molecular pharmaceutics

# Predicting Phospholipidosis Using Machine Learning

**Robert Lowe,[†] Robert C. Glen,[†] and John B. O. Mitchell*[,‡]**

*Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, U.K., and Centre for Biomolecular Sciences, University of St Andrews, North Haugh, St Andrews, Scotland, KY16 9ST, U.K.*

Received April 8, 2010; Revised Manuscript Received August 24, 2010; Accepted August 27, 2010

**Abstract:** Phospholipidosis is an adverse effect caused by numerous cationic amphiphilic drugs and can affect many cell types. It is characterized by the excess accumulation of phospholipids and is most reliably identified by electron microscopy of cells revealing the presence of lamellar inclusion bodies. The development of phospholipidosis can cause a delay in the drug development process, and the importance of computational approaches to the problem has been well documented. Previous work on predictive methods for phospholipidosis showed that state of the art machine learning methods produced the best results. Here we extend this work by looking at a larger data set mined from the literature. We find that circular fingerprints lead to better models than either E-Dragon descriptors or a combination of the two. We also observe very similar performance in general between Random Forest and Support Vector Machine models.

**Keywords:** Phospholipidosis; machine learning; Random Forest; Support Vector Machine; in silico; prediction

## Introduction

Phospholipidosis was first believed to be observed by Nelson and Fitzhugh in 1948,[1] when they reported the accumulation of foam macrophages in rats after long-term treatment with chloroquine. It has been observed since then that numerous cationic amphiphilic drugs can induce phospholipidosis in several cell types and that it can be characterized by the excess accumulation of phospholipids.[2−4] Electron microscopy is the most reliable method of identifying whether a compound has induced phospholipidosis, by the presence of lamellar inclusion bodies.[5] However, it may also be identified by light microscopy in which cells appear vacuolated and contain foamy macrophages.[6] The observation of compound-induced phospholipidosis in the drug development process is considered manageable as the effect often only occurs at very high doses, many times that of the intended therapeutic dose.[7] There is at present no strong evidence that that the condition is harmful to human health, and it is reversible once treatment is terminated, the drug is expelled from the cell and phospholipid levels return to

* To whom correspondence should be addressed. Mailing address: Centre for Biomolecular Sciences, University of St Andrews, North Haugh, St Andrews, Scotland, KY16 9ST, U.K. Phone: +44 1334 467259, Fax: +44 1334 462595, E-mail: jbom@st-andrews.ac.uk.

[†] University of Cambridge.

[‡] University of St Andrews.

(1) Nelson, A. A.; Fitzhugh, O G. Chloroquine: Pathological changes observed in rats which for two years had been fed various proportions. *Arch. Pathol.* **1948**, *45*, 454–462.

(2) Halliwell, W. H. Cationic amphiphilic drug-induced phospholipidosis. *Toxicol. Pathol.* **1997**, *25*, 53–60.

(3) Lüllmann, H.; Lüllmann-Rauch, R.; Wassermann, O. Lipidosis induced by amphiphilic cationic drugs. *Biochem. Pharmacol.* **1978**, *27*, 1103–1108.

(4) Reasor, M. J. A review of the biology and toxicologic implications of the induction of lysosomal lamellar bodies by drugs. *Toxicol. Appl. Pharmacol.* **1989**, *97*, 47–56.

(5) 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*, 315–324.

(6) Reasor, M. J.; Kacew, S. Drug-Induced Phospholipidosis: Are There Functional Consequences. *Exp. Biol. Med.* **2001**, *226*, 825–830.

(7) 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*, 1196–1205.

10.1021/mp100103e © 2010 American Chemical Society
Published on Web 08/27/2010

normal.[8] This process can take weeks, however, and in some cases has been reported to last several months. It can be especially important in the context of the nervous system, where phospholipids may disrupt cell signaling in neurons and could possibly be linked to several genetic diseases such as Niemann−Pick disease.[8,9] The occurrence of phospholipidosis in the drug development process therefore can cause delays as more tests need to be carried out to satisfy regulatory bodies. It is also possible that the occurrence may sometimes stop the drug development process altogether. A recent minireview[6] shows that the method by which compounds induce phospholipidosis is still not well understood and indeed suggests that the underlying mechanism is not exactly the same for each compound. The most common mechanism is the inhibition of lysosomal phospholipase activity leading to the accumulation of several classes of phospholipids;[3,10] it has also been shown, however, that an increase in synthesis of acidic phospholipids may occur leading to the "redirection of phospholipid synthesis".[11,12]

The application of an in silico model for predicting phospholipidosis to produce an accurate and fast method could be of great use to the pharmaceutical industry, where early screening is of great importance. Indeed, this has already been recognized with many early attempts at doing just this. First attempts at producing such a model by Ploemen et al.[13] used p$K_a$ and ClogP. Here Ploemen et al. suggested that a compound would be phospholipidosis-inducing (PPL+) provided that p$K_a$ > 8 and ClogP > 1 and also that the inequality, eq 1, is satisfied.

$$(\text{ClogP})^2 + (\text{calculated p}K_a)^2 > 90 \qquad (1)$$

Another simple model was suggested by Tomizawa et al.,[5] in which a modification to the Ploemen model involved replacing p$K_a$ with NC, the sum of the charge of all dissociable functional groups in a molecule. A tree model was created to predict phospholipidosis based on these two features. This Ploemen model was further tested by Pelletier et al.[7] with a larger data set, and it was shown that improvements in the rules could be made to produce better predictivity. The inclusion of additional descriptors and the use of a Bayesian model also produced even better predictivity than that of the modified Ploemen model. Kruhlak et al.[14] used two commercially available software packages, MC4PC and MDL-QSAR, to produce predictive models for phospholipidosis and on a 10-fold cross-validation test produced {76% positive, 78% negative predictivity} and {65% positive, 87% negative predictivity}, respectively. All these approaches are based on relatively simple models, and as phospholipidosis is clearly a complex effect, applying state of the art machine learning techniques could therefore produce even better predictive results.

Ivanciuc[15] produced a comprehensive list of state of the art machine learning techniques and their ability to predict phospholipidosis, using the nonproprietary data from the Pelletier data set; the models were tested on a 10-fold cross validation. The best model for prediction was a Support Vector Machine[16] with an RBF kernel and $\gamma = 0.01$, producing on the validation fold 97% accuracy and 0.94 Matthews Correlation Coefficient. Here we propose testing the machine learning models on a larger data set than the Pelletier one.

## Phospholipidosis Database

A phospholipidosis database was created from various literature sources. It was mainly created from a combination of two data sets producing a total of 185 compounds, of which 102 were positive for phospholipidosis (PPL+) and 83 were negative (PPL−). The literature mined data from the Pelletier data set were used as the basis for our data set of 117 compounds, and this was supplemented by data taken from Kruhlak et al.,[14] consisting of compounds compiled from 12 other sources. The Kruhlak et al. data include compounds marked as negative solely due to the absence of a reported positive result, and these compounds, which may in fact be untested, were excluded from our database to reduce the possibility of erroneous data. This means that all of the 83 PPL− compounds were reported negative by electron microscopy. Out of the 102 PPL+ compounds, 34 of these compounds are reported positive by the presence of foamy macrophages or vacuolations and the remaining

[8] Nioi, P.; Perry, B. K.; Wang, E.-J.; Gu, Y.-Z.; Snyder, R. D. In Vitro Detection of Drug-Induced Phospholipidosis Using Gene Expression and Fluorescent Phospholipid Based Methodologies. *Toxicol. Sci.* **2007**, *99*, 162–173.

[9] Sawada, H.; Takami, K.; Asahi, S. A Toxicogenomic Approach to Drug-Induced Phospholipidosis: Analysis of Its Induction Mechanism and Establishment of a Novel in Vitro Screening System. *Toxicol. Sci.* **2005**, *83*, 282–292.

[10] Reasor, M. J.; McCloud, C. M.; Beard, T. L.; Ebert, D. C.; Kacew, S.; Gardner, M. F.; Aldern, K. A.; Hostetler, K. Y. Comparative evaluation of amiodarone-induced phospholipidosis and drug accumulation in Fischer-344 and Sprague-Dawley rats. *Toxicology* **1996**, *106*, 139–147.

[11] Eichberg, J.; Gates, J.; Hauser, G. The mechanism of modification by propranolol of the metabolism of phosphatidyl-CMP (CDP-diacylglycerol) and other lipids in the rat pineal gland. *Biochim. Biophys. Acta* **1979**, *573*, 90–106.

[12] Pappu, A.; Hostetler, K. Y. Effect of cationic amphiphilic drugs on the hydrolysis of acidic and neutral phospholipids by liver lysosomal phospholipase A. *Biochem. Pharmacol.* **1984**, *33*, 1639–1644.

[13] 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.

[14] 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.

[15] Ivanciuc, O. Weka machine learning for predicting the phospholipidosis inducing potential. *Curr. Topics Med. Chem.* **2008**, *8*, 1691–1709.

[16] Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learning* **1995**, *20*, 273–297.

68 of these compounds were confirmed by electron microscopy. The phospholipidosis-inducing compounds have been observed to act on a variety of species including humans, rats, mice, dogs, rabbits, hamsters and monkeys as well as a variety of tissue types including lungs, kidney and liver. A full breakdown of the positive compounds in the Pelletier data set is shown in Table 1 of Pelletier et al.[7] Negative compounds from the Pelletier data set included "druglike" molecules searched from the literature, contrasting with the Kruhlak data included in our data set, in which the negative compounds were all drugs. As reported by Perez,[17] a good measure of the dissimilarity is the average value of the dissimilarity of the members of the set:

$$\mathrm{div}(A) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} [1 - T(i,j)]/[N(N-1)] \qquad (2)$$

where $T(i, j)$ is the Tanimoto coefficient. Here a value of 1 would represent a very diverse set and a value of 0 would be a set of very similar compounds. The value obtained for our data set was calculated as 0.833 and therefore suggests that we have a reasonable diversity of compounds contained within our data. The structures of the compounds were stored in SMILES[18] format, and missing SMILES strings were obtained from PubChem (http://pubchem.ncbi.nlm.nih.gov) using the name of the compound as a search term. The database is supplied as Supporting Information with structures stored as SMILES.

## Method

The SMILES for each molecule was converted into SDF format and was standardized using tools from ChemAxon.[19] The standardization involved removing fragments, rearomatizing, removing explicit hydrogens and cleaning the 2D structure. Two different sets of descriptors, E-Dragon[20] and Circular Fingerprints,[21] were calculated for the data set. The E-Dragon descriptors were calculated using the online Java program. Circular Fingerprints were calculated to a depth of 2 bonds using a Python script. $pK_a$, ClogP and $(\mathrm{ClogP})^2 + (pK_a)^2$ were also calculated using ChemAxon tools, and these were added as descriptors to both of the original sets of descriptors. This leads to a total of 1,669 descriptors for E-Dragon and 320 for Circular Fingerprints. A final descriptor set was created by combining both of the different descriptor sets, creating a total of 1986 descriptors. A stratified 10-fold cross validation was used for each run: 8 folds were used for training, one was used for internal validation and one was used as a test set. Stratification was used to maintain the proportion of positive and negative compounds in the folds. The test and validation folds were cyclically rotated so that each fold was used once each as a test set and as an internal validation set. This internal validation was used to tune parameters without causing bias in the prediction of the test fold. Feature selection was then performed on the training data using the Weka[22] function SVMAttributeEval with default parameters to select the top 50 features. Attributes are ranked by the square of the weight assigned by SVM.[23] We compare two different Machine Learning algorithms and their power to predict phospholipidosis; the R[24] implementation of Random Forest[25] and the R implementation of Support Vector Machines (SVM)[16] which uses LIBSVM.[26]

The number of trees in Random Forest was set to 1000. The generalization error for forests converges as the number of trees in the forest become large.[25] We can therefore simply choose a large number of trees, and as our data set is small, our computation time is not limiting. The variable, *mtry*, which controls the number of features selected at random at each node, of which the feature providing the best split is chosen, was varied from 1 to 50 with a step size of 1. SVM was run with an RBF kernel meaning that two parameters needed tuning, $\gamma$ and $C$. These were varied by ($2^{-19}$, $2^{-18.75}$, ..., $2^{2.75}$, $2^{3}$) and ($2^{-5}$, $2^{-4.75}$, ..., $2^{14.75}$, $2^{15}$) respectively.

The Matthews Correlation Coefficient (MCC)[27] was used as a measure of predictivity. Its value can range from −1 to 1, where −1 is a perfect anticorrelation, 0 is the equivalent of random guessing and 1 is a perfect correlation. It is used here for comparison between the two algorithms as it is arguably the best single valued metric that describes the confusion matrix of a binary classification problem.

[17] Perez, J. J. Managing molecular diversity. *Chem. Soc. Rev.* **2005**, *34*, 143–152.

[18] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[19] ChemAxon, Standardizer, JChem 5.2.5.1, 2009. http://www.chemaxon.com.

[20] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–463.

[21] Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.

[22] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *SIGKDD Explor.* **2009**, *11*, 10–19.

[23] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learning* **2002**, *46*, 389–422.

[24] R Development Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2006, ISBN 3-900051-07-0.

[25] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

[26] Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines; 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[27] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{3}$$

It can be an especially useful measure where classes are unbalanced as can be seen in eq 3; it takes into account not only true positives (TP) and true negatives (TN) but also false positives (FP) and false negatives (FN). Accuracy, the percentage of correctly predicted instances, is not a good measure of prediction quality for unbalanced classes. For example, when a problem has a very skewed proportion of two classes, predicting each instance to belong to the larger class will lead to a high percentage accuracy. This prediction is, however, not very useful or informative, especially if one is interested in predicting membership of the smaller class correctly. The MCC gives a value close to zero for such uninformative predictions. Indeed, even if either the predicted or real data contain no members of one class, causing one of the four sums in the denominator to go to zero, one can show that the correct limiting value of the MCC is still zero[28] and pragmatically one can specify a minimum value of one for the denominator.[29] Hence the uninformative prediction of the same class for every instance is seen to be no more useful or informative than random guessing. Ten independent runs of both methods were used for each fold, and the average MCC on both the validation and test sets for each of the parameters was calculated. Initially it was seen that a single 10-fold cross-validation run could give large deviations across each individual fold. Therefore in order to get a more realistic value of performance, we repeated the 10-fold cross validation 10 times with different fold definitions. Each fold definition was created using random stratified sampling with a different input seed.

## Results

We report here in Tables 1−3 the average MCC and the standard deviation for each of the 10-fold cross validations performed for the three different descriptor sets used. The parameters are chosen so that the maximum MCC value is obtained for the internal validation. This is done by calculating the MCC of the internal validation for all tested parameters over each of the 10 repeated runs of each fold. This value is then averaged to give a value for the performance on that fold. This is done for all of the 10 folds, in each of the 10 different 10-fold cross validations. These MCC values are then averaged over their respective 10 folds, and the value is summed for each parameter across the 10 different definitions of folds. The parameter which produced the highest averaged MCC value over the different fold definitions was selected. We report the averaged MCC value over each of the 10 repeated 10-fold validations and 10-fold tests, for that parameter. We also report the standard deviation across the 10 folds for each of the fold definitions, $\sigma_T$. $\sigma$ (MCC) is the standard deviation of the MCC of the test set, across the 10 different fold definitions.

***Table 1.*** Average MCC Value for the Internal Validation and Test Set for Each of 10 Different Definitions of a 10-fold Cross Validation for the E-Dragon Descriptors[a]

| | RF (*mtry* = 5) | | | | SVM ($\gamma$ = 0.0003, *C* = 6.727) | | | |
|---|---|---|---|---|---|---|---|---|
| fold definitions | int vals | $\sigma_V$ | test | $\sigma_T$ | int vals | $\sigma_V$ | test | $\sigma_T$ |
| 1 | 0.524 | 0.162 | **0.532** | 0.144 | 0.517 | 0.208 | 0.500 | 0.228 |
| 2 | 0.538 | 0.143 | **0.488** | 0.211 | 0.451 | 0.274 | 0.481 | 0.191 |
| 3 | 0.558 | 0.179 | **0.548** | 0.176 | 0.570 | 0.212 | 0.462 | 0.182 |
| 4 | 0.538 | 0.111 | **0.523** | 0.133 | 0.434 | 0.230 | 0.474 | 0.207 |
| 5 | 0.519 | 0.218 | **0.552** | 0.194 | 0.439 | 0.258 | 0.476 | 0.245 |
| 6 | 0.543 | 0.259 | **0.507** | 0.257 | 0.434 | 0.188 | 0.409 | 0.206 |
| 7 | 0.457 | 0.161 | **0.550** | 0.188 | 0.521 | 0.132 | 0.519 | 0.198 |
| 8 | 0.497 | 0.147 | **0.573** | 0.229 | 0.534 | 0.130 | 0.464 | 0.238 |
| 9 | 0.452 | 0.182 | 0.485 | 0.225 | 0.543 | 0.111 | **0.551** | 0.293 |
| 10 | 0.567 | 0.155 | **0.560** | 0.176 | 0.554 | 0.179 | 0.517 | 0.196 |
| av | 0.519 | 0.172 | 0.532 | 0.193 | 0.500 | 0.192 | 0.485 | 0.218 |

[a] The average reported at the bottom is the averaged value of the columns. $\sigma_V$ represents the standard deviation across the 10 different folds for the internal validation. Similarly, $\sigma_T$ represents the standard deviation for the test set. Highlighted in bold is the highest MCC on the test fold for each fold definition.

***Table 2.*** Average MCC Value for the Internal Validation and Test Set for Each of 10 Different Definitions of a 10-fold Cross Validation for the CFP Descriptors[a]

| | RF (*mtry* = 4) | | | | SVM ($\gamma$ = 0.0110, *C* = 0.841) | | | |
|---|---|---|---|---|---|---|---|---|
| fold definitions | int vals | $\sigma_V$ | test | $\sigma_T$ | int vals | $\sigma_V$ | test | $\sigma_T$ |
| 1 | 0.650 | 0.191 | 0.639 | 0.207 | 0.706 | 0.182 | **0.719** | 0.212 |
| 2 | 0.638 | 0.259 | **0.647** | 0.227 | 0.679 | 0.188 | 0.629 | 0.198 |
| 3 | 0.634 | 0.223 | 0.619 | 0.238 | 0.648 | 0.188 | **0.686** | 0.205 |
| 4 | 0.591 | 0.167 | 0.639 | 0.171 | 0.600 | 0.142 | **0.644** | 0.180 |
| 5 | 0.512 | 0.396 | 0.536 | 0.361 | 0.680 | 0.080 | **0.623** | 0.152 |
| 6 | 0.650 | 0.264 | 0.626 | 0.243 | 0.668 | 0.223 | **0.658** | 0.182 |
| 7 | 0.696 | 0.169 | 0.607 | 0.191 | 0.672 | 0.179 | **0.633** | 0.168 |
| 8 | 0.624 | 0.188 | 0.610 | 0.190 | 0.663 | 0.182 | **0.622** | 0.206 |
| 9 | 0.643 | 0.222 | 0.611 | 0.204 | 0.696 | 0.296 | **0.636** | 0.234 |
| 10 | 0.675 | 0.161 | 0.653 | 0.180 | 0.708 | 0.146 | 0.653 | 0.202 |
| av | 0.631 | 0.224 | 0.619 | 0.221 | 0.672 | 0.181 | 0.650 | 0.194 |

[a] The average reported at the bottom is the averaged value of the columns. $\sigma_V$ represents the standard deviation across the 10 different folds for the internal validation. Similarly, $\sigma_T$ represents the standard deviation for the test set. Highlighted in bold is the highest MCC on the test fold for each fold definition.

The results for the E-Dragon descriptor set are shown in Table 1. Random Forest produces the best result with an averaged MCC of 0.532 across the 10 separate 10-folds. SVM only produces a better result on one of the different fold definitions and produces an averaged MCC of 0.485. Random Forest is also more reliable, producing a smaller averaged standard deviation on the test folds, $\bar{\sigma}_T$, of 0.193 compared to that of SVM. Standard deviations for different fold definitions range from 0.133 to 0.257 for Random Forest

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

[29] Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. O. Chemoinformatics-Based Classification of Prohibited Substances Employed for Doping in Sport. *J. Chem. Inf. Model.* **2006**, *46*, 2369–2380.

***Table 3.*** Average MCC Value for the Internal Validation and Test Set for Each of 10 Different Definitions of a 10-fold Cross Validation for the Combination of Descriptors[a]

| | RF (*mtry* = 4) | | | | SVM ($\gamma = 0.019$, $C = 0.354$) | | | |
|---|---|---|---|---|---|---|---|---|
| fold definitions | int vals | $\sigma_V$ | test | $\sigma_T$ | int vals | $\sigma_V$ | test | $\sigma_T$ |
| 1 | 0.502 | 0.171 | **0.586** | 0.134 | 0.553 | 0.223 | 0.403 | 0.215 |
| 2 | 0.535 | 0.227 | 0.506 | 0.233 | 0.527 | 0.151 | 0.506 | 0.220 |
| 3 | 0.564 | 0.218 | **0.565** | 0.165 | 0.584 | 0.229 | 0.516 | 0.140 |
| 4 | 0.512 | 0.070 | **0.589** | 0.117 | 0.581 | 0.220 | 0.564 | 0.217 |
| 5 | 0.449 | 0.200 | **0.523** | 0.140 | 0.429 | 0.260 | 0.478 | 0.221 |
| 6 | 0.511 | 0.169 | **0.508** | 0.212 | 0.529 | 0.242 | 0.481 | 0.240 |
| 7 | 0.511 | 0.224 | 0.545 | 0.158 | 0.509 | 0.235 | **0.590** | 0.214 |
| 8 | 0.476 | 0.221 | **0.495** | 0.147 | 0.522 | 0.239 | 0.430 | 0.162 |
| 9 | 0.542 | 0.226 | 0.517 | 0.272 | 0.579 | 0.177 | **0.536** | 0.269 |
| 10 | 0.531 | 0.096 | **0.557** | 0.199 | 0.546 | 0.173 | 0.540 | 0.190 |
| av | 0.513 | 0.182 | 0.539 | 0.178 | 0.536 | 0.215 | 0.505 | 0.209 |

[a] The average reported at the bottom is the averaged value of the columns. $\sigma_V$ represents the standard deviation across the 10 different folds for the internal validation. Similarly, $\sigma_T$ represents the standard deviation of the test set. Highlighted in bold is the highest MCC on the test fold for each fold definition.

and from 0.182 to 0.293 for SVM. For the CFP descriptors, the results of the 10 independent 10-fold cross validations are shown in Table 2. SVM produces the best result overall with an averaged MCC of 0.650 compared to an MCC of 0.619 for Random Forest. Random Forest also has a higher $\sigma_T$ on average at 0.221. The standard deviations for the averaged MCC for each fold definition ($\sigma$(MCC)) are much lower for both SVM and Random Forest, with values 0.031 and 0.033 respectively.

For the combination descriptor set the results are shown in Table 3. Here we see again Random Forest producing a higher averaged MCC value on the test set compared to that of SVM. Again the best *mtry* parameter selected is 4 and Random Forest produces more reliable results with a $\bar{\sigma}_T =$ 0.178. The standard deviation of the average MCC across each fold definition is small for both Random Forest and SVM, with values 0.034 and 0.058.

Overall the most predictive method is an SVM model with $\gamma = 0.011$ and $C = 0.841$ using CFP descriptors.

## Discussion

The use of machine learning algorithms and a more sophisticated descriptor set can lead to improved prediction. While it is hard to distinguish which had the larger effect, it has been shown that using more sophisticated descriptors than simple "dumb" descriptors leads to an increase in predictivity.[30] It has also been shown that both Random Forest and SVM do significantly better than simpler methods such as linear trees.[31,32] Both SVM and Random Forest produce good predictivity. Using a repeated 10-fold cross validation with 10 different definitions allows for a more reliable result to be obtained. This can be seen by the small deviation in results across different fold definitions ($\sigma$(MCC)). Another option could have been to select the folds so that dissimilar compounds appear in each. While this produces a good test for the algorithm, it may not be similar to how the algorithms would be used for a real world problem. Often these techniques will be trained on all possible data available, therefore when a new molecule is tested it will not always be highly dissimilar to those molecules in the training set. When artificially selecting folds it can be the case that the test set has a larger amount of unseen molecules than when used in the real world, and hence an artificially lower MCC value could occur. A stratified cross validation, in which compounds are selected randomly while maintaining their class proportions, can still give the variation in folds necessary to test the algorithm. This can be seen from the large standard deviation across individual folds ($\sigma_T$). These large variations across individual folds can suggest that certain molecules are particularly difficult to predict. A confidence index was derived which can be used as a guide to which molecules were hard to predict. For the CFP data set we calculated the proportion of times the majority prediction was made for each compound over all the runs. This produces a value between 0.5, suggesting that over the runs both classes were predicted equally, and 1.0, for which over the runs only one class was predicted for this compound. A table of the compounds and their respective indexes for both SVM and Random Forest is included in the Supporting Information.

In order to validate that our model was not overfitting, we investigated *y*-scrambling[33] of the data. This was performed on the CFP data set as this produced our best model. The data were initially split into a training set and a test set (60%/40%) using a weighted random sampling. In the training set the class column was permuted randomly, and then the same procedure as before was followed. 50 features were selected using SVMAttributeEval with default parameters, and then SVM with $\gamma = 0.011$ and $C = 0.841$ and Random Forest with *mtry* = 4 were trained on the data. These models were then used to predict the original test set. This was repeated 50 times. The MCC and the fraction of correct predictions, ACC, are shown in Figure 1 for Random Forest and SVM respectively. The model was also trained without the initial scrambling following the same procedure

[30] Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369–1375.

[31] Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

[32] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14.

[33] Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

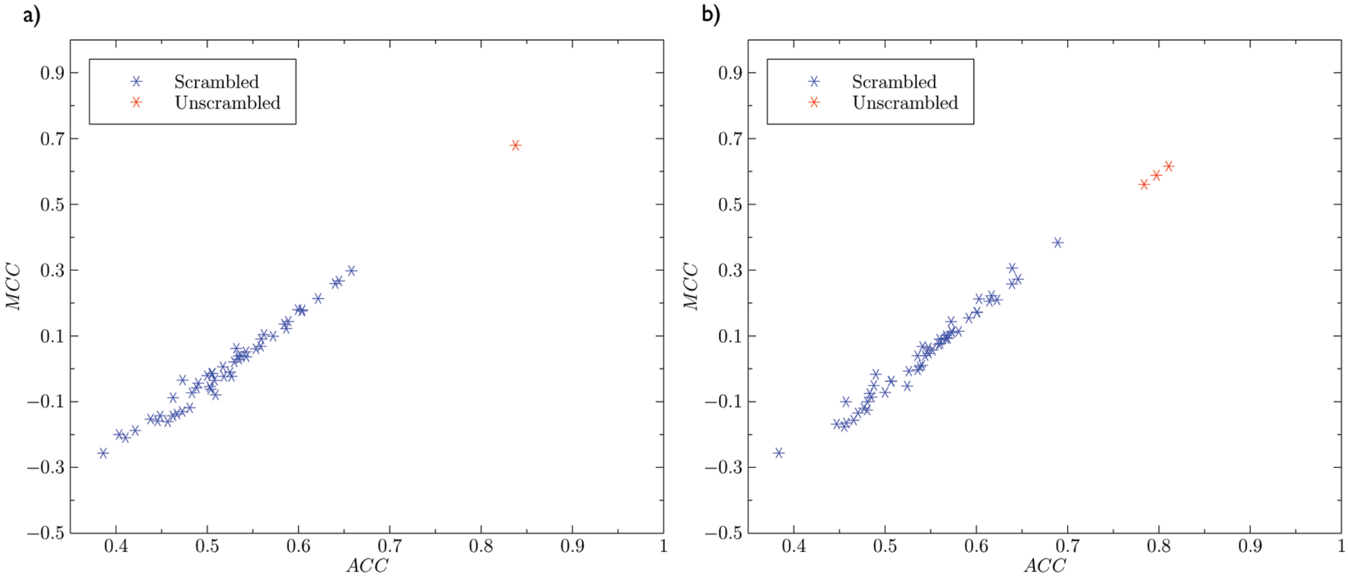

***Figure 1.*** (a) SVM ($\gamma$ = 0.0110, *C* = 0.841) results for *y*-scrambling. The MCC is plotted against the accuracy, ACC (the fraction of correct predictions). Blue stars show the repeated runs of different scrambled data. Red stars show our best model run on this split of the unscrambled data which is repeated 10 times. For SVM this produces the same confusion matrix for all runs as expected. (b) Random Forest (*mtry* = 4) results for *y*-scrambling. Here Random Forest produces 3 distinct confusion matrices for the 10 runs on the unscrambled data.

with 10 repeats, and the results are shown as red stars. It can be seen that our methods do produce both higher MCC and higher ACC, fraction of correct predictions, than those with *y*-scrambling, suggesting that information is gained in the process of learning. The results from the *y*-scrambled models are statistically in line with those expected on the basis of purely random prediction. We also see a good correlation between MCC and ACC, which is as expected for a reasonably balanced data set. The best result for scrambling has an MCC value of 0.384 and an ACC around 0.689. The mean of the scrambled data for both SVM and Random Forest lies around the (0.5,0.0) point as expected, and all points lie within 2 standard deviations. With a larger data set and therefore test set we would expect a reduction in the deviation of the MCC and ACC values for scrambling.

Table 4 shows the top 10 ranked features across all of the runs for the combination of descriptors data set. Out of the top 10, 8 features are selected from CFP, and 2 are selected from the E-Dragon descriptors which are the sophisticated descriptors: 3D-MoRSE—signal 23/unweighted (Mor23u) and the Ghose—Viswanadhan—Wendoloski antihypertensive-like index at 50% (Hypertens-50). For the top 50 features selected for the combination data set, 25 were on average chosen as CFP descriptors and 25 were chosen as E-Dragon. Despite the use of feature selection, as can be seen from the results, using a larger number of descriptors can lead to a less predictive model. Surprisingly, as they have been suggested as important descriptors in previous work,[7,13] p$K_a$ and ClogP do not appear in the top 10 selected features. Looking at the top 50 features for the CFP data set, however, both ClogP (top) and p$K_a$ (third) appear. This could be the main reason why we see a large improvement with CFP, as these descriptors seem important and are picked out more easily with the smaller number of descriptors to choose from.

***Table 4.*** Average Rank of the Features across All Runs for the Combined Data Set[a]

| | feature | average rank |
|---|---|---|
| 1 | 0-Cac;1-C3;1-O.co2;1-O.co2 | 9.86 |
| 2 | 0-N3;1-C3;1-C3;1-C3 | 18.71 |
| 3 | 0-C2;1-C2 | 21.98 |
| 4 | 0-Car;1-Car;1-Car;1-Nar | 23.09 |
| 5 | 0-C3;1-C3 | 23.58 |
| 6 | Mor23u | 28.51 |
| 7 | 0-C3;1-C2;1-C3;1-N3 | 29.35 |
| 8 | 0-O.co2;1-Cac | 29.60 |
| 9 | Hypertens-50 | 29.70 |
| 10 | 0-C2;1-Nam;1-Nam;1-O2 | 30.12 |

[a] The rank is determined from the feature selection performed using SVMAttributeEval on the training folds. Here the top 10 highest ranked features are shown. All features apart from 6 and 9 are represented in circular fingerprint notation. Mor23u relates to 3D-MoRSE—signal 23/unweighted and Hypertens-50 relates to Ghose—Viswanadhan—Wendoloski antihypertensive-like index at 50%. Both are descriptors calculated by E-Dragon.

This also gives a good justification for believing that just using a large number of features, and hoping that feature selection will pick the most descriptive features out, is a bad approach to machine learning.

We have made numerous attempts to repeat Ivanciuc's study[15] of the Pelletier database. Structures were downloaded using the CAS numbers supplied by Pelletier et al.[7] using PubChem (http://pubchem.ncbi.nlm.nih.gov). E-Dragon[20] descriptors were calculated from these structures once any fragments had been removed. We have used the same feature selection method from Weka,[22] SVMAttributeEval with default parameters, to choose the top 50 E-Dragon descriptors, just as Ivanciuc did. We have implemented a workflow which is effectively identical to that with which Dr. Ovidiu

Ivanciuc kindly supplied us, using Weka to build an SVM model with a Gaussian radial basis function kernel. As previously discussed, this involves the use of two parameters $\gamma$ and $C$. We chose the same $\gamma$ parameters as Ivanciuc reports and chose $C = 100$ just as he did (private communication). We randomly split the Pelletier database into 10 folds for cross validation, and repeated this splitting 10 times, so that we have carried out 10 independent 10-fold cross validations on the 117-molecule data set.

This procedure generated an average prediction accuracy of 0.820 and MCC of 0.658. This is a significant difference compared with the accuracy of 0.970 and MCC of 0.944 reported by Ivanciuc, but is in line with what we report for our own models. To ensure that our own results can be reproduced, we supply as Supporting Information our database containing structures in SMILES format and the necessary scripts used to run our models.

As can be seen from the results reported for our models on the smaller 117 molecule data set, an increase in the size of the data set can cause a difference in predictivity. One possible reason for the large deviation across folds is the relatively small number of molecules in each fold (~18 in the full data set and only ~12 in the smaller one) and hence in the test set. A small number of molecules implies that the MCC will vary to a much greater extent when a single molecule's prediction is changed. Therefore despite the ~50% increase in size of our full 185 molecule data set, an individual molecule can still have a large effect on the results. Figure 1 shows that, with a small test set, in this case 74 molecules, there is a greater chance of a random *y*-scrambled model predicting reasonably well. Hence, with a larger data set more reliable measures of performance can be calculated. A further future increase in the size of the available data sets for phospholipidosis would also hopefully cause an increase in the size of chemical space sampled and allow for more general rules to be learned.

## Conclusion

We have used SVM and Random Forest to generate predictive models for phospholipidosis inducing potential. SVM produces the best predictive model using CFP descriptors giving an average MCC of 0.650 in a 10-fold cross validation. Indeed, we obtain universally better results with CFP than either with E-Dragon descriptors or with a combination of the two. The results of the *y*-scrambling tests confirm that our models have actually learned and that their success is not due to chance correlations. A relatively large deviation occurs between individual folds in each set of 10, suggesting that some individual molecules could be hard to predict. However, the deviation is small between the averaged results from the different partitions of the whole data set into folds. This suggests that we have a reliable value for the MCC describing the overall predictivity of our models. We find lower MCC values for the larger 185 molecule data set than for the 117 molecule Pelletier data set. This suggests, for more reliable and robust predictivity, the need for a much larger publicly available database of phospholipidosis inducing potential.

**Acknowledgment.** We thank the EPSRC (Grant EP/F049102/1) and the Scottish Universities Life Sciences Alliance (SULSA) for funding and Unilever for supporting the Centre for Molecular Science Informatics.

**Supporting Information Available:** The database in SMILES format and all scripts used to run our experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

MP100103E