Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays

Rajarshi Guha · Stephan C. Schürer

Received: 4 October 2007 / Accepted: 30 January 2008 / Published online: 19 February 2008

Abstract Computational toxicology is emerging as an encouraging alternative to experimental testing. The Molecular Libraries Screening Center Network (MLSCN) as part of the NIH Molecular Libraries Roadmap has recently started generating large and diverse screening datasets, which are publicly available in PubChem. In this report, we investigate various aspects of developing computational models to predict cell toxicity based on cell proliferation screening data generated in the MLSCN. By capturing feature-based information in those datasets, such predictive models would be useful in evaluating cell-based screening results in general (for example from reporter assays) and could be used as an aid to identify and eliminate potentially undesired compounds. Specifically we present the results of random forest ensemble models developed using different cell proliferation datasets and highlight protocols to take into account their extremely imbalanced nature. Depending on the nature of the datasets and the descriptors employed we were able to achieve percentage correct classification rates between 70% and 85% on the prediction set, though the accuracy rate dropped significantly when the models were applied to in vivo data. In this context we also compare the MLSCN cell proliferation results with animal acute toxicity data to investigate to what extent animal toxicity can be correlated and potentially predicted by proliferation results. Finally, we present a visualization technique that allows one to compare a new dataset to the training set of the models to decide whether the new dataset may be reliably predicted.

Keywords Domain applicability · HTS assay · QSAR · Cell proliferation · Animal toxicity · Jurkat cell line

Introduction

Toxicological effects play an important role in various fields such as drug discovery, environmental chemistry as well as in various industrial processes as a regulatory factor. Traditionally, toxicity studies have been experimental in nature and in most cases have involved animal studies. Such studies can be time-consuming and expensive. As a result, computationally predicting the toxicity of a given molecule has been intensively studied, as a means to avoid animal testing. One of the major problems faced by computational models is the complexity involved in the phenomenon of toxicity. That is, animal toxicity exhibited by a compound can be caused by—often unexpected—interactions with 'off-targets' involved in a variety of cellular pathways and processes; toxicity can also be the result of a series of embolic processes. Traditional animal testing does not always provide us with information regarding these aspects. Though techniques [1, 2] have been devised to predict protein targets for a given structure, they have only been recently applied to toxicology and require experimental validation. Of course, the situation is not as dire as described here. For example, there are certain well-known protein targets that can lead to toxicity, such as the human Ether-a-go-go Related Gene (hERG) [3, 4] or acteylcholinesterase (AChE) [5, 6]. For such scenarios, one can apply a number of methods to decide whether a

R. Guha (✉)
School of Informatics, Indiana University, Bloomington,
IN 47406, USA
e-mail: rguha@indiana.edu

S. C. Schürer
Department of Scientific Computing, The Scripps Research
Institute, Jupiter, FL 33458, USA

compound will be toxic by virtue of interacting with hERG, for example. However, this approach is only suitable when one is interested in predicting the toxicity of a compound with respect to a specific target.

In contrast to animal toxicity, cellular toxicity studies have recently become available as part of the Molecular Libraries Initiative (MLI). Cell-viability studies—as an indication of general cellular toxicity—can be carried out in the form of high-throughput screens; in an established general approach cell proliferation is determined by measuring ATP concentration. The results of various of such studies have been made publicly available in the PubChem bioassay repository. It should be noted however that such generic cell proliferation assays do not address any specific mechanisms or targets involved in toxicity. Although it may be possible to gain some mechanistic insights by analysis of cell proliferation results from different cell lines, the current study attempts to correlate generic cell-proliferation to chemical features. Cell-based studies can be very cost-effective as well as rapid. One can therefore imagine the results of such studies being used as surrogates for whole animal testing, though this would be associated with obvious problems [7]. A much more straight-forward use of data from cell proliferation assays is to predict cellular toxicity. As before, one is hampered by lack of knowledge of specific targets and mechanism, although the number of confounding factors is significantly reduced. Predictive cellular toxicity models would be a useful addition to the MLSCN data to quickly screen out compounds that could be reliably predicted as toxic and therefore be of no further interest or to remove false positives (for example in cell-based reporter assays).

The general problem of identifying whether a molecule will be toxic or not has been studied extensively using a variety of methods. Some studies have focused on specific species such as fathead minnow [8, 9] and T, pyriformis [10–12]. Other studies have focused on specific classes of compounds such as polycyclic aromatic hydrocarbons [13–16] and halocarbons [17–20]. A number of computational approaches have been employed including linear regression models [21, 22], neural networks [23, 24], Kohonen maps [25, 26] and expert systems [27, 28]. As noted one of the fundamental problems is identifying toxic compounds in the absence of knowledge of toxicity mechanisms. Given that a number of known toxic compounds are available one approach is to identify structural features that are correlated to toxicity. This approach was used by Sander et al. [29] to assign compounds to different toxicity classes (such as mutagenic or carcinogenic) based on the occurrence of toxicity indicating substructural fragments. Other approaches are similar in nature, but use a variety of molecular descriptors to build predictive models. In some cases such models are used to predict LD50 values whereas in other cases they are used to perform classification. More recently a number of studies have been carried out which aim to predict toxicity mechanisms and, in some cases, actual protein targets [2]. The approach used in these cases is fundamentally a QSAR. However, so-called ‘target fishing’ methods [1] that were originally devised to identify drug targets of therapeutic interest could profitably be applied to the problem of target identification in toxicity studies.

We had a number of goals for the current study. First, we aim to provide a comparison of the structures used in different HTS cell proliferation assays and animal toxicity studies. Second, we wanted to examine the feasibility of building predictive models using cell proliferation data to predict cellular toxicity. Given the expected growth of such data due to the efforts of the MLI, we believe that there is a need to investigate protocols and best practices for modeling HTS data that is characterized by noise and significant imbalances in the composition of active (i.e. toxic) and inactive (non-toxic) classes. Our approach involves the use of ensembles of random forest, which can alleviate the imbalanced nature of the problem. In addition, we were also interested in identifying structural features that are correlated to toxicity. Third, we wanted to explore the feasibility of utilizing cell proliferation data to predicting animal toxicity. Given the above discussion, the results of such a task would appear to be a foregone conclusion. However, an analysis of toxicity data indicates that a number of molecules are toxic in both cell proliferation and animal studies, suggesting data from cell toxicity studies in some cases can be correlated to animal toxicity. Finally, we were also interested in developing methods to define the domain of applicability of a model, such that one would be able to decide whether a model can be used to provide reliable predictions for a new dataset.

Datasets

Human T-cell (Jurkat) proliferation data generated by the The Scripps Research Institute Molecular Screening Center were extracted from PubChem AID’s 364, 463, 464. These entries include approximately 60,000 data points with primary percent inhibition (measured at 4 μM) and about 800 IC50 data points. For the initial studies we used 775 pIC50 values since we could not evaluate descriptors for some of the molecules. We then selected a cutoff of 5.5, such that molecules with a pIC50 greater than this value were classified as toxic and the remainder as non-toxic.

To investigate the generality of our workflow we also considered datasets from the National Center for Chemical Genomics (NCGC). Cell proliferation QHTS IC50 data points [30] for 1,334 compounds against 13 cell lines were extracted from various PubChem AID’s. Class labels were

Table 1 A summary of the cutoffs and subsequent class distributions for the three datasets considered in the study

Dataset	Subgroup	Cutoff	No. toxic	No. non-toxic
MDL Toxnet	Mouse/IP	4.05	1,823	44,820
Scripps Jurkat		5.50	276	499
NCGC	BJ	3.63	106	1,228
	Jurkat	4.68	59	1,275
	Hek293	4.07	103	1,231
	HepG2	3.57	110	1,224
	MRC5	3.46	101	1,233
	SKNSH	4.10	106	1,228
	N2a	4.39	112	1,222
	NIH3T3	4.53	70	1,264
	HUVECC	3.84	125	1,209
	H4IIE	4.59	62	1,272
	SHSY5Y	4.80	59	1,275
	RenProxTube	4.08	166	1,168
	Mesenchym	3.74	117	1,217

automatically generated by choosing a cutoff equal to two standard deviations above the mean pIC50 for a given cell line and labeling compounds with a pIC50 greater than the cutoff as toxic and those below as non-toxic. We then developed models for each individual cell line. Table 1 summarizes the cutoff's and subsequent class distributions for the datasets used in this study.

The acute animal toxicity data set was extracted and reformatted from the Registry of Toxic Effects of Chemical Substances (RTECS) available through the MDL Toxicity Database (2006.2) [31], which reports toxicity endpoints for various species and routes of administration. From this set we extracted and aggregated LD50 results for mouse and rat and four routes of administration (oral, intravenous, intraperitoneal, subcutaneous) for 103,041 chemical structures (154,019 LD50 data points; compounds do not have reported values for all combinations of species and modes of administration); LD50 values in mg/kg were converted into concentrations using the reported molecular weight. For the purposes of this study, we focused on the mouse and intraperitoneal mode of administration (referred to as MIP) since it had the largest number of observations. We selected a cutoff, such that any molecule whose measured pLD50 was two standard deviations greater than the mean was classified as toxic and the rest as non-toxic.

It should be noted that the compounds tested in the NCGC proliferation assays showed little overlap with the MLSMR collection (at the time these datasets were generated), which is routinely screened in the MLSCN. In addition, since the data from the RTECS database is from a different source, we compared the similarities between the datasets in a pairwise fashion using extended connectivity fingerprints (ECFP_12), summarized in Fig. 1. If we consider the first graph (a) we see that close to 80% of the dataset has an average similarity of 0.2 to the RTECS dataset and 50% of the dataset has an average similarity of 0.35 to the NCGC dataset. There are no compounds in the MLSMR that have a similarity greater than 0.6 to either of the other two databases. If we consider the next plot (b), we

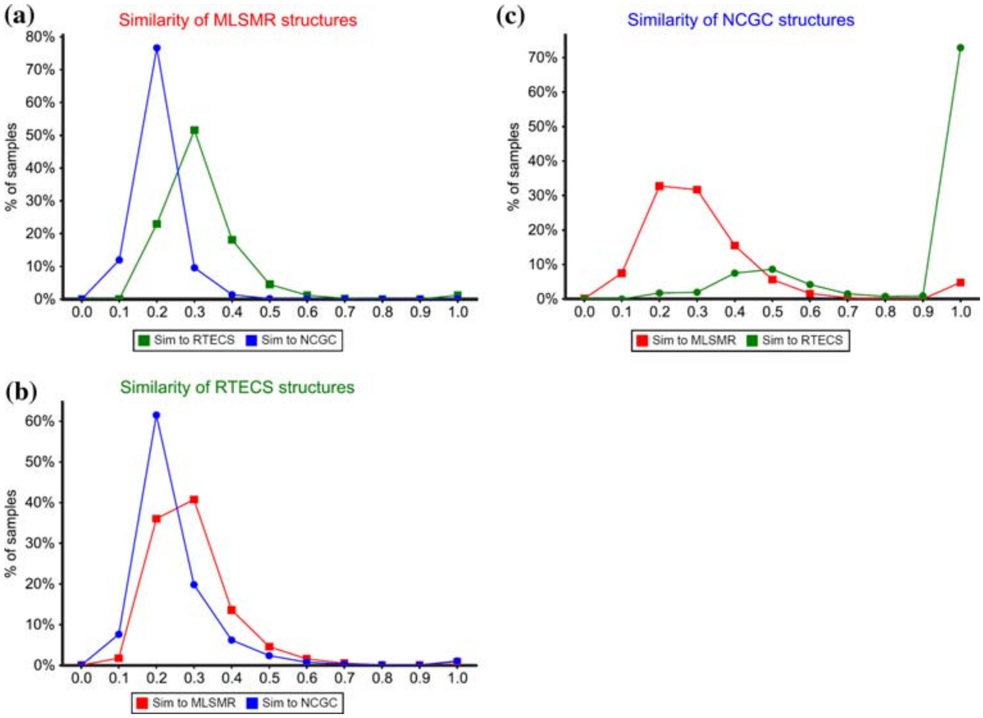


Fig. 1 Pair-wise similarity histograms between the three libraries from the (**a**) MLSMR (59,805 structures screened by Scripps), (**b**) RTECS database (103,040 structures), and (**c**) NCGC (1,334 structures) using ECFP_12 fingerprints and the Tanimoto metric

see a similar situation. The bulk of the RTECS database has very low similarity to either of the other two datasets. As before there are very few compounds that exhibit a similarity higher than 0.6 to the compounds in the MLSMR or NCGC datasets. Finally, if we consider the similarities of the NCGC dataset (Fig. 1c), we see that the bulk of the dataset has a low average similarity to the MLSMR dataset, but does exhibit a very high (greater than 0.9) similarity to the RTECS dataset; in fact the compounds tested by NCGC were specifically selected with respect to known toxicity.

Descriptors

One of the goals of this study was to try to identify toxicity-indicating patterns [29]. As a result, we focused on structural key fingerprints. We used the BCI 1,052-bit structural keys and evaluated these fingerprints for all three of the datasets mentioned above. We also considered the CATS2D [32] descriptors, which can be considered pharmacophoric fingerprints that are based on topological distance constraints rather than 3D Euclidean distance constraints. We employed an in house implementation that utilizes the five pharmacophoric groups, defined by Renner et al. [32] and a maximum path length of 9, thus generating a descriptor vector of length 150 for each molecule. Finally, we also investigated the use of real-valued holistic descriptors, namely, Molconn-Z (Edusoft LC.) However, the results obtained using these descriptors did not lead to an appreciable improvement in the results when compared to the fingerprint or pharmacophoric descriptors and thus we do not describe the results in this paper.

Methodology

Model development

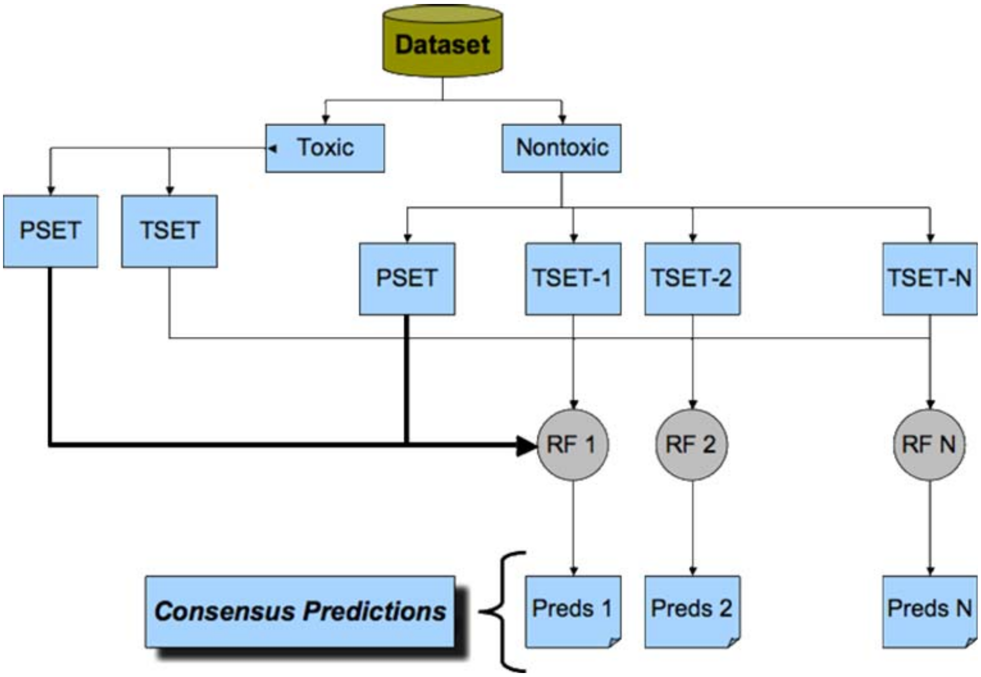
The first step in our modeling protocol was to develop a set of random forest [33] models. Our choice of the random forest as our primary predictive model was guided by a number of attractive features of this model type. First, the random forest method does not require a priori feature selection. Furthermore, the algorithm is not hampered by correlated descriptors. Thus, we were able to use all 1,052 bits of the BCI fingerprints or all 150 CATS2D descriptors (within which the largest correlation coefficient (R^2) was 0.85) for modeling purposes. Second, the random forest algorithm has been mathematically proven to avoid overfitting of the data [34]. Finally, the descriptor importance measure that can be obtained from a random forest model allows us to identify important bits of the fingerprint, which can be subsequently used in other modeling approaches that do require feature selection.

It is clear from Table 1 that we are faced with an imbalanced classification problem. Naively building models for a given dataset without taking into account the imbalanced nature will result in models that exhibit very good performance on the major class (i.e., the non-toxic class) and significantly poorer performance on the minor (toxic) class.

To address the imbalanced nature of the problem we developed a protocol (summarized graphically in Fig. 2) to build ensembles of random forest models. Given a dataset we first consider the toxic compounds and place 20% of them into a fixed prediction set (PSET). The remaining 80% are used for training. We then randomly selected non-toxic compounds for the PSET such that the ratio of the toxic to non-toxic compounds in the PSET corresponded to the ratio in the dataset overall. The underlying goal for this approach is to create a fixed PSET such that the predictive performance over the ensemble can be measured and to provide a realistic predictive test of the models. Next, we used the remaining toxic compounds as a training set (TSET) and an equal number of randomly selected non-toxic compounds are placed in the TSET. A random forest model is then built using the resultant TSET. The process is repeated 30 times so that the non-toxic class is sufficiently sampled, resulting in a 30-member ensemble of random forest models. The ensemble is then tested on the fixed PSET whose statistics are then reported.

We used the implementation of the random forest algorithm available in R 2.5.1 [35]. The algorithm contains a number of user-definable parameters. We investigated the effect of the number of trees and the number of descriptors selected at each split point. We saw no significant improvement in the predictive ability by increasing the number of trees or the number of descriptors and thus went with the default values of 500 and 32 (for the BCI fingerprints) or 12 (for the CATS2D descriptors) respectively. These results are not surprising. As shown by Breiman et al. [34], random forests do not overfit, and so, increasing the number of trees beyond a certain value will not exhibit any changes in the accuracy. In contrast, for other model types, one can increase the number of degrees of freedom (such as number of descriptors in a linear regression model) to get arbitrarily good accuracy. However, such increases in accuracy invariably lead to an overfit model. For the case of the number of descriptors, we note that the descriptors are sampled randomly at a given split point. Given sufficient split points (over all the trees) the entire descriptor pool can be expected to be sampled, using the default value equal to the square root of the total number of descriptors available. Since the algorithm will try and identify the descriptors that are most important for predictive ability, increasing the size of the descriptor samples may lead these “important” descriptors being found faster,

Fig. 2 A graphical depiction of the model development protocol designed to take into account severely imbalanced datasets



but will not necessarily increase the accuracy of the final model.

We also considered the use Naïve Bayes (NB) models as a simpler alternative to random forest models. Though one could use all 1,052 bits to develop the NB model, it is clear that not all bits in the fingerprint play a role in the prediction of toxicity. Instead, one must select a subset of bits from the pool of 1,052 bits. Similar reasoning can be applied to the CATS2D descriptors. This is termed the feature selection problem and many approaches have been described in the literature including genetic algorithms [36–38], stepwise methods [39], simulated annealing [40, 41] and information theoretic methods. We approached this problem in a slightly different manner. We used the random forest models we built to perform feature selection for the NB models; it is described in detail below. After having selected a relevant subset of bits, we followed the procedure described above to develop a 30-member ensemble of Naïve Bayes models available in R.

Feature selection

The random forest algorithm allows one to derive a measure of importance [34] of the input descriptors. Since we had built 30 models for each dataset, we considered the 10 most important descriptors for each of the models for a given data set. Now one might expect that since each model is built on a slightly different subset of the original training sets, the 10 most important descriptors in each of the models will be slightly different. In the worst case, one might 300 unique descriptors to be selected. However, this is not the case as the training data supplied to each model is not significantly different and thus if we consider the *unique* set of 10 most important features across the ensemble, we will generally get fewer than 300 features. The procedure is summarized in Fig. 3. We also conjecture that the size of the unique set of important features correlates to the stability of the ensemble and sampling coverage. Thus, an ensemble that has sampled the non-toxic class well should have a smaller set of unique features than an ensemble that has not sampled it well.

We investigated the validity of the unique set of important features by using them to build an ensemble of Naïve Bayes models. Though the reduced set of features (for both, the BCI fingerprints and CATS2D descriptors) performed better compared to using the full set of features, the Naïve Bayes models did not perform better than the random forest models. Hence, we do not include any further discussion of these models.

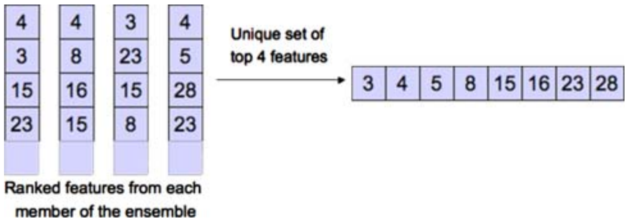

4 3 15 23
4 8 16 15
3 23 15 8
4 5 28 23
Unique set of top 4 features
3 4 5 8 15 16 23 28
Ranked features from each member of the ensemble

Fig. 3 A schematic diagram of the random forest based feature selection procedure. In the above example, we consider an ensemble of four random forest models. The resultant 'important feature' set is the unique set of the top four important features from each model in the ensemble

Results

Random forest models

We first present the results for the random forest ensemble developed using the Scripps cell proliferation pIC50 data. Table 2 presents the confusion matrix for the prediction set, where the results are obtained using a consensus vote of the predictions from the individual models in the ensemble. Overall, the percentage correct classification is 70%. However, the percentage correct for the toxic class is just 55% with an appreciable number of false positives (non-toxic compounds predicted as toxic). Figure 4 highlights the behavior of the ensemble in terms of ROC curves [42] as well as error rate and false negative rate. The area under the curve (AUC [42]) averaged over the ensemble was 0.73 indicating satisfactory classification performance, though it should be kept in mind that this is in reality a result of getting the non-toxic class correct, rather than the toxic class. The random forest algorithm provides *probabilities* of class membership. By default the cutoff is 0.5, so that a prediction whose probability is greater than 0.5 is assigned to the toxic class. Figure 4b highlights the variation in the overall error with the probability cutoff. It is apparent that the optimal error rate is obtained when the cutoff is set to approximately 0.6, rather than 0.5. Given the mediocre performance of the current model and the inability to increase the number of toxic compounds, we investigated the possibility of including more negative (non-toxic) examples to emphasize the difference between the two classes. Based on this assumption we randomly selected 10,000 inactive compounds from the primary screen that were not followed up in secondary screens and placed them in the non-toxic class and developed a 30-member ensemble. However, this approach did not lead to a significantly improved model. In fact, the overall percentage correct consensus classification was 69%, though there is a slight increase in prediction accuracy of the toxic class to 56%. One possible reason for the lack of improvement in this model is due to insufficient sampling of the non-toxic class. Given that the non-toxic class contains 10,000 compounds, a 30-member ensemble would select 8,280 unique non-toxic compounds in the best case and in reality fewer.

Table 2 The confusion matrix for the 30-member random forest ensemble developed using the Scripps Jurkat assay data

		True	
		Non-toxic	Toxic
Calculated	Non-toxic	70	31
	Toxic	17	39

The results are for the fixed prediction set using the consensus vote from the individual predictions

One approach to understanding why the increase in the size of the non-toxic class did not improve the results is to consider the distribution of features. To perform this visualization we created *bit spectra*. These are graphical representations of feature distributions within a dataset, obtained using binary fingerprints. Briefly, for a given dataset we count the number of times a given bit position is set to one. This is repeated for all the bit positions in the fingerprint and the resultant counts are then normalized by the size of the dataset. The normalized counts are then plotted against the bit position. We evaluated bit spectra using BCI fingerprints for the original 775 compound dataset and the 10,775 compound dataset and they are shown in Fig. 5. Visual inspection indicates that there is very little difference in the feature distributions between the two datasets, explaining the nearly identical performance. We numerically quantified the difference between the two spectra using a normalized Manhattan distance and the distance between the two spectra was 0.016. This is certainly quite small; however, at this point, we do not have a measure of how large the distance between two such spectra should be to indicate, in an absolute manner, that two datasets are significantly different. Given the lackluster performance of this model on the toxic class, one might assume that it would have little utility in real-world situations. However, we believe that such models can be indicative of general trends in the dataset. Moreover, one could use such models as broad pre-filters, reducing the size of the non-toxic class, allowing one to apply more intensive analysis to the remaining dataset.

We next consider the results obtained for the NCGC dataset. Recall that this dataset actually consisted of thirteen cell lines and that we developed random forest ensemble models for each cell line. Figure 6 presents a graphical summary of the prediction set performance of the models (using BCI 1,052-bit fingerprints and CATS2D descriptors) over all the cell lines. For both sets of models, it is clear that models for certain cell lines perform significantly better than those for other cell lines. Specifically, the fingerprint-based model for the HUVECC and RenProxTube cell lines exhibits close to 80% correct consensus classification, though the performance on the toxic class is slightly degraded. On the other hand, the model built for the Jurkat cell line exhibits significantly improved predictive performance for the toxic class. It is clear, that overall, the CATS2D based models exhibit a slightly improved predictive performance. However, for certain cell lines, the performance has deteriorated, mainly in predictions of the toxic class (RenProxTube and NIH3T3). The deterioration in performance could be ascribed to the distribution of CATS2D features between non-toxic and toxic classes.

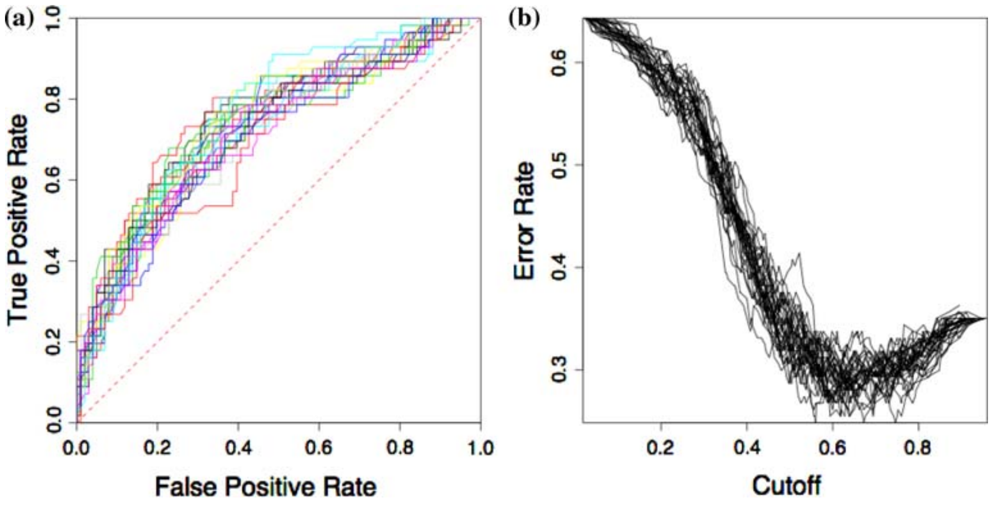


Fig. 4 Summary of the performance of the 30-member random forest ensemble built using the Scripps Jurkat cytotoxicity dataset. **(a)** are the ROC curves for each member of the ensemble. **(b)** is the variation in the overall error rate versus the probability cutoff

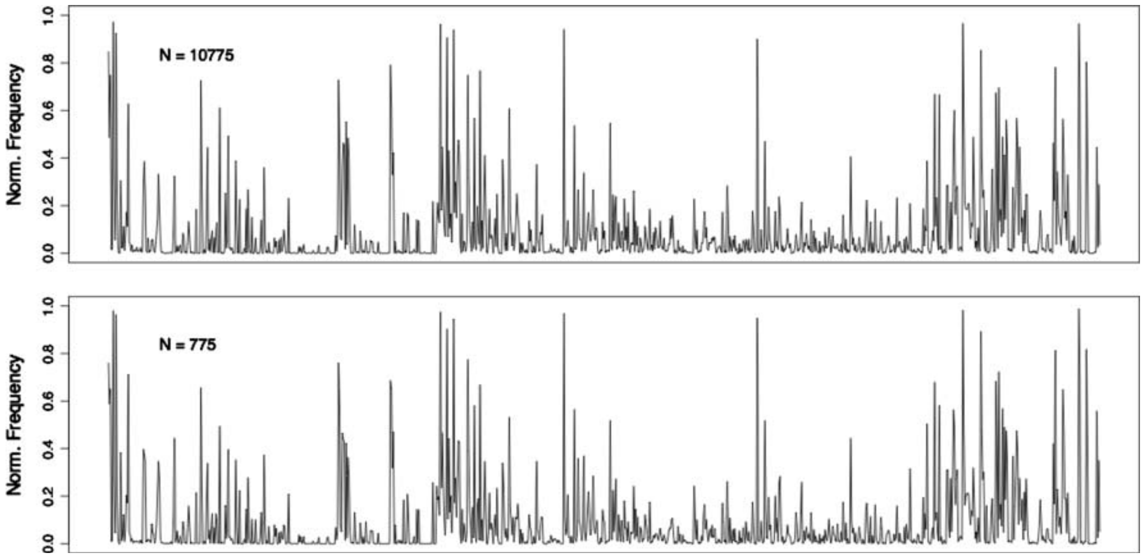


Fig. 5 A plot of the bit spectra for the Scripps cytotoxicity dataset generated using 1,052-bit BCI fingerprints. The top spectrum is for the augmented dataset (N = 10,775) and the lower one for the original dataset (N = 775). The x-axis represents bit position

It is interesting to note that for both the NCGC models described above, there is a very low rank correlation (less than 0.4) between the size of the toxic class and the model performance (either the consensus prediction accuracy overall or the accuracy on the toxic class). This would seem to imply that, the distribution of features in the dataset is more important than simply increasing the size of the toxic class. That is, discrimination between the classes is of higher priority than the absolute size of the classes.

We next attempted to predict the toxicity class of the Scripps 775 compound dataset, using both the fingerprint-based model and the CATS2D-based model developed for the NCGC Jurkat cell line. It should be noted that we chose the NCGC Jurkat model to ensure that the underlying systems were the same, which would not necessarily be the case if we used any other NCGC cell line model to predict the Scripps dataset. Also, the toxic class labels for the Scripps dataset were obtained using arbitrary cutoff's and thus do not match the class labels for the NCGC Jurkat dataset. To overcome this problem we relabeled the Scripps dataset using the cutoffs obtained for the NCGC Jurkat cell line data (see Table 1). The results of the prediction are shown in Table 3 where we see that the overall percentage correct classification is 68% and 76% on the toxic class. From Table 3 it is interesting to note the high false positive rate. From the point of view of feature distributions between the two datasets, we see that the toxic class of the NCGC Jurkat dataset is quite similar to the toxic class of

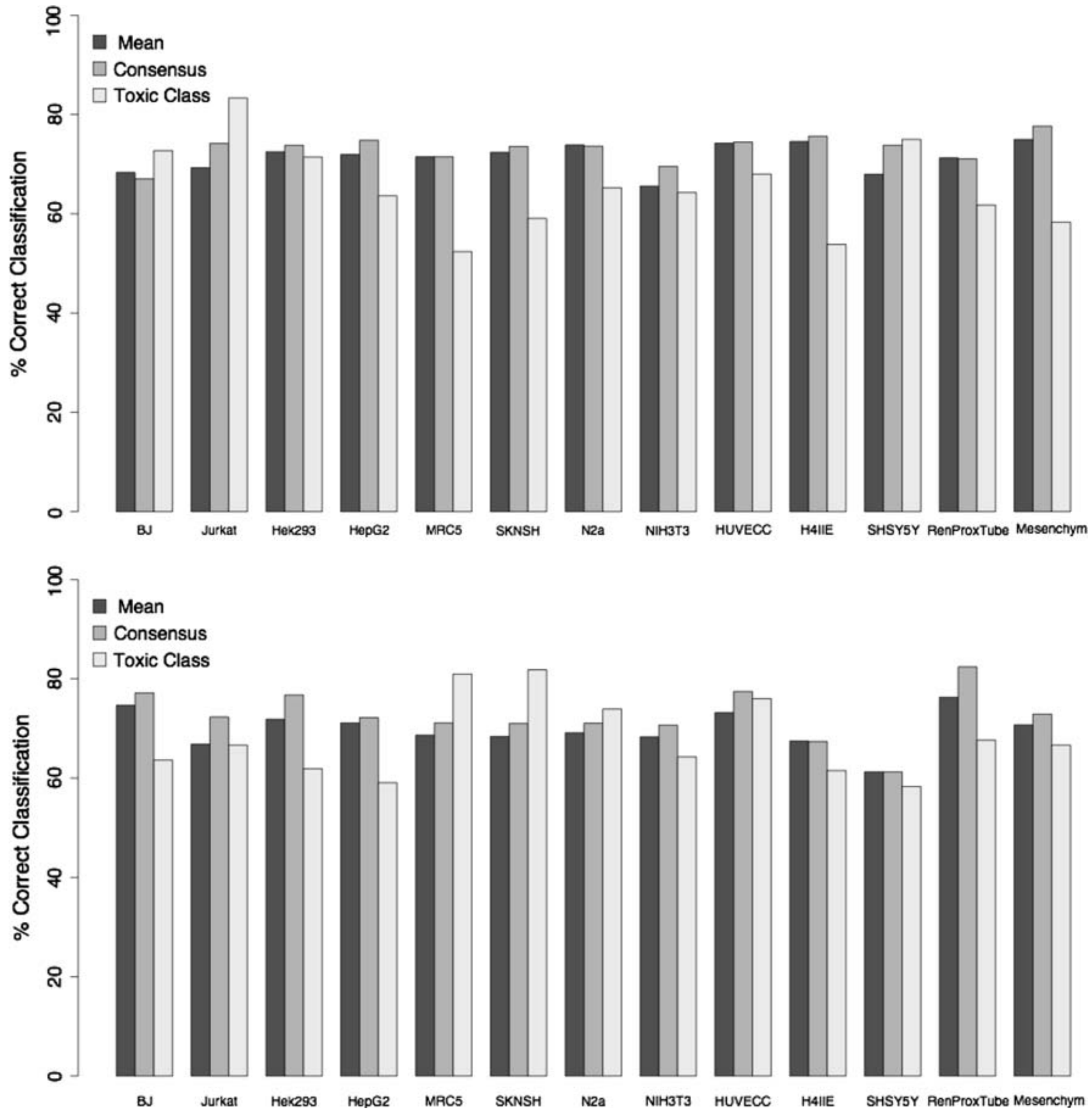


Fig. 6 A summary of the performance of the random forest ensemble models, for the thirteen cell lines in the NCGC dataset. The upper plot represents results for models developed using BCI 1,052-bit fingerprints and the lower plot represents results for models developed using CATS2D descriptors. In both cases, all models were 30-member ensembles

Table 3 The confusion matrix obtained for the prediction of the Scripps datasets using the 30-member random forest ensemble developed using CATS2D descriptors on the NCGC Jurkat cell line data

		True	
		Non-toxic	Toxic
Calculated	Non-toxic	35	152
	Toxic	100	488

the (relabeled) Scripps dataset. At the same time, the non-toxic class of the Scripps Jurkat dataset is also very similar to the toxic class of the NCGC Jurkat dataset. The bit spectra for the two classes of the NCGC Jurkat and Scripps datasets are shown in Fig. 7.

Cell viability and animal toxicity

One of our goals in this study was to explore to what extent cell-proliferation data such as generated by the MLSCN, can be compared and correlated to actual animal toxicity data. To directly compare the available datasets we identified all structures, which are part of at least two of the three datasets (RTECS, NCGC, MLSMR). Using Pipeline Pilot [43] for all records we standardized structural

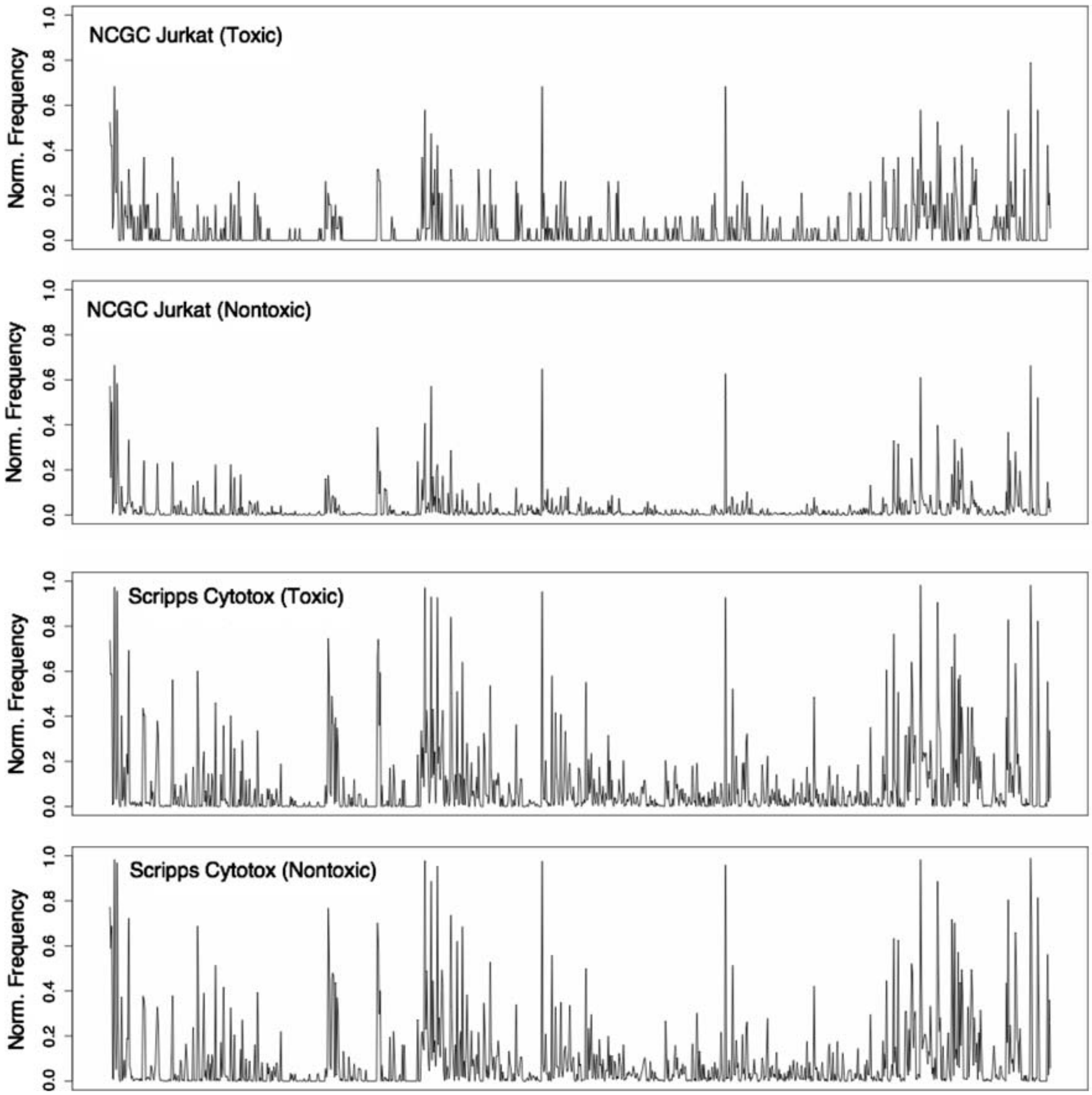


Fig. 7 A class-wise comparison of the bit spectra for the NCGC Jurkat dataset and the Scripps dataset, highlighting the fact that the both classes of the Scripps dataset are more similar to the NCGC toxic class than the NCGC non-toxic class

representations, stereo centers, charges, geometric isomers; we removed salts (using an in-house addend library of 168 structures), generated canonical tautomers and removed redundant fragments (the RTECS structures are stochiometric representations and therefore often include multiple occurrences of individual fragments). We identified about 1,700 exact structures that occur in more than one data set and aggregated the experimental result sets accordingly. As expected the global correlations between any of the 14 cell-based data sets against any of the eight animal data sets is rather low, although not random with the lowest *P*-value of 5×10^{36} between N2a cell line (derived from mouse neuroblastoma; AID 540) and the MIP dataset followed by the Jurkat cell line and the MIP dataset (*P*-value 6×10^{32}). As an example, Fig. 8 shows the Jurkat cell proliferation pIC50 data against the MIP pLD50 results; the Jurkat cell line is among the most sensitive ones with respect to number of actives and also was screened in both libraries and therefore represents the largest cell dataset; for this representation we summarized the two Jurkat datasets (NCGC and Scripps) with few overlapping structures of comparable results that were averaged. Compounds that were not active at the lowest screening concentration in ether of the two data sets or (in case of the Scripps MLSMR data) identified as inactive in a primary screen are shown as inactive/not tested. Various known bioactive and/or toxic compounds are identified in the graph with the corresponding structures and PubChem SIDs provided in Table 4. Clearly there are cases where cell toxicity strongly relates to acute animal toxicity along with structures that

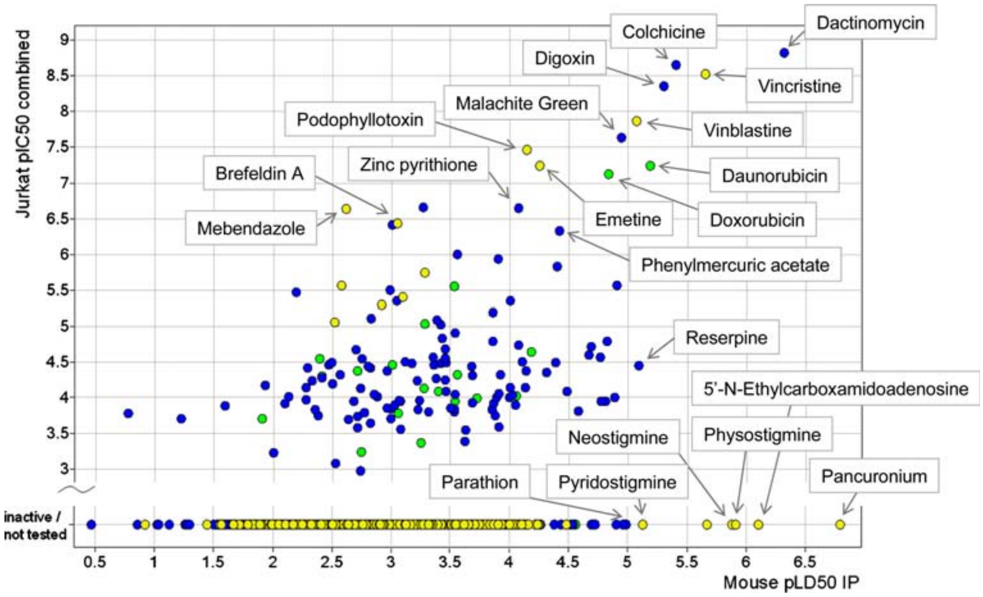


Fig. 8 Scatter plot of cell-based Jurkat pIC50 (moles/l) cell proliferation vs. mouse IP acute toxicity pLD50 (moles/kg) results; yellow—MLSMR compounds screened at Scripps, blue—structures screened by NCGC, green—screened by both centers (average value given); data points labeled inactive/not tested were not active at the lowest screening concentration or not active in the primary assay and therefore not screened in concentration response

are toxic in animals, but not in cells and vice versa. Among the most active compounds in cells and animals is dactinomycin a cyclic peptidic antineoplastic that inhibits protein synthesis by impairing mRNA synthesis by DNA binding. Colchicine is used for treatment of gouty arthritis and Mediterranean fever; its toxic effect is related to tubuline depolymerization. Digoxin is a cardiotonic glycoside used to control ventricular rate in atrial fibrillation with a small margin between toxic and therapeutic doses; it acts on the Na+/K+ ATPase receptor [44]. Vincristine and vinblastine are anti-tumor microtubule modulators. The anthracycline aminoglycosides daunorubicin and doxorubicin are DNA intercalating antineoplastics. Emetine inhibits protein synthesis and can cause serious cardiac, hepatic, or renal damage and violent diarrhea and vomiting. Podophyllotoxin is a potent spindle poison (tubuline modulator). Examples of compounds less active in animals, but still significantly toxic in cells include Zinc pyrithione, a keratolytic agent. Brefeldin A inhibits protein transport to the Golgi. Mebendazole is a glucose uptake inhibitor and microtubule inhibitor. The lower animal toxicity of these compounds may be related to metabolic instability of the lactone, methyl carbamate and perhaps the ketone. The cell-inactive but animal toxic pancuronium is a non-depolarizing muscle relaxant inhibiting the nicotinic acetylcholine receptor. Neostigmine [45], pyridostigmine [46] and physostigmine are reversible inhibitors of cholinesterase; physostigmine can penetrate the blood–brain barrier.

The examples in Fig. 8 illustrate that the relationship of cell toxicity and animal toxicity can be understood in many cases by mechanism of action and perhaps metabolic properties of the chemical entities. With appropriate categorization of target class and physiochemical properties—which in many cases can be related to chemical structures—it may be possible to develop models while taking into consideration specific categories of mechanism and other compound properties.

Given these considerations and the above models derived from cell proliferation data, we attempted to see if we could use those models to predict in vivo data. We thus predicted the toxicity classes for the MIP subset from the MDL RTECS dataset. Table 5 summarizes the results using the Scripps model. Overall, the percentage correct classification is 74% but this is clearly due to good performance on the non-toxic class. If we consider just the toxic class, we get an accuracy of just 41%. It is clear that the compounds correctly predicted as toxic are swamped by the large number of false positives. Given the data in Fig. 8 this is an expected result. Probably the most important reason is that our cell proliferation models—that are derived using relatively simple descriptors and no biological or structural categorization—simply do not sufficiently characterize phenomena such as metabolism and absorption or structural features that are related to mechanisms of action (for example tubuline modulation or cholinergics). Another, technical reason for the poor predictive performance is the issue of the probability cutoff discussed in Sect. 4.1. The results in Table 5 were obtained using the default cutoff of 0.5. However, if we refer to Fig. 4 and use the suggested optimal cutoff of 0.6, we observed that overall accuracy was now 93%. But this value is purely due to very good performance on the non-toxic class (96%) whereas the performance on the toxic class become significantly degraded (7%). At the same time, the false positive rate was reduced to 4%. It is clear that though one can investigate the use of different probability cut off's, their effect is to improve overall accuracy and not necessarily accuracy with respect to the class of interest (the toxic class in this case).

Table 4 Structures and Pubchem SIDs of the highlighted compounds in Fig. 12

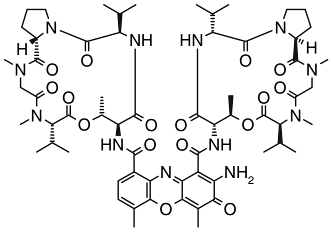
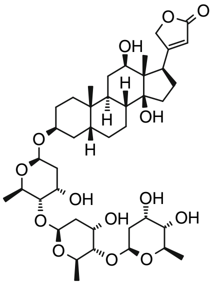
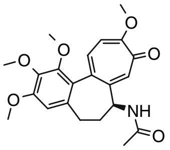
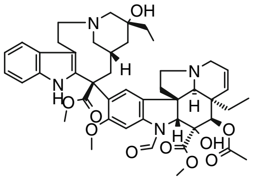
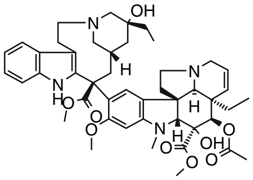
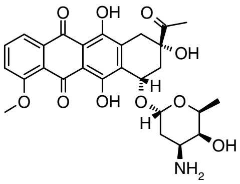
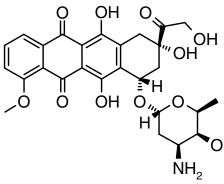
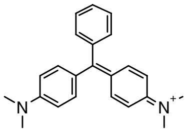
Name	Structure	SIDs
Dactinomycin		7851523
Digoxin		7854036
Colchicine		11076280
Vincristine		855866
Vinblastine		855758
Daunorubicin		11076302 855543
Doxorubicin		11076176 855944
Malachite Green		11076432

Table 4 continued

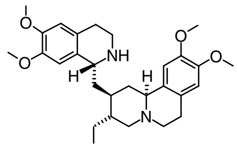
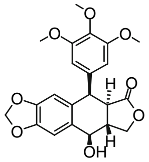
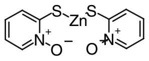
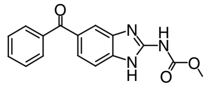
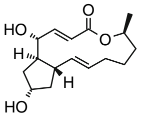
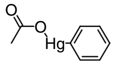
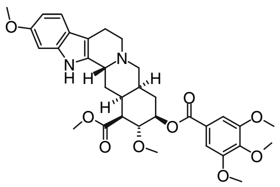
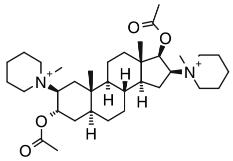
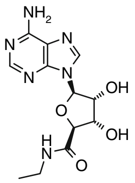
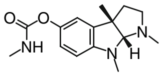
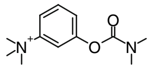
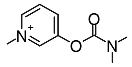
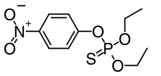
Name	Structure	SIDs
Emetine		855836
Podophyllotoxin		855658
Zinc pyrithione		11076713
Mebendazole		855610
Brefeldin A		855810
Phenylmercuric acetate		7852642
Reserpine		7852729
Pancuronium (Vecuronium Bromide)		855826
5’-N-Ethylcarboxamido adenosine		7975554
Physostigmine		855875
Neostigmine		855599 855838
Pyridostigmine		855974
Parathion		11076576

Table 5 The confusion matrix obtained for the MIP subset of the MDL ToxNet database using the Scripps model (built using BCI 1,052-bit fingerprints)

		True	
		Non-toxic	Toxic
Calculated	Non-toxic	33,390	1,066
	Toxic	11,430	757

Table 6 The confusion matrix obtained for the MIP subset of the MDL ToxNet database using the NCGC Jurkat model (built using CATS2D descriptors)

		True	
		Non-toxic	Toxic
Calculated	Non-toxic	12,634	267
	Toxic	32,605	1,137

We then used the NCGC Jurkat model developed using CATS2D descriptors to predict the animal data and the results are summarized in Table 6. We see that in this case, the overall percentage correct is 29%, but this is influenced by the large number of false positives. If we consider just the toxic class, we see that we achieved an accuracy of 80%. Though this is an impressive level of accuracy, it must be noted that in practice it is probably not very useful since the true toxic compounds are swamped by the large number of false positives.

In fact, this result is quite surprising, since in general, a classifier will tend to assign a new observation to the major class by default. That is, one would expect that the non-toxic class would be very well predicted. Instead, we observe the opposite case. We can gain some insight into the behavior of the NCGC Jurkat model with respect to the animal data by considering the feature distributions of the animal dataset and the two classes of the NCGC Jurkat dataset, as shown in Fig. 9. From the bit spectra we see that, overall, the MIP dataset is much more similar to the toxic class of the NCGC Jurkat dataset than the non-toxic class. Given the similarity in features, it is thus not surprising that the model will tend to predict the MIP dataset as toxic. It should be noted that the bit spectra were derived using the BCI fingerprints, whereas the model we considered here was built using the CATS2D descriptors. We also considered the fingerprint-based model and applied it to the animal dataset and we observed similar behavior. Thus, though the bit spectra do not correspond directly to the

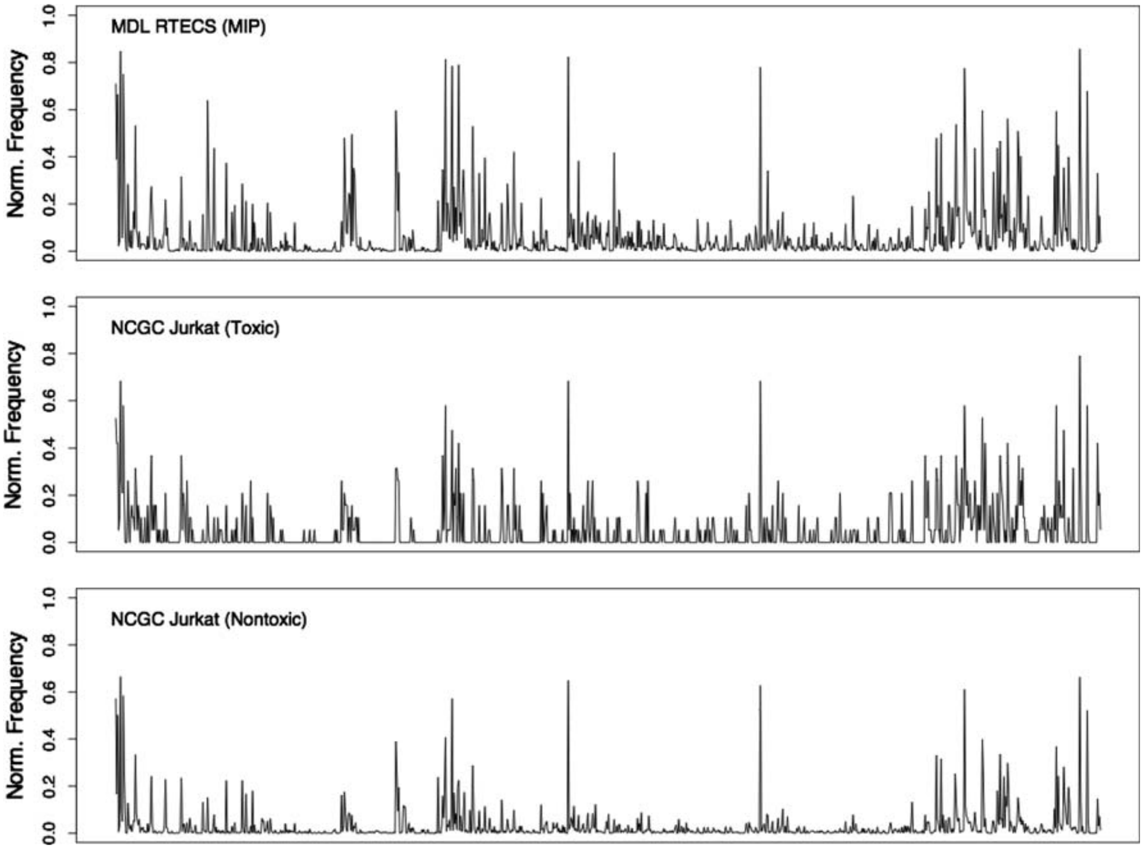


Fig. 9 A comparison of the MIP subset of the MDL RTECS database (top) with the toxic and non-toxic classes of the NCGC Jurkat dataset (middle and bottom respectively)

CATS2D model, we believe that the underlying comparison of the datasets is still valid.

We can do a similar comparison between the Scripps dataset and the MIP dataset and this is shown in Fig. 10. The most distinct feature of the bit spectra is that there is very little difference between the plots for the toxic and non-toxic classes of the Scripps dataset. If we consider the bit spectra of the MIP dataset, we thus see that it is quite similar to both the classes of the Scripps dataset. As a result, a classifier built on these features, will tend to assign the class label of the major class (in this case, the non-toxic class) by default. Indeed, this is what the confusion matrix in Table 5 indicates. Table 6

Model applicability & toxicity indicating features

From the above discussion, it is clear that the composition of the dataset in terms of structural features must be taken into account when predicting new datasets with a previously trained model. The bit spectrum plots represent one approach to comparing datasets in bulk and judging whether a model can be used on a new dataset. Indeed, as described above, such plots can be used to explain model performance as well.

However, given that our descriptors encode structural feature directly it makes sense to investigate the correlation between “important” features and toxicity. One of the simplest approaches [47] to the problem of model applicability has been to measure the applicability of a model to new compounds in terms of the similarity between the new compounds and the training set used to build the model. Our approach is similar in concept but rather than directly consider the complete structural similarity between a new compound and the training set, we consider the similarity in terms of the most important bits (which translates to structural features).

In the interest of space, we only discuss the NCGC Jurkat dataset. Before considering the issue of model applicability, we investigated the nature of the structural features that were deemed important. We first considered the BCI fingerprints, by asking for the unique set of the 10 most important bits. The size of the unique set was 53 bits, which corresponded to 72 actual structural features (represented in the form of SMARTS). We noted that the toxic compounds exhibited a larger number of these 72 features

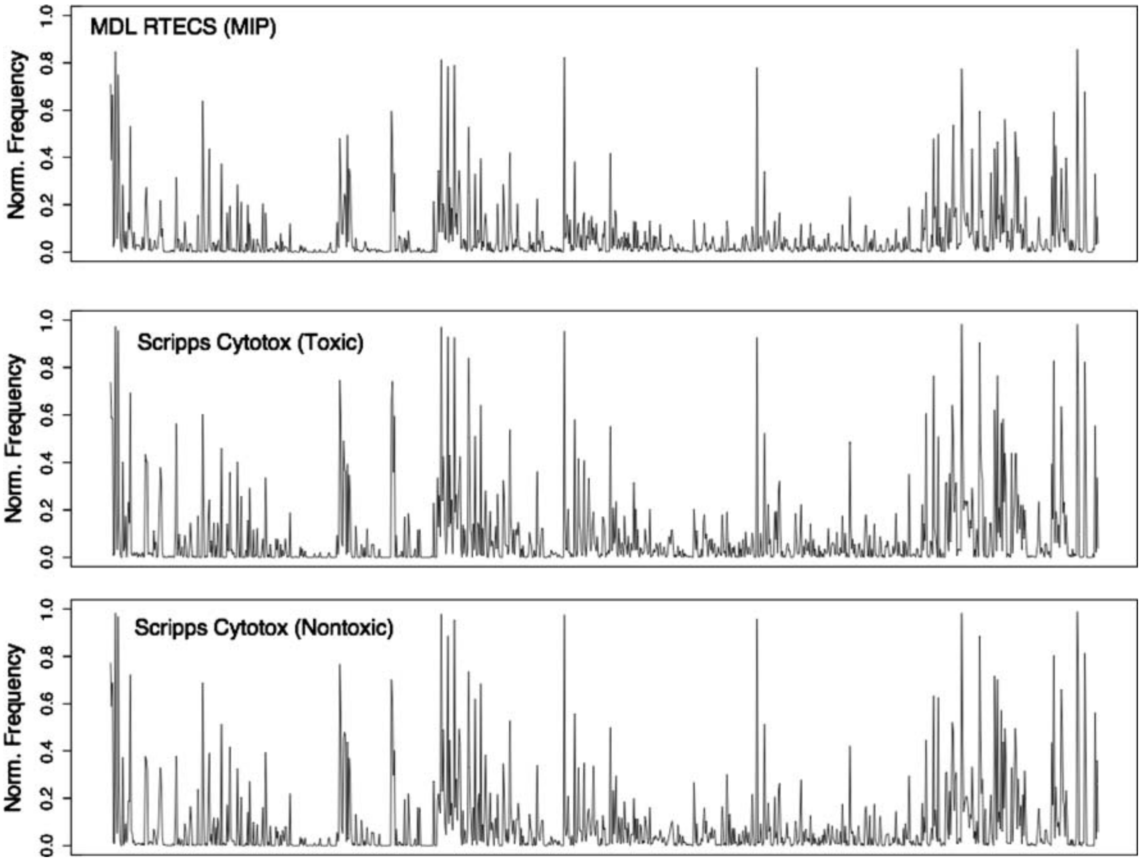
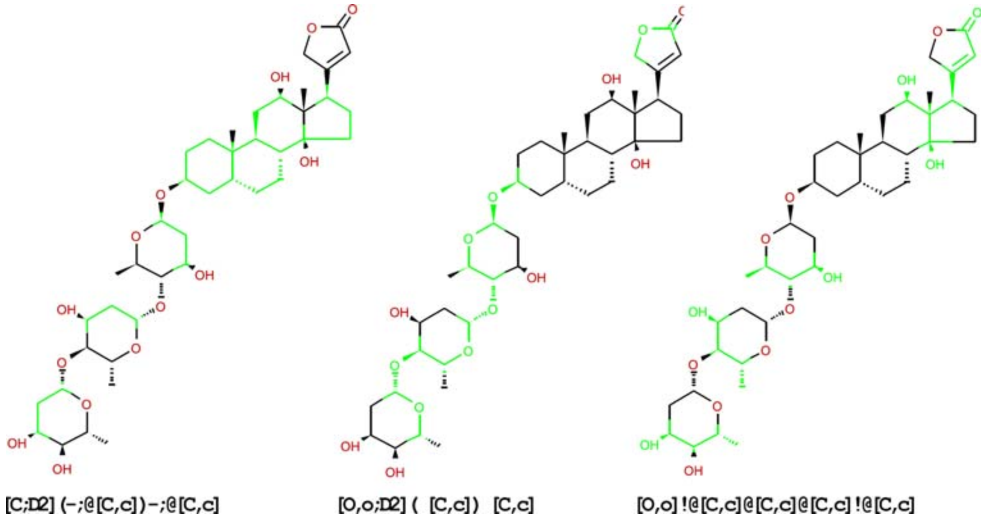


Fig. 10 A comparison of the MIP subset of the MDL RTECS database (top) with the toxic and non-toxic classes of the Scripps dataset (middle and bottom respectively)

on average, compared to the non-toxic compounds. The main problem that we faced was that many of the features were quite broad in nature. As an example, Fig. 11 displays digoxin with three of the 72 important structural features highlighted in green. If we then consider the CATS2D descriptors, but this time specifying the top five most important features over the ensemble we get a unique set of 39 pharmacophore point pairs (PPP). It was observed that the most frequent of these PPP was the A–A pair (where A represents H-bond acceptors), with topological paths of lengths ranging from 0 to 9. On the other hand, only 6 of the 39 features involve a positive-charge pharmacophore (in combination with other pharmacophores) and six involve a negative-charge pharmacophore. Figure 12 shows an example from the NCGC dataset that was predicted as active by the NCGC Jurkat model developed using the CATS2D descriptors. Consistent with the general trends observed for the individual CATS2D descriptors, the most prevalent pharmacophoric groups are the hydrogen bond donor and acceptor groups, with the next most frequent group being the lipophilic groups (not shown). As with the BCI fingerprints, the features identified are not significantly remarkable. However, the frequency of occurrence of the individual features appears to indicate whether a molecule will be predicted as toxic or non-toxic, though this correlation is not necessarily distinct.

Next, we considered how the occurrence of structural features compared between the NCGC Jurkat dataset and the MIP subset of the MDL RTECS dataset. For this comparison, we considered the top hundred important features, which led to a unique set of size 316. It was determined that of the 316 features regarded as important, 130 were in common with the structures from the animal dataset. However even though these features were, on average, more frequent in the NCGC toxic class than the non-toxic class, average number over the entire NCGC Jurkat dataset was just 18.8. Given this low number it would indicate, that in terms of important structural features, the animal dataset might not be reliably predicted by the models built on the NCGC Jurkat dataset.

Fig. 11 An example of some of the important structural features of Digoxin identified by the random forest ensemble for the NCGC Jurkat dataset using the 1,052-bit BCI fingerprints



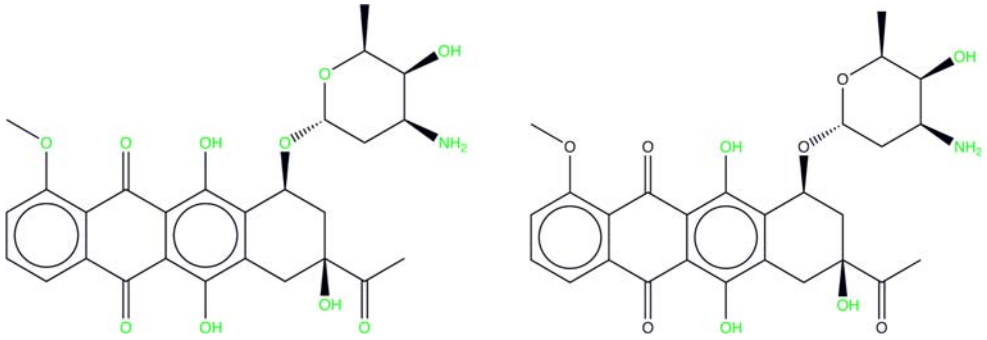


Fig. 12 One of the active molecules (Daunorubicin) from the NCGC dataset that was predicted toxic by the NCGC Jurkat model, using the CATS2D descriptors. The groups highlighted in green represent the topological pharmacophores. The structure on the left has the CATS2D hydrogen bond acceptor groups highlighted and on the right the donor groups have been highlighted

Conclusions

Though data from cell proliferation studies has become more available, developing models based on these datasets can be challenging. It is clear from the discussion above, that such datasets are highly imbalanced and any modeling protocol must take the class distributions into account. Our approach has been to use ensembles of models, where each model sees an equal distribution of the two toxic and non-toxic classes. Such an approach has been shown to achieve relatively good predictive performance. On the other hand, accuracy rates range from 56% to 80%. Our results indicate that simply assuming that a small toxic class will lead to poor models, is a hasty conclusion, as a number of the NCGC models do exhibit good predictive performance. The use of bit spectra indicates that, in terms of binary fingerprints, the Scripps dataset does not exhibit significant differences between the toxic and non-toxic class. Thus, pure structural descriptors (be they binary fingerprints or topological pharmacophores) do not appear to allow a consistent discrimination between toxic and non-toxic classes. In this context we also investigated the use of Molconn-Z descriptors, but they did not lead to a significant improvement in the performance.

Modeling toxicity is further complicated because of the multiple mechanisms and biological targets by which a compound may inhibit cell proliferation. The study shows that simple structural or pharmacophoric descriptors cannot always describe the complexity of this problem accurately enough. This is particularly true when comparing cell-based and animal toxicity. Given that no target information was used to build the models this is an expected, yet important result. More focused [48] models seem appropriate, but require biological (target-based) and structural categorization. A growing biological categorization of the MLSMR library is being generated by the MLSCN and we expect it to become a valuable resource to approach complex problems like toxicity prediction. We are also currently investigating the feasibility of computational target identification using techniques such as PASS [2] and target fishing [1]. Though such approaches are probabilistic in nature, they would allow us to consider a significantly smaller set of possible toxicity mechanisms, which could then be followed up by alternative computational methods and experimental verification.

The above observations indicate that although structural descriptors may be useful in some cases, more broadly applicable models will probably need to be focused on specific mechanisms and will also require mechanistic descriptors. Such descriptors would include a variety of physicochemical types (such as cLogP, polar surface area, etc.) as well as reactivity-based descriptors that may facilitate categorization related to metabolism. In addition to traditional molecular descriptors of structure and activity, one should also take into account data derived from proteomics and gene expression (microarray) experiments, to obtain detailed understanding of the possible interactions that a toxicant may have with a biological system.

Our investigation of the correlation between cell proliferation data and animal acute toxicity further emphasizes that categorization of structures based on biological target and mechanism of action and also based on chemical reactivity and metabolism may be a useful approach to model such complex system more accurately. Not surprisingly cell-based models did not exhibit very good global performance on animal data. Our analysis however has shown that there are a number of instances where compounds are toxic in both a cell proliferation assay as well in an animal system. This would seem to indicate that one could use cell proliferation data to predict animal toxicity, but for restricted cases. One possible approach to developing models based on cellular data for the prediction of animal data would be to focus on cellular datasets that are restricted to specific mechanisms or targets. In all these cases, the key decision is whether to use a model to predict a set of animal data in the first place. That is, does the applicability domain of the model encompass the animal data that is to be predicted. There has been much work on the issue of domain applicability [49]. We have presented a simple approach that attempts to provide a graphical summary in form of feature bit spectra of the match between the chemical spaces of the training data and the data to be predicted. We also investigated whether counts of the important features could be used as a measure of domain applicability and our results appear to indicate this is indeed possible. Future work will involve providing a more rigorous quantitative measure of applicability using both bit spectra as well as counts of important bits.

Acknowledgements RG would like to acknowledge funding from NIH Grant No. P20 HG003894-01. SCS acknowledges the support by the National Institutes of Health Molecular Library Screening Center Network (Grant No U54 MH074404-01, Prof. Hugh Rosen, Principle Investigator).

References

1. Nidhi GM, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46(3):1124–1133
2. Poroikov V, Filimonov D, Lagunin A, Gloriozova T, Zakharov A (2007) PASS: identification of probable targets and mechanisms of toxicity. SAR QSAR Environ Res 18:101–110
3. Paakkari I (2002) Cardiotoxicity of new antihistamines and cisapride. Toxicol Lett 127(1–3):279–284
4. Vandenberg JI, Walker BD, Campbell TJ (2001) Herg K+ channels: friend and foe. Trends Pharmacol Sci 22(5):240–246

5. Maxwell DM, Brecht KM, Koplovitz I, Sweeney RE (2006) Acetylcholinesterase inhibition: does it explain the toxicity of organophosphorus compounds? Arch Toxicol 80(11):756–760
6. Taylor P, Kovarik Z, Reiner E, Radic Z (2007) Acetylcholinesterase: converting a vulnerable target to a template for antidotes and detection of inhibitor exposure. Toxicology 233(1–3):70–78
7. Clark RD, Wolohan PRN, Hodgkin EE, Kelly JH, Sussman NL (2004) Modelling in vitro hepatotoxicity using molecular interaction fields and SIMCA J Mol Graph Model 22(6):487–497
8. Hodges G, Roberts DW, Marshall SJ, Dearden JC (2006) Defining the toxic mode of action of esther sulphonates using the joint toxicity of mixtures. Chemosphere 64(1):17–25
9. Ankley GT, Villeneuve DL (2006) The fathead minnow in aquatic toxicology: past, present and future. Aquat Toxicol 78(1):91–102
10. Lagunin AA, Zakharov AV, Filimonov DA, Poroikov VV (2007) A new approach to QSAR modelling of acute toxicity. Sar QSAR Environ Res 18(3–4):285–298
11. Pasha FA, Srivastava HK, Srivastava A, Singh PP (2007) QSTR study of small organic molecules against Tetrahymena pyriformis. QSAR Comb Sci 26(1):69–84
12. Yan XF, Xiao HM (2007) QSAR study of nitrobenzenes' toxicity to tetrahymena pyriformis using semi-empirical quantum chemical methods. Chin J Struct Chem 26(1):7–14
13. Park SY, Lee SM, Ye SK, Yoon SH, Chung MH, Choi J (2006) Benzo[a]pyrene-induced DNA damage and p53 modulation in human hepatoma HepG2 cells for the identification of potential biomarkers for PAH monitoring and risk assessment. Toxicol Lett 167(1):27–33
14. Roos PH, Tschirbs S, Pfeifer F, Welge P, Hack A, Wilhelm M, Bolt HM (2004) Risk potentials for humans of original and remediated PAH-contaminated soils: application of biomarkers of effect. Toxicology 205(3):181–194
15. Niu J, Yu G (2004) Molecular structural characteristics governing biocatalytic chlorination of PAHs by chloroperoxidase from Caldariomyces fumago. SAR QSAR Environ Res 15(3):159–167
16. Perugini M, Visciano P, Giammarino A, Manera M, Di Nardo W, Amorena M (2007) Polycyclic aromatic hydrocarbons in marine organisms from the Adriatic Sea, Italy. Chemosphere 66(10): 1904–1910
17. Bohonowych JE, Denison MS (2007) Persistent binding of ligands to the aryl hydrocarbon receptor. Toxicol Sci 98(1):99–109
18. Chroust K, Pavlova M, Prokop Z, Mendel J, Bozkova K, Kubat Z, Zajickova V, Damborsky J (2007) Quantitative structure-activity relationships for toxicity and genotoxicity of halogenated aliphatic compounds: wing spot test of *Drosophila melanogaster*. Chemosphere 67(1):152–159
19. Muellner MG, Wagner ED, McCalla K, Richardson SD, Woo YT, Plewa MJ (2007) Haloacetonitriles vs. regulated haloacetic acids: are nitrogen-containing DBPs more toxic? Environ Sci Technol 41(2):645–651
20. Lu GH, Wang C, Li YM (2006) QSARS for acute toxicity of halogenated benzenes to bacteria in natural waters. Biomed Environ Sci 19(6):457–460
21. Liu HX, Papa E, Gramatica P (2006) QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem Res Toxicol 19(11):1540–1548
22. Afantitis A, Melagraki G, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2006) A novel QSAR model for predicting induction of apoptosis by 4-aryl-4H-chromenes. Bioorg Med Chem 14(19):6686–6694
23. Mosier PD, Jurs PC (2002) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 42(6):1460–1470
24. Kaiser KLE, Niculescu SP, Schultz TW (2002) Probabilistic neural network modeling of the toxicity of chemicals to Tetrahymena pyriformis with molecular fragment descriptors. SAR QAR Environ Res 13(1):57–67
25. Roncaglioni A, Novic M, Vracko M, Benfenati E (2004) Classification of potential endocrine disrupters on the basis of molecular structure using a nonlinear modeling method. J Chem Inf Comput Sci 44(2):300–309
26. Mazzatorta P, Vracko M, Jezierska A, Benfenati E (2003) Modeling toxicity by using supervised Kohonen neural networks. J Chem Inf Comput Sci 43(2):485–492
27. Crettaz P, Benigni R (2005) Prediction of the rodent carcinogenicity of 60 pesticides by the DEREKfW expert system. J Chem Inf Model 45(6):1864–1873
28. Veith GD (2004) On the nature, evolution and future of quantitative structure-activity relationships (QSAR) in toxicology. SAR QSAR Environ Res 15(5–6):323–330
29. von Korff M, Sander T (2006) Toxicity-indicating structural patterns. J Chem Inf Model 46(2):536–544
30. Xia M, Huang R, Witt KL, Southall N, Fostel J, Cho MH, Jadhav A, Smith CS, Inglese J, Portier CJ, Tice RR, Austin CP (2007) Compound cytotoxicity profiling using quantitative high-throughput screening. Environ Health Perspect, in press, 10.1289/ehp.10727
31. MDL (2006) MDL Toxicity Database, MDL, San Ramon
32. Renner S, Fechner U, Schneider G (2006) Pharmacophores and pharmacophore searches. In: Langer T, Hoffmann RD (eds) Wiley-VCH, Wienheim, Germany 32:49–79
33. Breiman L (2001) Random forests. Machine Learning 45:5–32
34. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Chapman & Hall/CRC, Boca Raton, FL
35. R Development Core Team (2005) A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria
36. Cho SJ, Hermsmeier MA (2002) Genetic algorithm guided selection: variable selection and subset selection. J Chem Inf Comput Sci 42:927–936
37. Forrest S (1993) Genetic algorithms: principles of natural selection applied to computation. Science 261:872–878
38. Leardi R (2001) Genetic algorithms in chemometrics and chemistry. J Chemo 15:559–569
39. Derksen S, Keselman HJ (1992) Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Statis Psychol 45:265–282
40. Kirkpatrick S, Gelatt JCD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680
41. Sutter JM, Dixon SL, Jurs PC (1995) Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. J Chem Inf Comput Sci 35:77–84
42. Hanley JA, Mcneil BJ (1982) The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143:29–36
43. Accelrys Scitegic Pipeline Pilot, San Diego, 2007
44. Cerri A, Serra F, Ferrari P, Folpini E, Padoani G, Melloni P (1997) Synthesis, cardiotonic activity, and structure-activity relationships of 17 beta-guanylhydrazone derivatives of 5 beta-androstane-3 beta, 14 beta-diol acting on the Na+,K(+)-ATPase receptor. J Med Chem 40(21):3484–3488
45. Grove SJ, Kaur J, Muir AW, Pow E, Tarver GJ, Zhang MQ (2002) Oxyaniliniums as acetylcholinesterase inhibitors for the reversal of neuromuscular block. Bioorg Med Chem Lett 12(2):193–196
46. Leader H, Wolfe AD, Chiang PK, Gordon RK (2002) Pyridophens: binary pyridostigmine-aprophen prodrugs with differential inhibition of acetylcholinesterase, butyrylcholinesterase, and muscarinic receptors. J Med Chem 45(4):902–910
47. Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) Similarity to molecules in the training set is a good discriminator

for prediction accuracy in QSAR. J Chem Inf Comput Sci 44(6):1912–1928

48. Guha R, Dutta D, Jurs PC, Chen T (2006) Local lazy regression: making use of the neighborhood to improve QSAR predictions. J Chem Inf Model 46(4):1836–1847
49. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant CA, Myatt G, Nikolova-Jeliazkova N, Patlewicz GY, Perkins R, Roberts D, Schultz T, Stanton DW, van de Sandt JJM, Tong W, Veith G, Yang C (2005) Current status of methods for defining the applicability domain of (Quantitative) structure–activity relationships. The Report and Recommendations of ECVAM Workshop 52. Altern Lab Anim 33(2):155–173