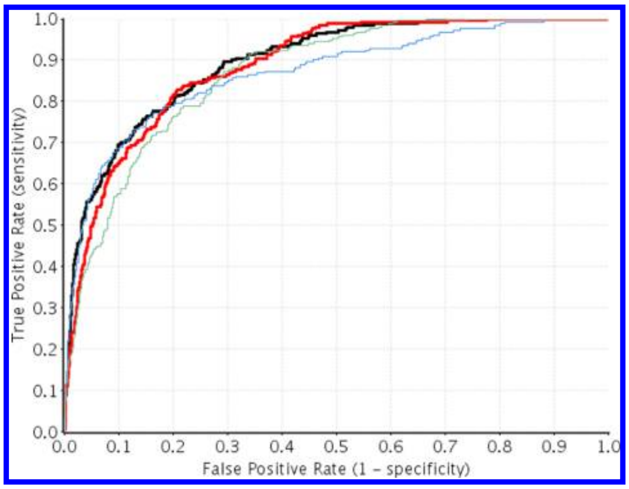


Figure 4. The ROC curves for the models derived using atom types (green line), MOE 2D descriptors (blue line), both atom types and MOE 2D descriptors (red thick line), and the consensus model (black thick line).

The validity of the SVM models was examined using Y-scrambling. The predictive power deteriorated sharply when the PLD activities in the training sets were randomly scrambled. For example, the AUC-ROC value dropped to 0.54, which was equivalent to a random model, from 0.87 in the 33%/67% training/test case (Figure 5).

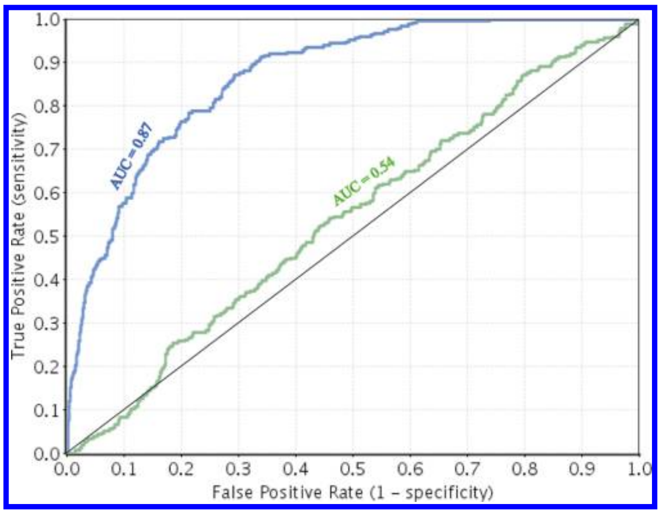


Figure 5. The ROC curves for the models of 33%/67% random division set before (blue line) and after (green line) Y-scrambling.

Impact of the Size of Training Data. Increment of sizes of training data will generally increase the coverage of feature space, and in turn, will improve the predictive power of the resulting models. This trend is, in general, independent of the molecular descriptors used to construct the models. When the N&T compounds were used to train the SVM models, the models predicted the PLD potential of the LOPAC compounds with improved accuracy of 0.88 and 0.91, respectively, as measured by the AUC-ROC, for the atom type and MOE 2D descriptors (Table 2). If the structural bias was reduced by randomly splitting the combined data set into training and test sets, the impact of the training data size became minimal — no significant difference in model performance was observed when one-third or two-thirds of the compounds were used as training data (Table 2). Further reducing the size of the training data to 25% and 10% of the total compounds did not impair the predictive power of the

Table 2. AUC-ROC Values for the Models Using Different Training Data and Molecular Descriptors

	training data	
descriptor	LOPAC/N&T	33%/67%
atom type	0.83/0.88	0.87/0.88
MOE 2D	0.87/0.91	0.87/0.88

derived models (Table 3). The results demonstrated that the predictive power fell into the same range when using 10% up to 67% of the data set to train the models, as long as the training compounds were randomly selected from the data set. It is interesting to note that the model trained by only 433 randomly selected compounds outperformed the one trained by 1,128 LOPAC compounds. These results suggest that the diversity and thus, structural coverage of the training data, has a greater impact on model performance than the size of the training data. This observation is in good agreement with the training set selection criteria outlined by other researchers,[27,28] albeit the approaches necessary to achieve diverse training sets vary depending on the size of the available data. Leonard and Roy[27] proposed the use of *K*-means clustering to maximize the diversity of the training set selected from entire data sets containing 35, 56, and 87 compounds, while the random division method failed to offer consistency in generating predictive models. In this study, random selection of 10% of data to serve as the training set was repeated three times, and the AUC-ROC values of the resulting models were 0.85, 0.86, and 0.86, respectively. Similarly, random selection of 25% of the data set as training data was also repeated three times, and the three resulting models again showed comparable performances with AUC-ROC of 0.85, 0.88, and 0.88. In conclusion, random division of a large data set into training and test sets offered sufficient diversity in the training data, leading to highly consistent model performance. The fact that as low as 10% of total data set is sufficient to train a model to perform equally as well as those trained with 33% and 67% of the data set indicates that the performance of a model cannot be continuously improved through increasing the size of training data, as the predictive accuracy is limited by the feature space and the experimental variations in the data.

Table 3. AUC-ROC Values for the Models Trained with Different Fraction of Randomly Selected Compounds

	training data			
percentage	67%	33%	25%	10%
count of cmpds	2766	1395	1056	433
AUC-ROC	0.88	0.87	0.87	0.86

Rebalance of a Skewed Data Set.

The PLD data set is heavily imbalanced with less than 10% of PLD active compounds. Imbalanced responses are commonly observed in biological data, such as toxicity data, while how to effectively learn from imbalanced data sets remains a challenge to the QSAR and machine learning community. Recently, substantial research has been carried out to develop strategies to efficiently learn from imbalanced data sets.[7,29] Among these strategies, including undersampling, oversampling, cost sensitivity, consensus, and ensemble learning, rebalancing a training set through undersampling the majority class exhibited the capability of consistently enhancing the performance of predictive models.[30,31]

Random undersampling was carried out as follows: the PLD inactive compounds in the training sets were randomly split into six equal-sized subsets, each subset was then combined with all the PLD active compounds to form six new training sets. The newly generated training sets were better balanced than the original training sets for both LOPAC/N&T and the randomly split data sets (Table 4). The results showed that most of the undersampled training sets outperformed the original imbalanced training sets in both cases, despite the fact that each undersampled training set used only a fraction of the PLD inactive compounds. The consensus models, where the consensus probability of a compound being PLD active was computed by averaging over the probabilities of the six individual rebalanced models, outperformed each individual undersampled model in both cases (Table 4). Figure 6 depicts the ROC curves of six individual rebalanced models together with the consensus model for the random division of 33%/67% case. The top 10% of compounds ranked by consensus probability contained nearly 70% of all the PLD actives, while the top 30% ranked compounds in the whole test set recovered 93% of all PLD actives, a percentage approaching the level of experimental errors, as estimated by the duplicated compounds in the data sets. The results suggested that the consensus model was superior to the individual models, since the consensus model exhibited higher AUC values than the individual models and it made use of the structural information of all the compounds in the training set.

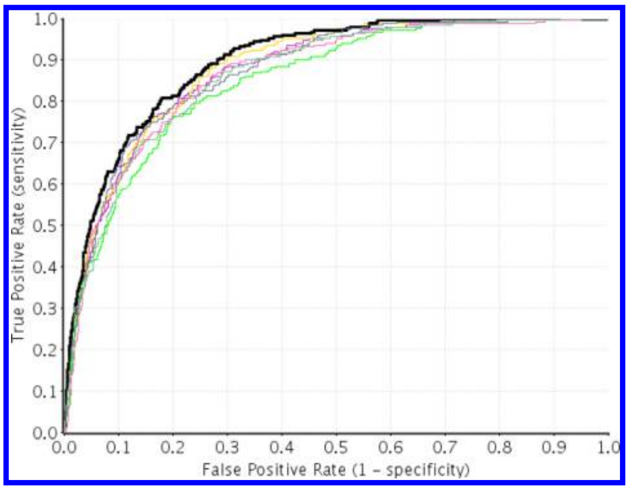


Figure 6. The ROC curves of the six rebalanced models in rainbow colors and the consensus model in a black thick line for random division data set (33%/67%).

Sensitivity and Specificity of the SVM Models.

Sensitivity and specificity are frequently employed to assess performances of binary classifiers, in terms of their ability to

Table 4. Percentages of PLD Active Compounds in the Training Sets and the AUC-ROC Values of the Models Derived Using the Original Imbalanced Training Data and the Six Rebalanced Training Sets for the LOPAC/N&T and the Random Division of 33%/67% Data Sets

		original	T1	T2	T3	T4	T5	T6	consensus
LOPAC/N&T	percentage of actives	12.6%	47.4%	48.9%	44.3%	45.2%	40.2%	45.8%	
	AUC-ROC	0.83	0.83	0.83	0.83	0.85	0.84	0.83	0.85
33.3%/66.6%	percentage of actives	9.4%	40.9%	41.3%	37.5%	37.9%	34.3%	39.0%	
	AUC-ROC	0.87	0.87	0.87	0.87	0.85	0.88	0.88	0.90

recognize true positives and negatives, respectively. As the sensitivity and specificity of a model depend on the predictor value used to make active or inactive calls, when imbalanced data are involved, adjusting the predictor value cutoff will have a significant impact on the resulting sensitivity and specificity.[32] Table 5 summarized such an impact on the SVM models before and after rebalancing, where the training set was 33% of randomly selected compounds. Without rebalancing, the classifier strongly biased toward the majority class (PLD negative, or specificity in this case), which is commonly observed in machine learning of heavily skewed data sets. Lowering the cutoff to 0.05 increased the sensitivity to 92.0% with specificity decreased to 63.7%. The rebalanced model re-established the balance between sensitivity and specificity at moderate thresholds − a high sensitivity of 92.4%, which is highly desired in predictive toxicology, was reached at the optimal cutoff of 0.29.

Table 5. Summary of the Impact of Decision Thresholds on the Sensitivity and Specificity of the SVM Classification Models before and after Rebalancing[b]

	threshold	AUC-ROC	sensitivity	specificity
prerebalance	0.06[a]	0.87	87.6%	69.4%
	0.2		45.0%	93.8%
	0.3		37.1%	96.7%
	0.4		23.5%	98.1%
	0.5		18.7%	99.0%
after-rebalance	0.2	0.90	96.4%	59.4%
	0.29[a]		92.4%	73.0%
	0.3		89.2%	74.1%
	0.4		79.7%	82.6%
	0.5		68.5%	88.5%

[a] The optimal cutoffs computed from the ROC curves. [b] The consensus model of the six random undersampling models was compared with the original model, where the training data was the random division of 33%/67% data set.

Interpretability of the SVM Models. One benefit of utilizing atom types as molecular descriptors is its interpretability. Feature selection[33] can be applied either to select a subset of descriptors for model construction or to rank the order of importance of features in terms of their contributions to the model. The top five important features extracted from the LOPAC training set are N16 (a positively charged nitrogen atom in a saturated ring, such as piperidine or piperizine), C3 (an aromatic carbon atom with no substitution and adjacent to two aromatic carbon atoms), H2 (a hydrogen bonded to an aromatic carbon), M12 (a number of aromatic rings), and S5 (sulfur in a ring bonded to two aromatic atoms).[25] These features summarized the two key characters of PLD inducing compounds − an aromatic moiety and a positive charge center. As a result, the two commonly used strategies to minimize the PLD potential of a compound are reducing its hydrophobicity and lowering its basicity.[34] Compounds containing two N16 nitrogen atoms have a > 60% chance of inducing PLD, and the probability dropped sharply to 7.8% for compounds without an N16 atom (Figure 7). Piperizine and piperidine are two moieties frequently used by medicinal chemists to reduce molecular flexibility and hydrophobilcity. However, both structure features have high PLD inducing potential, which could be alleviated by reducing the basicity of the ring nitrogen atoms with neighboring aromatic rings or carbonyl groups.

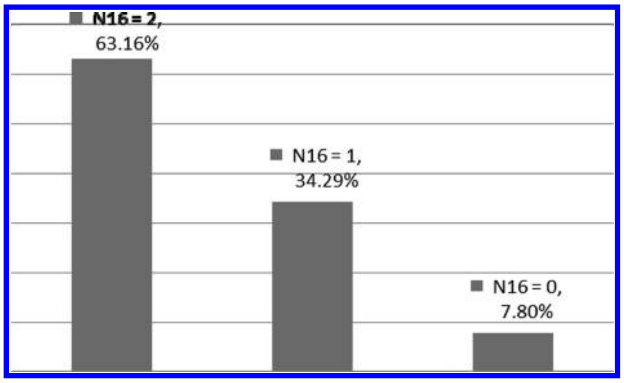


Figure 7. The declining trend of probability of being PLD active with decreasing count of N16 nitrogen atoms in the molecule.

All 11 compounds in LOPAC containing S5 sulfurs were PLD positive (Figure S2); however, only 23 of the 30 S5-containing compounds in the test set were PLD active. The rule learned from the training data that S5-containing compounds are PLD active is thus inapplicable to one-third of the S5-containing compounds in the test set. This example illustrates why the atom typing model trained from the LOPAC data set was less accurate than the MOE 2D model. Another important lesson learned is that machine learning is limited by the training data fed to the machine. All the compounds containing S5 in the training set happened to be CADs, but the presence of the atom type S5 is not necessarily linked to their PLD activity; for example, many compounds in the test set that contain the S5-containing tricyclic moiety without the aliphatic tail were neither CAD nor PLD active. The modeling algorithm by itself can hardly overcome this problem of chance correlation, but increasing the size and, more importantly, diversity of the training sets can help to reduce the probability of chance correlation.

■ SUMMARY

The demand of effective reduction of liability associated with drug induced PLD in the early stages of drug discovery calls for accurate prediction of PLD inducing potential of candidate drug molecules. Predictive models have been constructed based on the largest PLD data set known to date containing over four thousand nonredundant compounds. The SVM models built from atom types were sensitive to the structural feature coverage of the test data by the training data, while the models built from the whole-molecule based MOE 2D descriptors were less sensitive. The models trained with 10% of randomly selected compounds consistently achieved high accuracy comparable to those trained with up to 67% of the data set, implying the strong learning power of the SVM algorithm and the slim possibility of enhancing the predictive accuracy of a model through adding the redundancy of the structural features by increasing the size of training data. Our results also suggested that increment of structure diversity of a training set was more important than increasing its size in achieving high predictive performance. When one-third of the data set was randomly selected to train the model, most of the randomly undersampled models outperformed the one based on the original imbalanced training set, and the consensus model of the six individual undersampled models obtained an AUC-ROC accuracy of 0.90 in predicting the remaining two-thirds of data. The chemical features showing the highest discerning power are mostly in good agreement with those shared by CADs, yet chance correlation was observed in one atom type. The negative impact of chance correlation was minimized by random division of data into training and test sets.

ASSOCIATED CONTENT

Supporting Information

This work was supported by the Intramural Research Programs of the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 301-217-4675. Fax: 301-217-5736. E-mail: huangru@mail.nih.gov (R.H.), sunh7@mail.nih.gov (H.S.). Corresponding author address: NIH Chemical Genomics Center National Institutes of Health, 9800 Medical Center Drive, Rockville, MD 20850.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank in particular Anna Rossoshek for helpful comments and suggestions during the preparation of this manuscript.

REFERENCES

(1) Nonoyama, T.; Fukuda, R. Drug-induced Phospholipidosis-Pathological Aspects and Its Prediction. *J. Toxicol. Pathol.* **2008**, *21*, 9−24.

(2) Anderson, N.; Borlak, J. Drug-induced phospholipidosis. *FEBS Lett.* **2006**, *580*, 5533−40.

(3) Lullmann, H.; Lullmannrauch, R.; Wassermann, O. Lipidosis Induced by Amphiphilic Cationic Drugs. *Biochem. Pharmacol.* **1978**, *27*, 1103−1108.

(4) Kruhlak, N. L.; Choi, S. S.; Contrera, J. F.; Weaver, J. L.; Willard, J. M.; Hastings, K. L.; Sancilio, L. F. Development of a phospholipidosis database and predictive quantitative structure-activity relationship (QSAR) models. *Toxicol. Mech. Methods* **2008**, *18*, 217−227.

(5) Ploemen, J. P. H. T. M.; Kelder, J.; Hafmans, T.; Van De Sandt, H.; van Burgsteden, J. A.; Salemink, P. J. M.; Van Esch, E. Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential - A case study with structurally related piperazines. *Exp. Toxicol. Pathol.* **2004**, *55*, 347−355.

(6) Tomizawa, K.; Sugano, K.; Yamada, H.; Horii, I. Physicochemical and cell-based approach for early screening of phospholipidosis-inducing potential. *J. Toxicol. Sci.* **2006**, *31*, 315−24.

(7) Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)* **2008**, *8*, 1691−709.

(8) Pelletier, D. J.; Gehlhaar, D.; Tilloy-Ellul, A.; Johnson, T. O.; Greene, N. Evaluation of a published in silico model and construction of a novel Bayesian model for predicting phospholipidosis inducing potential. *J. Chem. Inf. Model.* **2007**, *47*, 1196−205.

(9) Huang, R.; Southall, N.; Wang, Y.; Yasgar, A.; Shinn, P.; Jadhav, A.; Nguyen, D. T.; Austin, C. P. The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med.* **2011**, *3*, 80ps16.

(10) Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Quantitative high-throughput screening: a titration-based approach that efficiently identifies biological activities in large chemical libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11473−8.

(11) Bhandari, N.; Figueroa, D. J.; Lawrence, J. W.; Gerhold, D. L. Phospholipidosis assay in HepG2 cells and rat or rhesus hepatocytes using phospholipid probe NBD-PE. *Assay Drug Dev. Technol.* **2008**, *6*, 407−19.

(12) Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/ (accessed April 2, 2012).

(13) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for cytochrome p450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51*, 2474−81.

(14) MOE. http://www.chemcomp.com (accessed April 2, 2012).

(15) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, 2005.

(16) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.

(17) Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons: New York, 1998.

(18) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and Non-inhibitors Using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996−1011.

(19) 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*, 982−92.

(20) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines* **2001**.

(21) Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861−874.

(22) Wang, Y.; Jadhav, A.; Southal, N.; Huang, R.; Nguyen, D. T. A grid algorithm for high throughput fitting of dose-response curve data. *Curr. Chem. Genomics* **2010**, *4*, 57−66.

(23) Kruhlak, N. L.; Choi, S. S.; Contrera, J. F.; Weaver, J. L.; Willard, J. M.; Hastings, K. L.; Sancilio, L. F. Development of a phospholipidosis database and predictive quantitative structure-activity relationship (QSAR) models. *Toxicol. Mech. Methods* **2008**, *18*, 217−27.

(24) Kuroda, Y.; Saito, M. Prediction of phospholipidosis-inducing potential of drugs by in vitro biochemical and physicochemical assays followed by multivariate analysis. *Toxicol. in Vitro* **2010**, *24*, 661−668.

(25) Sun, H. A universal molecular descriptor system for prediction of logP, logS, logBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 748−57.

(26) Ward, B.; Juehne, T. Combinatorial library diversity: probability assessment of library populations. *Nucleic Acids Res.* **1998**, *26*, 879−86.

(27) Leonard, J. T.; Roy, K. On selection of training and test sets for the development of predictive QSAR models. *QSAR Comb. Sci.* **2006**, *25*, 235−251.

(28) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−69.

(29) Rodriguez Sarmiento, R. M.; Nettekoven, M. H.; Taylor, S.; Plancher, J. M.; Richter, H.; Roche, O. Selective naphthalene H(3) receptor inverse agonists with reduced potential to induce phospholipidosis and their quinoline analogs. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4495−500.

(30) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Tice, R. R.; Kavlock, R. J.; Huang, R. Prediction of Cytochrome P450 Profiles of Environmental Chemicals with QSAR Models Built from Drug-like Molecules. Submitted to *Mol. Inf.*

(31) Drummond, C.; Holte, R. In *C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling*, Proceedings of the ICML'03 Workshops on Learning from Imbalanced Data Sets, 2003; 2003.

(32) Chawla, N. V.; Japkowicz, N.; Kotcz, A. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsl.* **2004**, *6*, 1−6.

(33) Chen, Y.-W.; Lin, C.-J. Combining {SVM}s with various feature selection strategies. In *Feature extraction, foundations and applications*; Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., Eds.; Springer: 2006.

(34) Ratcliffe, A. J. Medicinal Chemistry Strategies to Minimize Phospholipidosis. *Curr. Med. Chem.* **2009**, *16*, 2816−2823.