

Oversampling to Overcome Overfitting: Exploring the Relationship between Data Set Composition, Molecular Descriptors, and Predictive Modeling Methods

Chia-Yun Chang,[†] Ming-Tsung Hsu,[‡] Emilio Xavier Esposito,^{||} and Yufeng J. Tseng^{*,†,‡,§,#}

[†]School of Pharmacy, College of Medicine, National Taiwan University, No.1, Sec.1, Jen-Ai Road, Taipei, Taiwan 100

[‡]Genome and Systems Biology Degree Program, College of Life Science, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106

^{||}exeResearch, LLC, 32 University Drive, East Lansing, Michigan 48823, United States

[§]Department of Computer Science and Information Engineering, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106

[#]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, No.1 Sec.4, Roosevelt Road, Taipei, Taiwan 106

Supporting Information

ABSTRACT: The traditional biological assay is very time-consuming, and thus the ability to quickly screen large numbers of compounds against a specific biological target is appealing. To speed up the biological evaluation of compounds, high-throughput screening is widely used in the fields of biomedical, biological information, and drug discovery. The research presented in this study focuses on the use of support vector machines, a machine learning method, various classes of molecular descriptors, and different sampling techniques to overcome overfitting to classify compounds for cytotoxicity with respect to the Jurkat cell line. The cell cytotoxicity data set is imbalanced (a few active compounds and very many inactive compounds), and the ability of the predictive modeling methods is adversely affected in these situations. Commonly imbalanced data sets are overfit with respect to the dominant classified end point; in this study the models routinely overfit toward inactive (noncytotoxic) compounds when the imbalance was substantial. Support vector machine (SVM) models were used to probe the proficiency of different classes of molecular descriptors and oversampling ratios. The SVM models were constructed from 4D-FPs, MOE (1D, 2D, and 21/2D), noNP+MOE, and CATS2D trial descriptors pools and compared to the predictive abilities of CATS2D-based random forest models. Compared to previous results in the literature, the SVM models built from oversampled data sets exhibited better predictive abilities for the training and external test sets.



■ INTRODUCTION

With more and more experimentally (experimentally refers to wet or bench chemistry and biology) high-throughput screening (HTS) data becoming available, the need, ability, and desire to use it to construct classification predictive models is commonplace. Unfortunately, the HTS data are noisy and at times incomplete, but the biggest drawback attributed to the data is its inherent imbalanced nature. Typically, there are very few active (also known as “positive”) compounds (sometimes called samples), while there are a plethora of inactive (“negative”) compounds; thus the data are considered imbalanced. There are several methods and protocols available to accommodate a disproportioned data set and one of the more common methods is oversampling where the minority class – the smaller number of samples – is replicated to result in an equal number of samples in the two classes. This method has an opposing methodology fittingly called undersampling

where only a portion of the majority class – the larger number of samples – equal to the number of minority samples is selected to construct the training set. The research discussed herein examines the Jurkat cell cytotoxicity HTS data set, different structural (molecular) descriptor sets, and the effects of oversampling on classification predictive models. The protocols and methods presented outline best practices for approaching the construction and application of a classification predictive model from an imbalanced data set.

It is common for chemical compounds – especially drugs – used to treat various diseases to cause adverse effects and are at times cytotoxic. Thus, toxicity testing during the drug development process is necessary to ensure the safety of the patient and for the success of the research project. Tradition-

Received: January 23, 2013

Published: March 6, 2013

ally, animal models are used for the toxicity testing; however, animal trials are time and cost consuming. New directions and methodologies in toxicity testing for risk assessment have been widely addressed,^{1–4} and computational toxicology methodologies have become feasible to reduce the overall costs and need for animal toxicity testing. Using the large amount of HTS data and machine learning algorithms to construct predictive toxicity models aids in the early classification of compounds,^{1,5,6} and there are several examples of these computational methodologies being developed for early stage toxicity predictions and safety assessments.^{5,7} Successful examples of *in silico* cytotoxicity predictions via quantitative structure–activity relationship (QSAR) models include the prediction of the following: (i) phenoxyl radical-based toxicity in a fast growing murine leukemia cell line,⁸ (ii) toxicity of imidazolium-derived ionic liquids in a human Caco-2 cell line,⁹ and (iii) cellular toxicity in high throughput cell proliferation screening data for 13 cell lines.¹⁰ The power and reliability of predictive models relies on the fruitful interaction between the data set, the physicochemical descriptors, and the machine learning algorithm.¹¹

Quantitative high-throughput screening (qHTS) performed by the National Institutes of Health (NIH) Chemical Genomics Center (NCGC) has been developed to measure the cell responses of a large number of compounds in a short time period.^{2,12} The quality of data (experimental end points) from qHTS is more reliable than the data from traditional HTS. This improved reliability in experimental end points is due to the method used to obtain the data. By using concentration–response curves (CRCs) instead of a single concentration to determine whether or not a compound is active greatly reduces the false positive (*fp*) and false negative (*fn*) rate.² Although qHTS can rapidly generate high quality analyses for numerous compounds on cell viability, the screened compounds typically exhibit an active/inactive imbalance that is commonly seen in HTS results. The imbalance is due to the hypothesis that there are far fewer active compounds for a specific biological system than inactive compounds. While the reliability of biological (experimental) data from qHTS studies (compared to HTS) has been improved, the challenge remains the same: construct high quality classification models from imbalanced data sets.¹³

Imbalanced data set processing is an active area of research and has been explored and discussed in several studies,^{13–16} yet there are several ways to account for and safeguard against the obscuring of useful information contained in the data set. Support vector machines (SVMs) combined with resampling strategies redistribute the initially imbalanced data and improves the classification ability of the predictive model.^{17,18} To keep all the information embedded within a data set, oversampling methods are more suitable than undersampling methods that suffer from information loss^{16,17} due to their basic design. Oversampling brings the data set to a class equilibrium by simply duplicating samples from the minority class.¹⁹ The caveat to this seemingly simplistic method of increasing the number of minority class entities until there is an equal number of each class is that the best resampling ratio (active:inactive) varies for different data domains,¹⁴ and selecting the optimal sampling ratio can optimize and tune the classification model.

Several machine learning algorithms have been applied to cytotoxicity predictions, such as neural network,²⁰ random forest,²¹ and decision tree,²² and the performance of these toxicity classification models has been evaluated and compared.²³ Although SVM and artificial neural network

(ANN) algorithms perform well on toxicity prediction in the comparative analysis,²³ another comparison showed that SVM-based mutagenicity predictions are more accurate than ANN-based predictions.²⁴ Additionally, the random forest (RF)^{10,25,26} predictive modeling method has been used for compound classification and toxicity prediction as well as the SVM algorithm.²⁷ The SVM-based cytotoxicity classification models presented in this study are compared to the models constructed with the RF algorithm by Guha and Schürer.¹⁰

Random forest-based cytotoxicity classification models of screened compounds from NCGC have been curated and constructed for 13 different cell lines by Guha and Schürer.¹⁰ The NCGC Jurkat model was used to predict the toxicity classification of the Scripps Jurkat data set from Molecular Library Screening Center Network (MLSCN). The cytotoxicity classification accuracy of the Guha and Schürer¹⁰ CATS2D-based RF model applied to their test set (the Scripps/MLSCN data set that is similar to the test set used in this study) was 67.5%. This is a respectable ability for a classification model applied to an external test set, but examining the sensitivity (ability to predict known active compounds) and specificity (ability to predict known inactive compounds) indicates that their model was skewed toward being able to better predict known actives; 76.3% and 26.0%, respectively. As will be shown, the performance of toxicity classification models can be improved by using different machine learning algorithms, descriptor classes, and sampling strategies.

Cell proliferation, viability assays, and qHTS methods provide a large collection of bioassay data for constructing cytotoxicity classification models and thus allowing the prediction of chemical compounds. In this study, the influences of different trial descriptor pools (4D-Fingerprint; 1D, 2D, and 21/2D MOE; and 4D-Fingerprint+MOE), data set composition (Jurkat-specific cell line biological end points or a collection of compounds that are known to be cytotoxic), oversampling strategies (various oversampling ratios), and model construction methods (SVM and RF) for the cytotoxicity prediction based on an imbalanced data set from qHTS assays are explored and discussed. Cytotoxicity classification models that can accommodate imbalanced data can inform scientists (specifically biologists in this case) whether particular compounds are cytotoxic (active) prior to biological experimentation and thus aid in the prioritization of compounds for further ADME/toxicology testing before being advanced to *in vivo* assays.

MATERIAL AND METHODS

Experiment Measurement of Cell Viability Data.

Evaluating a compound's cytotoxicity is critical when developing potential human therapeutics for obvious reasons; most notably therapeutic induced death. To aid in the investigation of the physicochemical properties that make compounds cytotoxic a qHTS campaign was designed by the NIH's Chemical Genomics Center. The qHTS study measured the metabolic activity of a suspended cell line after 40 h of incubation at 37 °C with the compound of interest. The qHTS used the human T-cell line Jurkat Clone E6-1, and each compound was evaluated at a final concentration of 4 μ M (μ M). To determine if the cells occupying the well were viable (alive) a luciferase-based cell proliferation/viability assay was employed to measure the amount of ATP present in the microtiter plate well via luminescence; the concentration of ATP corresponds to the amount of luminescence. Wells that do

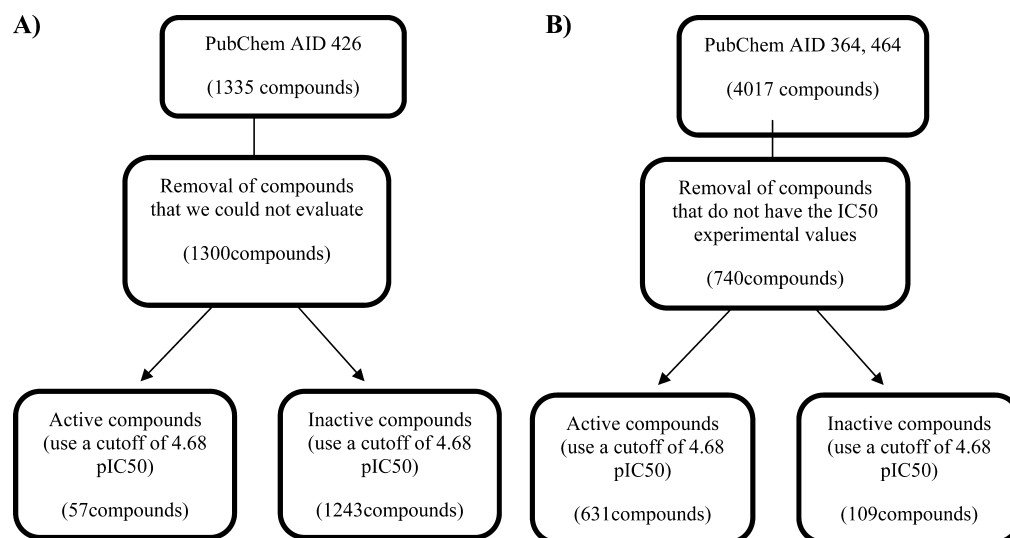


Figure 1. A) Training data set of PubChem BioAssay AID 426 that used a cutoff value of 4.68 pIC_{50} to classify the active and inactive compounds. B) Testing data set of PubChem BioAssay AID 364 and 464 that used a cutoff value of 4.68 pIC_{50} to classify the active and inactive compounds.

not luminescence — due to the lack of ATP — indicate that the catalytic conversion of luciferin into oxyluciferin has not occurred and the absence of ATP is correlated with cell death.

Data Sets. The training set for this cytotoxicity study is a subset of the PubChem BioAssay database AID: 426 — as indicated below — that has two components: (i) the raw data set that contains all the compounds and their designation as cytotoxic (active) or not (inactive) for the 13 cell lines and (ii) the compounds that have been tested for specific cytotoxicity on the Jurkat cell line. The test set is a combination of PubChem BioAssay databases AIDs 364 and 464 that focuses on cytotoxicity of the Jurkat cell line. The experimentally determined qHTS end points of the training set are considered superior to the HTS end points of the test set.

Training Set. The training data set is based on the PubChem BioAssay database AID: 426, a qHTS assay for cell viability of the Jurkat cell line curated by the NIH Chemical Genomics Center. The data set, at the time of access, contained 1335 compounds with IC_{50} experimental values for all the compounds. In order to compare our results with those in the literature,¹⁰ a cutoff value of 4.68 on the pIC_{50} ($-\log_{10} \text{IC}_{50}$) scale was selected. Compounds with a pIC_{50} value greater than 4.68 were classified as active (cytotoxic), and the remainder of the compounds were considered inactive. This method of classifying compounds as active or inactive is the same as the one described by Guha and Schürer¹⁰ where compounds were divided based on a cutoff value that is “two standard deviations above the mean pIC_{50} ” for the Jurkat cell line. Thirty-five compounds were removed from the training set because descriptors could not be calculated for them, resulting in a training set of 1300 compounds with 57 active and 1243 inactive compounds (Figure 1a).

Test Set. The test set is a fusion of two Jurkat cell PubChem BioAssay databases; AID 364 and AID 464. Fifty of database AID 364’s 3311 compounds possessed experimental IC_{50} values, while all of database AID 464’s 706 compounds have experimental IC_{50} values. After combining these two data sets and removing 16 compounds that were unable to have the required molecular descriptors calculated, the test set contained 740 compounds (50 compounds from AID 364 plus 706 compounds from AID 464 minus 16 compounds) with 631

active and 109 inactive compounds (Figure 1b) after applying the pIC_{50} cutoff value of 4.68. The data sets for the test set was originally provided by The Scripps Research Institute Molecular Screening Center to PubChem, and the end points—cytotoxic or not—were determined using traditional HTS.

Molecular Descriptors. Because the chemical structure of a compound determines the physicochemical properties,²⁸ commonly referred to as molecular features and descriptors, focusing on a specific substructure — toxicophores — within a series of compounds is a common method of constructing toxicity classification models.^{29,30} The selection of suitable molecular descriptors has a direct impact on the performance of predictive models and thus the quality of the classification models.^{31,32} The Jurkat cell cytotoxicity study by Guha and Schürer¹⁰ used the BCI fingerprints and CATS2D descriptors³³ to construct toxicity classification models. In our previous toxicity studies,^{34,35} successful classification models were constructed by employing 4D-Fingerprints³⁶ and Molecular Operating Environment (MOE)³⁷ molecular descriptors. The CATS2D molecular descriptors are only discussed with respect to the portion of this study concerned with the comparison of model construction methods — SVM compared to random forest. Comparing the performance of models developed using different descriptor sets is necessary to identify the descriptor class that best captures the key physicochemical properties for cytotoxicity and thus constructing a robust model.

Universal 4D-Fingerprints. The detailed formalism to compute the 4D-Fingerprints (4D-FPs) has been published in a previous research article,³⁶ and the methodology is only summarized herein. The 4D-FP method generates a set of molecular fingerprints that is divided into pharmacophore elements designed to capture the 3D size, shape, and conformational flexibility of a molecule while embedding conformational averaged molecular information.

The first step in constructing the 4D-FPs is the generation of the conformation ensemble profile (CEP) of each molecule via molecular dynamic simulation (MDS). For each compound 36 molecular similarity main distance-dependent matrix (MDDM) are constructed, from the same term or a cross-term, based on the interaction pharmacophore element (IPE) pair type from eight IPEs; they are as follows: all atoms, nonpolar atoms, polar

positive atoms, polar negative atoms, hydrogen bond acceptor atoms, hydrogen bond donor atoms, aromatic atoms, and non-hydrogen (or heavy) atoms. The eigenvector and corresponding eigenvalues are derived from the diagonalization of the MDDM. When the IPE types are the same, the MDDM is an upper/lower triangular matrix and can be directly diagonalized. The resulting eigenvalues are then normalized and sorted in numerically descending order. When the IPE types are different, the MDDM will likely be rectangular because the number of IPE elements is probably different. These MDDM matrices of different sizes are made square – like the MDDM matrices of same IPE type – by multiplying one of the two IPE matrices by the transpose of the other MDDM. All of the eigenvalues for all of the MDDMs for all IPE pairs are utilized as the universal 4D-Fingerprints for one molecule. For each test set molecule, the number of eigenvalues of each specific IPE type is set to the maximum size of the training set. If the number of eigenvalues for a test set compound is greater than those for the training set, then the excessive eigenvalues are disregarded. The test set descriptor matrix is also normalized using the same protocol used for the training set.

MOE Descriptors. A set of 306 1D, 2D, and 21/2D (3D molecular properties mapped to a single numerical value) descriptors from MOE 2010.10³⁸ was calculated for inclusion in the descriptor pool. The 1D molecular descriptors include the number of specific atoms, atom types (hydrogen bond acceptors and donors), and number of single, double, and triple bonds. The 2D molecular descriptors are numerical features derived from the connection table representing a molecule and include physical properties, subdivided surface areas, atom counts, bond counts, Kier and Hall connectivity and Kappa Shape indices, adjacency and distance matrix descriptors containing BCUT and GCUT descriptors, pharmacophore feature descriptors, and partial charge descriptors. The 21/2D molecular descriptors are dependent on the conformation of a molecule and include the following: potential energy, surface area, volume, shape, and charge descriptors. A description of MOE molecular descriptors can be found on the Chemical Computing Group, Inc. Web site.³⁹

CATS2D Descriptors. The Chemically Advanced Template Search (CATS) molecular descriptors⁴⁰ are based on the distance (number of bonds) between pairs of topological pharmacophore elements within the compound. The five-pharmacophore elements are as follows: anion, cation, hydrogen bond acceptor, hydrogen bond donor, and hydrophobic (lipophilic) atoms. The number of occurrences for a pharmacophore-pair is binned based on the number of atomic bonds between the two atoms. The number of bonds considered for inclusion spanned from 0 to 9 and when the number of bonds between two atoms was greater than 9, those distances were added to the 9 bonds bin. The 0 bond separation bin counts the number of each pharmacophore element in the compound. The 15-pharmacophore element interaction pairs along with the 10 bins for number of separating bonds (though, nonsimilar pharmacophore pairs with zero bonds separating the two atoms should not exist) results in 150 CATS2D molecular descriptors. Originally, these descriptors were created for virtual screening and thus their ability to aid in the construction of robust and applicable classification models is explored.

Rebalancing the Data Set. The cell cytotoxicity data set is imbalanced, and this situation is often experienced with HTS data sets. To counteract the imbalanced nature of the cell

cytotoxicity data set, oversampling was employed to increase the number of minority class members (cytotoxic compounds; active) and resulted in an evenly distributed data set. Oversampling increases the number of minority class members in the training set to a population size equal to – or a defined ratio of – the majority class. The advantage of oversampling is that no information from the original training set is lost since all members from the minority and majority classes are retained. Retaining all of the members – through the duplication of members in the smaller subsets – results in oversampling's main drawback, the effective size of the training set is greatly increased and, thus, so are the required computational resources.

The research presented herein took advantage of the oversampling methodology and created oversampled data sets with active-to-inactive ratios of 1:1, 1:2, and 2:3 to find the optimal condition for each descriptor set. It has been reported by Gazzah and Amara¹⁶ that the best active-to-inactive ratio for oversampling is dependent on the data set of interest. The method of overcoming the imbalanced data set in this study is significantly different than the method implemented by Guha and Schürer.¹⁰ The training and “validation” test sets of Guha and Schürer were constructed by randomly selecting a subset of compounds from the data set for the test set and using the remaining compounds for the training set. Specifically, the test set was constructed by selecting 20% of the known cytotoxic (actives) compounds followed by randomly selecting inactive compounds to replicate the active:inactive ratio of the complete data set. The leftover active compounds were combined with the same number of randomly selected inactive compounds to form the training set resulting in a 1:1 ratio of active-to-inactive compounds.

Model Creation Methods. Two types of predictive models were constructed from the imbalanced data set to explore the impact of oversampling. To provide an equivalent comparison to the Guha and Schürer¹⁰ study, random forest²¹ models were also constructed, while support vector machine^{41,42} models were constructed to compare the ability of the trial descriptor pools.

Random Forest (RF). The random forest²¹ methodology randomly selects a subset of samples (compounds) from a data set along with a random selection of molecular descriptors (independent variable) to construct a predisposed number of predictive models (ensemble) whose results are combined into a single model. Random forest models are appealing because the need for descriptor selection is not required, and they are equally adept at creating robust models for continuous and/or binary end points (biological measures of activity). Models were constructed and validated using the randomForest package v4.6–7⁴³ in R v2.15.2.⁴⁴ Two user definable tuning arguments are available for the randomForest function, *ntree* and *mtry*, and were used to construct optimal RF models. The *ntree* argument defines the number of ‘trees’ to create within the forest and was set from 500 to 1000 using an interval of 100 (i.e. 500, 600, 700, 800, 900, and 1000), while the *mtry* parameter that specifies the number of randomly selected independent variables to include in each model. The *mtry* value was set to a value between 1 and 16.

Support Vector Machine (SVM). A support vector machine^{41,42} is a supervised machine-learning technique that applies (creates) a hyperplane within the descriptor space in an attempt to separate (classify) the samples. The end points for each compound (sample) of the Jurkat-specific cell line data set

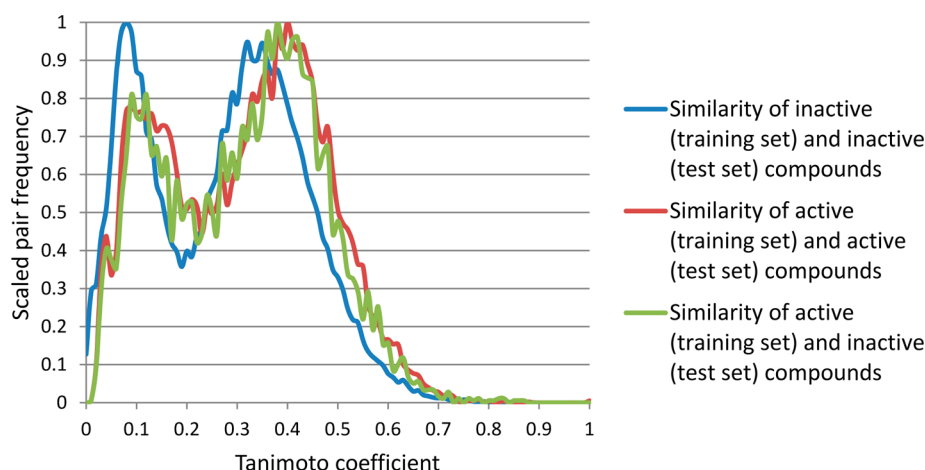


Figure 2. Similarity of active compounds (training data set) and inactive compounds (testing data set).

are binary – the compounds are classified as active (1s) or inactive (0s). Models were constructed and validated using the LIBSVM v2.91⁴⁵ application.

Model Evaluation. To assess the predictive performance of the constructed binary classification models several evaluation methods were employed. The common classification evaluation methods of *accuracy*, *sensitivity*, and *specificity* – eqs 1, 2, and 3, respectively – were calculated along with the more advanced methods of geometric means (*G-means*; eq 4) and Cohen's *kappa* (κ ; eq 5) that are not affected by imbalanced data sets. *Accuracy* represents the proportion of correctly predicted classifications for the entire data set, *sensitivity* is the fraction of active samples that were successfully classified, and *specificity* is the fraction of inactive samples that were properly classified. The equations to calculate these model evaluation methods are

$$\text{accuracy} = \frac{tp + tn}{tp + fn + tn + fp} = \frac{tp + tn}{\text{Total number of samples}} \quad (1)$$

$$\text{sensitivity} = \frac{tp}{tp + fn} = \frac{tp}{\text{Number of active samples}} \quad (2)$$

$$\text{specificity} = \frac{tn}{tn + fp} = \frac{tn}{\text{Number of inactive samples}} \quad (3)$$

The variables in eqs 1–3 are as follows: *tp* is the number of true positives (active compounds that are correctly predicted to be active); *fn* is the number of false negatives (active compounds that are incorrectly predicted to be nonactive); *tn* is the number of true negative (nonactive compounds that are correctly predicted to be nonactive); and *fp* is the number of false positive (nonactive compounds that are incorrectly predicted to be active). Because *accuracy* can be swayed by a data set that is heavily composed of active or inactive samples, solely evaluating a model based on the *accuracy* value is not advisable. For example, if a major portion of a data set is considered “active” (80% of the samples), then a model that classifies most of the samples as active regardless of their true category will have an *accuracy* value that is considered “good”. Instead, if the model is evaluated on its ability to correctly predict active samples and inactive samples using *sensitivity* and *specificity*, then a better understanding of the model's ability is displayed. Combining *sensitivity* and *specificity* into a single value via the geometric mean (*G-means*) allows for a simple way to evaluate

the model's ability to correctly classify active and inactive samples using the formula

$$G\text{-means} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (4)$$

In addition to the above model evaluation methods, Cohen's *kappa* (κ) can be used to measure the agreement between classification models or predicted and known classifications.⁴⁶ It is defined as

$$\text{Cohen's } \kappa = \frac{\text{Pr}(a) - \text{Pr}(e)}{1 - \text{Pr}(e)} \quad (5)$$

where $\text{Pr}(a)$ is the relative observed agreement between the predicted classification of the model and the known classification, and $\text{Pr}(e)$ is the hypothetical probability of chance agreement. The $\text{Pr}(a)$ and $\text{Pr}(e)$ values are calculated from a confusion matrix. Cohen's *kappa* analysis returns values between −1 (no agreement) and 1 (complete agreement). Predictive models, when compared to the known classification of the data set, with Cohen's *kappa* values between −1.0 and 0.4 indicate that the model is a poor predictor, values between 0.4 and 0.6 indicate that the model is average, values between 0.6 and 0.8 imply that the model is acceptable, and values between 0.8 and 1.0 denote that the model is highly predictive. While *G-means* is the primary model evaluation method for this study, this quintet of classification model evaluation measures is included to provide a complete view of the classification models' predictive abilities.

To further validate the ability of the predictive models, Y-scrambling⁴⁷ was applied to the training set. Y-scrambling is the well-known method of scrambling the known end points, refitting the previously constructed model, and evaluating the “new” model. Typically, and in this study, the known end points are scrambled 1000 times, and the predictive ability of the model is severely compromised (data not shown).

Molecular Similarity. Cytotoxicity may be caused by many reasons, one of which is the structural characteristics of different toxic compounds. Therefore, the similarity of active and inactive compounds between the training set and test set and the similarity between active compounds in the training set and inactive compounds in the test set were calculated. The molecular similarity between compounds was calculated using the provided PubChem 2D descriptors and resulted in a Tanimoto Coefficient value. The molecular similarity within the

Table 1. Performance of the SVM Model for 4D-Fingerprints, MOE, and noNP+MOE Descriptor Sets with Raw Data and Jurkat Cell Specific Data

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa
complete AID 426 data set	4D-FPs	training set	89.08 (1158/1300)	92.98 (53/57)	88.90 (1105/1243)	0.909	0.386
		test set	42.30 (313/740)	37.72 (238/631)	68.81 (75/109)	0.509	0.028
	MOE	training set	83.69 (1088/1300)	73.68 (42/57)	84.15 (1046/1243)	0.787	0.229
		test set	35.54 (263/740)	28.37 (179/631)	77.06 (84/109)	0.468	0.021
	noNP+MOE	training set	89.38 (1162/1300)	57.89 (33/57)	90.83 (1129/1243)	0.725	0.278
		test set	61.08 (452/740)	64.98 (410/631)	38.53 (42/109)	0.500	0.022
Jurkat subset of AID 426	4D-FPs	training set	92.00 (46/50)	84.62 (11/13)	94.59 (35/37)	0.895	0.792
		test set	55.14 (408/740)	56.26 (355/631)	48.62 (53/109)	0.523	0.027
	MOE	training set	62.00 (31/50)	100.00 (13/13)	48.65 (18/37)	0.697	0.330
		test set	66.08 (489/740)	70.52 (445/631)	40.37 (44/109)	0.534	0.075
	noNP+MOE	training set	52.00 (26/50)	100.00 (13/13)	35.14 (13/37)	0.593	0.220
		test set	72.43 (536/740)	80.51 (508/631)	25.69 (28/109)	0.455	0.053

training and test sets and between the two sets of compounds is summarized in Figure 2.

RESULTS AND DISCUSSION

The models presented and discussed below were constructed from either (i) the entire PubChem data set (AID 426) or (ii) a subset of this data set that specifically had Jurkat cell viability end point values. The test set is constructed from PubChem AIDs 364 and 464 and was evaluated with the classification models constructed from the full and Jurkat subset of AID 426 to determine the abilities of the models. External test sets are considered an excellent method to evaluate the predictive ability of models. Since the provided end points for the combined test set were determined via HTS and not *quantitative* HTS, the reduced quality of the experimental end points for the test set provides a real world example of “less than ideal” data. There is a striking contrast for the test set of Cohen’s *kappa* values – and to a lesser extent for the *G-means* values – compared to those for the training set. Based on molecular similarity analysis, the active training set compounds are not similar to the active test set compounds indicating that the test set might be beyond the “domain” of the predictive models. There are cases where compounds in the training set have similar molecular structure – based on molecular similarity analysis – to compounds in the test set yet are respectively considered active and inactive or vice versa.

Prediction Models of Cell Viability. Three descriptor pools were used to explore the important structural features relating to Jurkat cell cytotoxicity, specifically the following: 4D-Fingerprints, MOE (1D+2D+21/2D), and noNP+MOE (4D-Fingerprints (excluding NP) + MOE (1D+2D+21/2D)). The inclusion of these three trial descriptor pools along with the CATS2D descriptors used in the Guha and Schürer¹⁰ study provides the ability to compare the model creation method and the molecular descriptors while also exploring the ability of oversampling. The noNP+MOE descriptor pool consisted of MOE 1D, 2D, and 21/2D molecular descriptors and 4D-FPs except for those that included nonpolar (NP) interactions. Based on the results of the 4D-FPs only model, the independent variables containing information regarding non-polar IPEs in the 4D-FPs descriptor pool were removed from the combined 4D-FP + MOE descriptor pool. The classification models constructed from the noNP+MOE descriptor pool have better predictive ability for the test set based on the significantly better *sensitivity* (65.0%) compared to that of 4D-FP (37.7%) and MOE (28.4%) models. This trend was also seen for SVM

models constructed from the Jurkat subset of AID 426. It was therefore concluded that removing the NP descriptors of the 4D-FP descriptor pool and including the MOE descriptors provided a robust and mixed descriptor pool that will be better able to capture important physicochemical properties related to cytotoxicity. To alleviate the Jurkat cell cytotoxicity data set’s imbalanced nature, oversampling was used to address the inherent problems with constructing predictive models from skewed data sets. The contributions of molecular descriptors, sampling, filtering, and model creation methods were explored to better understand the impact of these components on an imbalanced data set.

SVM Model Constructed from the Complete AID 426 Data Set. Without applying any filters or sampling methods to AID 426, the training set contained 1300 compounds, the self-prediction *specificity* (inactive compounds) for the training set for each of the three descriptor sets 4D-Fingerprints, MOE, and noNP+MOE is 88.9%, 84.2%, and 90.8% while the *sensitivity* was 93.0%, 73.7%, and 57.9%, respectively (Table 1). Applying these models to the test set returned a different trend for the *specificity* (68.8%, 77.1%, 38.5%) and *sensitivity* (37.7%, 28.4%, 65.0%) values. Based on these values it can be stated that the three descriptor sets are overfitting with respect to the inactive compounds for the complete training set and the test set. This is most likely the result of inactive-state bias exhibited in the training set. Based on *G-means* and *kappa* values, it can be seen that the 4D-FPs provides slightly better results for models constructed from the entire AID 426 data set.

SVM Model Constructed from the Jurkat Cell Specific Data. Because the goal of this modeling effort is to better understand how molecular descriptor classes and curated data sets affect the predictive ability of the models, only compounds from AID 426 specifically indicated with end point values for the Jurkat cell line were retained for the next set of classification models. This resulted in a training set with 13 active compounds (toxic) and 37 inactive compounds. The 50 compounds were used to construct models using the 4D-Fingerprints, MOE, and noNP+MOE descriptor sets. Overall the *specificity* (94.6%, 48.7%, and 35.1%; inactive compounds) for the Jurkat cell line training set severely decreased, but the *sensitivity* (84.6%, 100%, and 100%; active compounds) was greatly improved. Upon further inspection, the *sensitivity* of the models constructed with only the MOE descriptors exhibited a marked improvement from 73.7% to 100%, and the noNP+MOE model’s *sensitivity* increased from 57.9% to 100% (see Table 1). These changes are directly attributed to the

Table 2. Performance of the SVM Model for the 4D-Fingerprints Descriptor Set with the Oversampling Ratios (Active:Inactive): 1:1, 1:2, and 2:3

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa
complete AID 426 data set	original	training set	89.08 (1158/1300)	92.98 (53/57)	88.90 (1105/1243)	0.909	0.386
		test set	42.30 (313/740)	37.72 (238/631)	68.81 (75/109)	0.509	0.028
	1:1	training set	91.19 (2277/2497)	100.00 (1254/1254)	82.30 (1023/1243)	0.907	0.824
		test set	45.41 (336/740)	42.00 (265/631)	65.14 (71/109)	0.523	0.032
	1:2	training set	90.05 (1684/1870)	98.25 (616/627)	85.92 (1068/1243)	0.919	0.790
		test set	45.27 (335/740)	41.84 (264/631)	65.14 (71/109)	0.522	0.031
	2:3	training set	92.14 (1933/2098)	98.25 (840/855)	87.93 (1093/1243)	0.929	0.841
		test set	44.73 (331/740)	40.73 (257/631)	67.89 (74/109)	0.526	0.038
	Jurkat subset of AID 426	original	92.00 (46/50)	84.62 (11/13)	94.59 (35/37)	0.895	0.792
		test set	55.14 (408/740)	56.26 (355/631)	48.62 (53/109)	0.523	0.027
	1:1	training set	90.79 (69/76)	100.00 (39/39)	81.08 (30/37)	0.900	0.815
		test set	44.05 (326/740)	40.41 (255/631)	65.14 (71/109)	0.513	0.024
	2:3	training set	88.89 (56/63)	84.62 (22/26)	91.89 (34/37)	0.882	0.769
		test set	56.89 (421/740)	58.48 (369/631)	47.71 (52/109)	0.528	0.035

composition of the training set. The complete 4D-FP training set experienced slight changes to the *specificity* and *sensitivity* values, while the *sensitivity* of MOE and noNP+MOE models were both improved at the cost of reduced *specificity*. Regardless of the varied changes in the predictive ability of the models, the MOE and noNP+MOE descriptor sets still exhibited overfitting. The results indicated that the predictive ability of the Jurkat-only models were better able to indicate active compounds (*sensitivity*: 56.3%, 70.5%, and 80.5%) within the test set compared to the SVM models constructed from all 1300 compounds of the AID 426 data set (*sensitivity*: 37.7%, 28.4%, and 65.0%). The moderate overall improvement seen in the prediction of active compounds from the test set does not indicate that any specific model is superior. The removal of compounds that were not tested specifically for cytotoxicity within the Jurkat cell line allows the model to emphasize important molecular features for Jurkat cell cytotoxicity but unfortunately the models exhibited. By focusing the data set — based on end point — the models have shifted the overfitting of the training set from the inactive compounds (complete AID 426 data set) to the active compounds (Jurkat subset of AID 426) with the exception of the models constructed from the 4D-FPs that remained statistically similar.

While the performance of the Jurkat cell line only models was similar to the complete AID 426 data set, the overfitting should not be considered due to “noise” in the data set (full AID 426 data set versus the Jurkat cell line specific compounds from the AID 426 data set) but instead it is most likely due to the imbalanced data set. Oversampling to adjust the ratio of active to inactive compounds is explored to see if overfitting is due to an imbalanced training set and if this imbalance can be reduced.

Oversampling the Imbalanced Data Sets. Three oversampling active to inactive ratios were explored to gauge the impact of balancing the training set. The initial ratio of active-to-inactive compounds for the complete AID 426 data set was 1:22, while the Jurkat specific data set had an active:inactive ratio of 1:3. It could be argued that the Jurkat cell line specific data set is not imbalanced due to its low number of compounds compared to the AID 426 data set, 50 compounds compared to 1300, and the relatively low ratio of active-to-inactive compounds, 1:3 compared to 1:22. But when taking into consideration the overfitting experience within its SVM models toward active compounds, the potential

contributions from an imbalanced data set should be addressed. The first oversampling ratio explored was 1:1 and increased the number of active compounds in the training set to 1254 from 57 (for the complete AID 426 data set) and 39 from 13 (for the Jurkat subset) by replicating all of the active compounds. To retain the complete collection of compounds from the minority class, the minority class compounds were duplicated an integer number of times. For example, to create as close as possible an active:inactive ratio of 1:1 the 57 active compounds of the full AID 426 data set were replicated 22 times resulting in 1254 active compounds. While this provided slightly more — 11 compounds — active compounds than inactive compounds, the data set should be considered balanced. The other oversampling ratios for the full training set were 1:2 and 2:3 increasing the number of active compounds to 627 and 855 active compounds, respectively, while retaining 1243 inactive compounds. The Jurkat cell line training set is constrained to the active:inactive ratios of 1:1 and 2:3 due to the initial active:inactive ratio of 1:3, thus to satisfy the 1:2 ratio increasing both the number of active and inactive compounds is required. The number of active compounds for the Jurkat training set becomes 39 and 26 for the 1:1 and 2:3 active:inactive ratios with the number of inactive compounds held at 37. The oversampling of the full AID 426 data set had an immediate improvement on the ability of the models as seen in Tables 2–4. Overall the oversampling ratios exhibited increases in the *sensitivity* values and a reduction in *specificity* values compared to the original full AID 426. Remember, the *accuracy* value is based on all the predictions (active and inactive), and an imbalanced data set can influence the *accuracy* values. The *accuracy* values presented in Tables 2–4 are now more realistic because the number of active and inactive compounds is closer to equivalent. Additionally, the *G-means* and Cohen's *kappa* values signal that the models have had a significant improvement for the prediction of the training sets.

Descriptors, Sampling, Modeling, and Filtering Contribution in a Classification Model. The 4D-FP models constructed from the entire AID 426 data set had a *G-means* value of 0.909 that reduced to 0.895 for the Jurkat subset (Table 1). Oversampling improved the *G-means* values for 1:2 and 2:3 ratios (actives:inactives) when using the entire AID 426 data set and 4D-FP descriptors with an average value of 0.92 (Table 2). A similar increase for the Jurkat subset is also seen with *G-means* values of approximately 0.89 (Table 2). While the

Table 3. Performance of the SVM Model for the MOE Descriptor Set with the Oversampling Ratios (Active:Inactive): 1:1, 1:2, and 2:3

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa	
complete AID 426 data set	original	training set	83.69 (1088/1300)	73.68 (42/57)	84.15 (1046/1243)	0.787	0.229	
		test set	35.54 (263/740)	28.37 (179/631)	77.06 (84/109)	0.468	0.021	
	1:1	training set	84.74 (2116/2497)	100.00 (1254/1254)	69.35 (862/1243)	0.833	0.694	
		test set	37.30 (276/740)	30.43 (192/631)	77.06 (84/109)	0.484	0.029	
	1:2	training set	79.63 (1489/1870)	100.00 (627/627)	69.35 (862/1243)	0.833	0.603	
		test set	37.30 (276/740)	30.43 (192/631)	77.06 (84/109)	0.484	0.029	
	2:3	training set	81.41 (1708/2098)	94.74 (810/855)	72.24 (898/1243)	0.827	0.635	
		test set	32.84 (243/740)	24.25 (153/631)	82.57 (90/109)	0.447	0.025	
	Jurkat subset of AID 426	original	training set	62.00 (31/50)	100.00 (13/13)	48.65 (18/37)	0.697	0.330
			test set	66.08 (489/740)	70.52 (445/631)	40.37 (44/109)	0.534	0.075
		1:1	training set	78.95 (60/76)	100.00 (39/39)	56.76 (21/37)	0.753	0.574
			test set	56.89 (421/740)	57.37 (362/631)	54.13 (59/109)	0.557	0.063
		2:3	training set	69.84 (44/63)	100.00 (26/26)	48.65 (18/37)	0.697	0.439
			test set	66.08 (489/740)	70.52 (445/631)	40.37 (44/109)	0.534	0.075

Table 4. Performance of the SVM Model for the noNP+MOE Descriptor Set with the Oversampling Ratios (Active:Inactive): 1:1, 1:2, and 2:3

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa	
complete AID 426 data set	original	training set	89.38 (1162/1300)	57.89 (33/57)	90.83 (1129/1243)	0.725	0.278	
		test set	61.08 (452/740)	64.98 (410/631)	38.53 (42/109)	0.500	0.022	
	1:1	training set	88.67 (2214/2497)	100.00 (1254/1254)	77.23 (960/1243)	0.879	0.773	
		test set	45.54 (337/740)	42.95 (271/631)	60.55 (66/109)	0.510	0.016	
	1:2	training set	84.87 (1587/1870)	100.00 (627/627)	77.23 (960/1243)	0.879	0.695	
		test set	45.54 (337/740)	42.95 (271/631)	60.55 (66/109)	0.510	0.016	
	2:3	training set	78.65 (1650/2098)	87.72 (750/855)	72.41 (900/1243)	0.797	0.576	
		test set	60.14 (445/740)	63.71 (402/631)	39.45 (43/109)	0.501	0.020	
	Jurkat Subset of AID 426	original	training set	52.00 (26/50)	100.00 (13/13)	35.14 (13/37)	0.593	0.220
			test set	72.43 (536/740)	80.51 (508/631)	25.69 (28/109)	0.455	0.053
		1:1	training set	72.37 (55/76)	100.00 (39/39)	43.24 (16/37)	0.658	0.439
			test set	66.76 (494/740)	73.06 (461/631)	30.28 (33/109)	0.470	0.025
		2:3	training set	61.90 (39/63)	100.00 (26/26)	35.14 (13/37)	0.593	0.309
			test set	72.43 (536/740)	80.51 (508/631)	25.69 (28/109)	0.455	0.053

Table 5. Best Models for Each of the Three Types of 4D-Fingerprints, MOE, and noNP+MOE Descriptor Sets with the Complete AID 426 Data Set and Jurkat-Specific Subset of AID 426

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa
complete AID 426 data set	4D-FPs (2:3, SVM)	training set	92.14 (1933/2098)	98.25 (840/855)	87.93 (1093/1243)	0.929	0.841
		test set	44.73 (331/740)	40.73 (257/631)	67.89 (74/109)	0.526	0.038
	MOE (1:1, SVM)	training set	84.74 (2116/2497)	100.00 (1254/1254)	69.35 (862/1243)	0.833	0.694
		test set	37.30 (276/740)	30.43 (192/631)	77.06 (84/109)	0.484	0.029
	noNP+MOE (1:1, SVM)	training set	88.67 (2214/2497)	100.00 (1254/1254)	77.23 (960/1243)	0.879	0.773
		test set	45.54 (337/740)	42.95 (271/631)	60.55 (66/109)	0.510	0.016
Jurkat subset of AID 426	4D-FPs (2:3, SVM)	training set	88.89 (56/63)	84.62 (22/26)	91.89 (34/37)	0.882	0.769
		test set	56.89 (421/740)	58.48 (369/631)	47.71 (52/109)	0.528	0.035
	MOE (1:1, SVM)	training set	78.95 (60/76)	100.00 (39/39)	56.76 (21/37)	0.753	0.574
		test set	56.89 (421/740)	57.37 (362/631)	54.13 (59/109)	0.557	0.063
	noNP+MOE (1:1, SVM)	training set	72.37 (55/76)	100.00 (39/39)	43.24 (16/37)	0.658	0.439
		test set	66.76 (494/740)	73.06 (461/631)	30.28 (33/109)	0.470	0.025

increase in *G-means* values for the 4D-FP models (Jurkat subset) is not as substantial of an increase for the full data set, the difference in model ability is most likely due to focusing the data set to compounds with end points for the Jurkat subset.

The predictive models created with MOE and noNP+MOE descriptor pools experienced similar improvements in their *G-means* values for the complete AID 426 and the Jurkat cell line data sets, Tables 3 and 4. It is impressive how well the

oversampled models performed based on their ability to classify active and inactive compounds, but evaluating the models' abilities based on their *G-means* values illustrates how the model performs on the whole data set. Regardless of the descriptors and oversampling ratios being employed, the overall *G-means* performance of the oversampling models outperforms the models constructed from the original complete AID 426 and Jurkat cell line data sets.

Table 6. Comparative Models between SVM and RF for 4D-Fingerprints (the Best in This Study) and CATS2D Descriptor Sets (Guha and Schürer¹⁰) with the Same Oversampling Ratio (Active:Inactive = 1:1) of the Complete AID 426 Data Set

			Acc (%)	Sen (%)	Spe (%)	G-means	kappa
complete AID 426 data set	4D-FPs (1:1, SVM) ^a	training set	91.19 (2277/2497)	100.00 (1254/1254)	82.30 (1023/1243)	0.907	0.824
		test set	45.41 (336/740)	42.00 (265/631)	65.14 (71/109)	0.523	0.032
	4D-FPs (1:1, RF) ^a	training set	74.33 (1856/2497)	49.12 (616/1254)	99.76 (1240/1243)	0.700	0.488
		test set	25.00 (185/740)	14.74 (93/631)	84.40 (92/109)	0.353	−0.003
	CATS2D (1:1, SVM) ^a	training set	75.17 (1877/2497)	61.40 (770/1254)	89.06 (1107/1243)	0.739	0.504
		test set	68.24 (505/740)	75.28 (475/631)	27.52 (30/109)	0.455	0.022
	CATS2D (1:1, RF) ^a	training set	79.98 (1997/2497)	63.16 (792/1254)	96.94 (1205/1243)	0.782	0.600
		test set	36.22 (268/740)	28.53 (180/631)	80.73 (88/109)	0.480	0.035
	CATS2D ^b (1:1, RF) ^b	NCGC (test set)	67.48 (523/775)	76.25 (488/640)	25.93 (35/135)	0.445	0.019

^aResults from this study. ^bResults from Guha and Schürer.¹⁰

A common result among all of the oversampling models is their similar and improved ability to classify test set compounds. The best test set prediction for cytotoxic compounds (active compound predictive ability is measured with *sensitivity*) based on models created from the complete AID 426 data set was 65.0% (noNP+MOE descriptors), while the MOE descriptor pool SVM model was the least capable with a *sensitivity* value of 28.4% (see Tables 2–4). Classification models constructed from 4D-FPs using the oversampled data sets for the training sets constructed of Jurkat specific compounds and end points were able to indicate if a compound was not cytotoxic at a high degree, while the models constructed from the other descriptor pools performed poorly. The best *specificity* results for the Jurkat subset SVM models were obtained from the model created with the 4D-FPs descriptor pool and an oversampling ratio of 2:3 that was able to correctly identify 91.9% of the inactive compounds. The Jurkat specific training set SVM model constructed from MOE descriptors with an oversampling ratio of 1:1 returned a *specificity* of 56.8% compared to 48.7% for the original Jurkat specific MOE-SVM models, while the noNP+MOE SVM model with a 1:1 oversampling had a *specificity* of 43.2% compared to 35.1% for the original Jurkat specific noNP+MOE-SVM models. All of these improvements with respect to the ability of the models to reproduce the training set did not translate into equal success for the test set with *specificity* values ranging from 30.3% to 77.1% (Table 5) with all but one, the complete AID 426 data set SVM model constructed from the MOE descriptor pool, *specificity* value for the test set being less than 70%.

When using oversampling, the models perform quite well in self-prediction, while the test set results are not ideal; therefore, combinations of different descriptor sets are applied hoping to determine the corresponding reasons. The six best overfitted models constructed from the three descriptor pools and various active:inactive ratios are provided in Table 5 along with their ability to classify the test set. While all the models can be classified as ‘Acceptable’ or better based on Cohen’s *kappa*, the models’ ability to reproduce the classification of the test set compounds is considered ‘poor’. This is likely due to clashes of molecular similarity – discussed below – between the active compounds in the training set and the inactive compounds in the test set.

To provide an equivalent comparison to the Guha and Schürer study,¹⁰ RF and SVM models were constructed using the 1300 compounds of the AID 426 data set (the complete data set) with an oversampling ratio of 1:1 to compare the best

descriptor set in this study, the complete 4D-FPs pool, to models constructed from CATS2D descriptors. The 4D-FPs SVM model significantly outperformed the 4D-FP RF model based on training set *G-mean* and Cohen’s *kappa* values; 0.907 and 0.824 for the SVM model compared to 0.700 and 0.488 for the RF model, respectively. The ability to predict the class for the test set was also markedly better for the SVM model compared to the RF model with a *G-mean* score of 0.523 compared to 0.353 and a *kappa* value of 0.032 compared to −0.003. The SVM model for the training set can be considered ‘highly predictive’ based on the Cohen’s *kappa* criteria, while the RF model is considered ‘average.’ The test set abilities of these models is not as impressive, and once again the SVM model is considered superior to the RF model but should not be considered outstanding based on the test set evaluation. Using the same model creation methodologies, SVM and RF and the CATS2D molecular descriptors models for the complete training set were constructed. In this scenario the RF model outperformed the SVM model based on *G-means* (0.782 versus 0.739) and Cohen’s *kappa* (0.600 versus 0.504) evaluation methods but not by the same performance gap seen for the SVM and RF models constructed from 4D-FPs. This indicates the 4D-FPs contains more detailed molecular information, and the SVM method of model creation is better able to extract and apply that information. It can also be viewed that – in this example – the RF method is better able to extract pertinent information from a less information dense set of molecular descriptors than the SVM method.

Comparing these results to the Guha and Schürer¹⁰ study is not straightforward due to the manner that their training set was constructed. As discussed above, the Guha and Schürer training set, with a 1:1 active:inactive ratio, was constructed by randomly selecting 20% of the active compounds for the test set and using the remaining 80% of the actives and an equal number of inactives compounds for the training set. Unfortunately they did not indicate which compounds were assigned to the training or test set making it difficult to perform a true one-to-one comparison. The Guha and Schürer¹⁰ CATS2D-RF model based on the Jurkat subset of the NCGC AID 426 data set (this study’s Jurkat subset of AID 426 training set) when tested against The Scripps Jurkat data set (the combined AID 364 and 464 data set) was not as robust as either of the SVM or RF models created for this study. Their model exhibited overfitting for the active compounds (similar to what was experienced here before oversampling) but did not perform as well as the models presented here with *G-means* and *kappa* values of 0.445 and 0.019, respectively. A summary of the

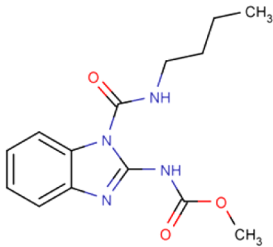
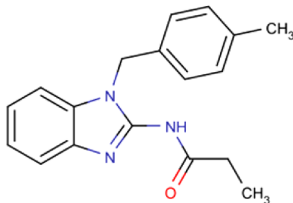
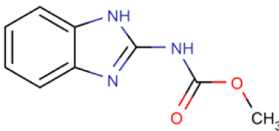
Active Training Set Compounds	Inactive Test Set Compound
 <p style="text-align: center;">CID 28780</p>	 <p style="text-align: center;">CID 780272</p>
 <p style="text-align: center;">CID 25429</p>	

Figure 3. The active compounds of the training set (CID 28780, 25429) and the inactive compound of the test set (CID 780272).

model evaluation methods for the 1:1 SVM and RF models constructed from the 4D-FPs and CATS2D descriptor pools are provided in Table 6.

Similarity of Training Data Set and Testing Data Set.

Often the ability of the model is impacted by the molecular similarity between the active and inactive compounds that compose the training set and the molecular similarity between the training set and the test set. In a data set of this size, it is expected that active compounds would be moderately to highly similar to each other, while this same trend would be seen for the inactive compounds regardless if they are part of the training or test set. When comparing the active compounds to the inactive compounds it is expected that there would be little molecular similarity between the two sets of compounds, and this would be the ideal situation for the construction of a classification model. Unfortunately, the real world of predictive modeling is far from ideal, and the experimental data that are provided is typically somewhat muddled; the compounds that are molecularly similar to each other yet possess opposing classifications cause problems for predictive models.

A combination of events – specifically the ability of the models – indicated that model performance may be limited by the data, which means that the features within the data may not have a good explanation for Jurkat cell viability. It is known that several different structural features may cause cytotoxicity, and thus molecular similarity analysis of active and inactive compounds was conducted. To explore the molecular similarity between active and inactive compounds within the training and test sets and between the training and test sets the PubChem fingerprints⁴⁸ were calculated and the molecular pairwise similarity was calculated resulting in Tanimoto coefficients. Figure 2 displays the scaled pairwise molecular similarity between the compounds in the training and test sets, specifically the similarity between the inactive compounds in the training and test sets (nontoxic; blue line) and the active compounds in the training and test sets (cytotoxic; red line). It is expected that the active compounds in the training and test sets would be similar to each other, while the inactive

compounds in the training and test sets would be similar to each other. Based on the scaled frequency of pairwise Tanimoto coefficients, both sets of compounds are weakly related to each other based on the PubChem fingerprints. A majority of active and inactive compounds in the training and test sets do not have a large number of molecularly similar compounds. It is expected that for each class of compounds (active and inactive) the majority of Tanimoto coefficient pairs would be greater than 0.60 not less than 0.40. The molecular comparison between active training set compounds and inactive test set compounds results in the expected molecular similarity (green line in Figure 2) where – albeit the same level of molecular similarity seen for active and inactive compounds between the training and test sets – there is little molecular similarity. The low amount of molecular similarity within the active and inactive compounds could be one of the reasons why the classification models are not able to definitively classify compounds within the test set.

Using the pairwise molecular similarity results to compare compounds, active and inactive compounds that are significantly similar, a Tanimoto coefficient greater than 0.40 was investigated. In Figure 3, CID 28780 and CID 25429 are active compounds in the training set, and CID 780272 is an inactive compound of the test set. The Tanimoto coefficient between the two active training set compounds is 0.90, the Tanimoto coefficient between CID 28780 (active compound in the training set) and CID 780272 (inactive compound from the test set) is 0.82, and the Tanimoto coefficient between CID 25429 (active compound in the training set) and CID 780272 is 0.78. The core of these three compounds is a benzimidazole, and the associated substituent groups are very similar. The structural similarity between these three compounds would most likely result in the inactive compound of the test set being predicted as ‘active’ due to its high molecular similarity to the two active compounds of the training set.

Another example of high molecular similarity is depicted in Figure 4 where three active training set compounds (CID 22586, CID 5930, and CID 5288209) are compared to two

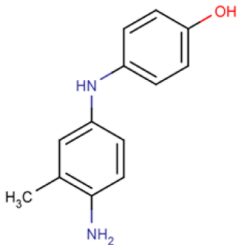
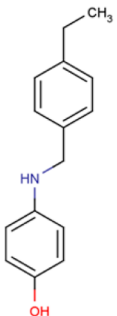
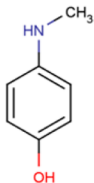
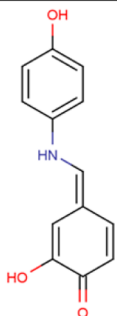
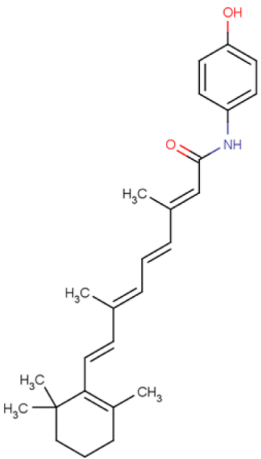
Active Training Set Compounds	Inactive Test Set Compound
 <p data-bbox="500 578 638 611">CID 22586</p>	 <p data-bbox="1028 615 1184 648">CID 833970</p>
 <p data-bbox="507 945 635 978">CID 5930</p>	 <p data-bbox="1021 1030 1196 1063">CID 5338656</p>
 <p data-bbox="483 1594 658 1627">CID 5288209</p>	

Figure 4. The active compounds of the training set (CID 22586, 5930, 5288209) and the inactive compound of the test set (CID 833970, 5338656).

inactive test set compounds (CID 833970 and CID 5338656). The pairwise Tanimoto coefficients for these five compounds are provided in Table 7; shaded values indicate pairwise Tanimoto coefficients for compounds within the same group (training or test set). The molecular similarity between the active and inactive compounds ranges from 0.81 to 0.86 indicating a very high level of similarity based on the PubChem fingerprints. All five compounds share a 4-aminophenol functional group, and hydrophobic molecular features are connected to the amine group for all the compounds. In this

situation, the three active compounds of the training set have a very similar molecular composition to the two inactive compounds in the test set making it difficult for the classification model(s) to correctly predict these test set compound as inactive. This is an example of a problem common in many classification models; molecularly similar compounds are part of opposite classes.

These close molecular similarities between active and inactive compounds cause problems for classification models, because it

Table 7. Pairwise Tanimoto Coefficients for Active Training Set Compounds and Inactive Test Set Compounds

	CID 5338656 ^a	CID 833970 ^a	CID 5288209	CID 5930
CID 22586	0.86	0.85	0.80	0.82
CID 5930	0.86	0.85	0.78	
CID 5288209	0.81	0.83		
CID 833970 ^a	0.81			

^aInactive test set compound.

becomes difficult for the underlying classification methods to discern the gross molecular features between the two classes.

Significant Molecular Features of the Optimal Cytotoxic Predictive Model. To extract the important molecular features from the optimal predictive model, we applied the linear SVM method to the data set. Unlike the radial-distribution function (RDF) SVM that was used, linear SVM can deduce the significant molecular descriptors by using weighting and linear combinations of the descriptors during the classification process. The top 20 significant molecular descriptors were identified from the best-fit model by using training the complete data set with an oversampling ratio of 2:3 (see Table S1). By once again using all data combinations – descriptor pools, oversampling ratios, and the original and subset data sets – an optimal linear SVM model emerged and was composed of 4D-FP descriptors. The *accuracy*, *sensitivity*, *specificity*, *G-means*, and *kappa* for this model applied to the 2:3 oversampled training set are 88.0%, 90.3%, 86.4%, 0.883 and 0.755, respectively, and 71.2%, 78.6%, 28.4%, 0.473, and 0.058, respectively, for the test set. The top 20 most important descriptors contained a variety of different pairwise pharmacophore interactions (see Table S1) with a mix of through-space distances separating the atoms. Half of the atomic distances are greater than 8 Å between the two atoms within the molecule. This indicates that the important molecular features, for cytotoxicity models derived from 4D-FP descriptor pools, are a combination of short- and long-range interactions, 3–8 Å and greater than 8 Å interactions, respectively.

The significant molecular descriptors have been classified according to their weighted contributions – based on the influence that the descriptor has in the classification of molecules with respect to cytotoxicity – to the linear SVM model. A positive weight (coefficient) for a molecular descriptor increases its contribution to the classification model and thus increases a compound's predicted cytotoxicity, while a negative weight decreases the descriptor's overall contribution to the classification model and correspondingly decreases a compound's predicted cytotoxicity. While the $\epsilon 4(\text{PP}, \text{HBD})$, $\epsilon 8(\text{NP}, \text{NP})$, $\epsilon 9(\text{NP}, \text{HS})$, $\epsilon 10(\text{NP}, \text{HS})$, $\epsilon 11(\text{NP}, \text{HS})$, and $\epsilon 4(\text{All}, \text{All})$ 4D-FPs are constructive descriptors, increasing the predicted end point, the other 14 descriptors reduce the predicted cytotoxicity.

To demonstrate the interpretation of a 4D-FP as a molecular descriptor, $\epsilon 2(\text{All}, \text{HBD})$, a significant descriptor with a negative weight, is illustrated for a cytotoxic and inactive compound, compounds with this molecular feature. The compound CID_33528, shown in Figure S1 in its 2D depiction and low energy 3D conformation, is a potent cytotoxic compound. Notice there are no hydrogen bond donors and thus no $\epsilon 2(\text{All}, \text{HBD})$ interactions within the compound. The opposite is seen in CID_8640, an inactive compound for cytotoxicity, and is displayed in Figure S2. It has several atomic

pairwise interactions between its hydrogen bond donor and the other atoms that satisfy the $\epsilon 2(\text{All}, \text{HBD})$ term. The dashed yellow lines represented the interaction between the atomic pairs. The other significant 4D-FP descriptors can be used in the same way, as guidelines, to design and refine potential drug candidates with respect to their cytotoxic activity.

CONCLUSION

In this study, SVM models constructed from a complete set of 4D-Fingerprints constructed a better classification model than the MOE (1D, 2D, and 21/2D) and the combined 4D-FP (sans nonpolar IPEs) and MOE descriptors when working with the complete and *imbalanced* AID 426 data set and the *imbalanced* Jurkat subset based on *G-means* and Cohen's *kappa* evaluation methods for the training set. These evaluation methods focus on the ability of the predictive models to correctly classify known active (cytotoxic) and inactive compounds. The ability for the 4D-FPs SVM models to correctly classify the divergent molecular similarity test set was also tops for the complete training set (AID 426) and ranked second only to the MOE SVM model, for the Jurkat subset, applied to the test set. The molecular information contained within the 4D-FPs when combined with a SVM provides enough key data to construct a classification model from an imbalanced data set that is comparable to oversampled SVM models.

Focusing the training set to contain only compounds with Jurkat cell specific data increases the performance of 4D-FPs and MOE SVM predictive models when applied to the test set. The advantage of the oversampling method is that it reduced the overfitting of the inactive compounds as demonstrated through the *specificity* evaluation values and increased the performance of self-prediction. The optimal ratio of oversampling is data set dependent, and a definitive active-to-inactive ratio could not be concluded from the presented study for cell cytotoxicity. With respect to the predictive modeling method and based on the training set, the SVM-trained models performed better than the random forest models,¹⁶ and compared to the results of the random forest study, the SVM method was able to decrease overfitting of inactive compounds.

It is possible that the predictive ability of the SVM and RF models is limited by the data set. Specifically, the compounds and end points contained within the data set may not provide a good explanation of the molecular features required for Jurkat cell viability. Cytotoxicity may be the result of the compounds of interest initiating several factors such as apoptosis,⁴⁹ damage to the cell membrane,⁵⁰ and DNA cleavage and enzyme inhibition.⁵¹ Additionally, different structural features account for different toxicities, and the molecular similarity of the compounds within the training and test sets have an impact on the predictive ability of the models when applied to a test set as shown in Figure 2. The protocols, methodologies, and results presented herein can be used as an outline for future imbalanced classification predictive models.

ASSOCIATED CONTENT

Supporting Information

Table S1 and Figures S1 and S2. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +886.2.3366.4888#529. Fax: +886.2.23628167. E-mail: yjtseng@csie.ntu.edu.tw.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The Taiwan National Science Council, Grants 101-2325-B-002-005- and 100-2325-B-002-004-, funded this work. Resources of the Laboratory of Computational Molecular Design and Detection, Department of Computer Science and Information Engineering, and Graduate Institute of Biomedical Engineering and Bioinformatics of National Taiwan University were used in performing these studies.

■ REFERENCES

- (1) Rusyn, I.; Daston, G. P. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ. Health Perspect.* **2010**, *118*, 1047–1050.
- (2) Shukla, S. J.; Huang, R.; Austin, C. P.; Xia, M. The future of toxicity testing: a focus on in vitro methods using a quantitative high-throughput screening platform. *Drug Discovery Today* **2010**, *15*, 997–1007.
- (3) Krewski, D.; Westphal, M.; Al-Zoughool, M.; Croteau, M. C.; Andersen, M. E. New directions in toxicity testing. *Annu. Rev. Public Health* **2011**, *32*, 161–178.
- (4) Sun, H.; Xia, M.; Austin, C. P.; Huang, R. Paradigm shift in toxicity testing and modeling. *AAPS J.* **2012**, *14*, 473–480.
- (5) Muster, W.; et al. Computational toxicology in drug development. *Drug Discovery Today* **2008**, *13*, 303–310.
- (6) Merlot, C. Computational toxicology—a tool for early safety evaluation. *Drug Discovery Today* **2010**, *15*, 16–22.
- (7) Modi, S.; Hughes, M.; Garrow, A.; White, A. The value of in silico chemistry in the safety assessment of chemicals in the consumer goods and pharmaceutical industries. *Drug Discovery Today* **2012**, *17*, 135–142.
- (8) Selassie, C. D.; et al. On the toxicity of phenols to fast growing cells. A QSAR model for a radical-based toxicity. *J. Chem. Soc., Perkin Trans.* **1999**, *2*, 2729–2733.
- (9) Garcia-Lorenzo, A.; et al. Cytotoxicity of selected imidazolium-derived ionic liquids in the human Caco-2 cell line. Sub-structural toxicological interpretation through a QSAR study. *Green Chem.* **2008**, *10*, 508–516.
- (10) Guha, R.; Schurer, S. C. Utilizing high throughput screening data for predictive toxicology models: protocols and application to MLSCN assays. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 367–384.
- (11) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct.: THEOCHEM* **2003**, *622*, 39–51.
- (12) Xia, M.; et al. Compound cytotoxicity profiling using quantitative high-throughput screening. *Environ. Health Perspect.* **2008**, *116*, 284–291.
- (13) Li, Q.; Wang, Y.; Bryant, S. H. A novel method for mining highly imbalanced high-throughput screening data in PubChem. *Bioinformatics* **2009**, *25*, 3310–3316.
- (14) Estabrooks, A.; Jo, T. H.; Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.* **2004**, *20*, 18–36.
- (15) Ertekin, S.; Huang, J.; Bottou, L.; Giles, L. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, Lisbon, Portugal, 2007; pp 127–136.
- (16) Gazzah, S.; Amara, N. E. B. New Oversampling Approaches Based on Polynomial Fitting for Imbalanced Data Sets. In *Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems*, IEEE Computer Society, 2008; pp 677–684.
- (17) Tang, Y.; Zhang, Y.-Q.; Chawla, N. V.; Krasser, S. SVMs modeling for highly imbalanced classification. *Trans. Sys. Man Cyber. Part B* **2009**, *39*, 281–288.
- (18) Sun, A. X.; Lim, E. P.; Liu, Y. On strategies for imbalanced text classification using SVM: a comparative study. *Decis. Support Syst.* **2009**, *48*, 191–201.
- (19) Xie, J. G.; Qiu, Z. D. The effect of imbalanced data sets on LDA: a theoretical and empirical analysis. *Pattern Recognit.* **2007**, *40*, 557–562.
- (20) Haykin, S. S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall: Englewood Cliffs, NJ, 1998.
- (21) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (22) Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106.
- (23) Judson, R.; Elloumi, F.; Setzer, R. W.; Li, Z.; Shah, I. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. *BMC Bioinf.* **2008**, *9*, 241.
- (24) Sharma, A.; Kumar, R.; Varadwaj, P. K.; Ahmad, A.; Ashraf, G. M. A comparative study of support vector machine, artificial neural network and bayesian classifier for mutagenicity prediction. *Interdiscip. Sci., Comput. Life Sci.* **2011**, *3*, 232–239.
- (25) 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.
- (26) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz'min, V. E. Application of random forest approach to QSAR prediction of aquatic toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.
- (27) Zhao, C. Y.; et al. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* **2006**, *217*, 105–119.
- (28) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D. Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction. *Chem. Rev.* **2010**, *110*, 5714–5789.
- (29) von Korff, M.; Sander, T. Toxicity-indicating structural patterns. *J. Chem. Inf. Model.* **2006**, *46*, 536–544.
- (30) Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.* **2005**, *48*, 312–320.
- (31) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630–1638.
- (32) Tseng, Y. J.; Hopfinger, A. J.; Esposito, E. X. The great descriptor melting pot: mixing descriptors for the common good of QSAR models. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 39–43.
- (33) Renner, S.; Fechner, U.; Schneider, G. *Pharmacophores and Pharmacophore Searches*; Wiley-VCH: Weinheim, Germany, 2006.
- (34) Su, B. H.; Tu, Y. S.; Esposito, E. X.; Tseng, Y. J. Predictive toxicology modeling: protocols for exploring hERG classification and Tetrahymena pyriformis end point predictions. *J. Chem. Inf. Model.* **2012**, *52*, 1660–1673.
- (35) Shen, M. Y.; Su, B. H.; Esposito, E. X.; Hopfinger, A. J.; Tseng, Y. J. A comprehensive support vector machine binary hERG classification model based on extensive but biased end point hERG data sets. *Chem. Res. Toxicol.* **2011**, *24*, 934–949.
- (36) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526–1539.
- (37) Molecular Operating Environment (MOE). Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2011.
- (38) MOE (Molecular Operating Environment). Chemical Computing Group, Inc.: Montreal, Canada, 2008.
- (39) Lin, A. QuaSAR-Descriptor. November 2012. Available from <http://www.chemcomp.com/journal/descr.htm>.

- (40) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-hopping" by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem., Int. Ed. Engl.* **1999**, 38, 2894–2896.
- (41) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
- (42) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: 2000.
- (43) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, Vol. 2, 18–22.
- (44) Team, R. D. C. R: *A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2011.
- (45) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, 2, 1–27.
- (46) Cohen, J. A. Coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, 20, 37–46.
- (47) Rucker, C.; Rucker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, 47, 2345–2357.
- (48) PubChem Substructure Fingerprint, 2009.
- (49) Schwartzman, R. A.; Cidlowski, J. A. Apoptosis: the biochemistry and molecular biology of programmed cell death. *Endocr. Rev.* **1993**, 14, 133–151.
- (50) Groot, R. D.; Rabone, K. L. Mesoscopic simulation of cell membrane damage, morphology change and rupture by nonionic surfactants. *Biophys. J.* **2001**, 81, 725–736.
- (51) Nakamura, E.; Tokuyama, H.; Yamago, S.; Shiraki, T.; Sugiura, Y. Biological activity of water-soluble fullerenes. Structural dependence of DNA cleavage, cytotoxicity, and enzyme inhibitory activities including HIV-protease inhibition. *Bull. Chem. Soc. Jpn.* **1996**, 69, 2143–2151.