Practical Outcomes of Applying Ensemble Machine Learning Classifiers to High-Throughput Screening (HTS) Data Analysis and Screening

Kirk Simmons,*,† John Kinney,‡ Aaron Owens,§ Daniel A. Kleier,|| Karen Bloch,§ Dave Argentar,⊥ Alicia Walsh,§ and Ganesh Vaidyanathan#

Simmons Consulting, 52 Windybush Way, Titusville, New Jersey 08560, DuPont Stine Haskell Research Laboratories, 1090 Elkton Road, Newark, Delaware 19711, DuPont Engineering Research and Technology, POB 80249, Wilmington, Delaware 19880, Drexel University, 3141 Chestnut Street, Philadelphia, Pennsylvania 19104, Sun Edge, LLC, 147 Tuckahoe Lane, Bear, Delaware 19701, and Quantum Leap Innovations, 3 Innovation Way, Suite 100, Newark, Delaware 19711

Received May 14, 2008

Over the years numerous papers have presented the effectiveness of various machine learning methods in analyzing drug discovery biological screening data. The predictive performance of models developed using these methods has traditionally been evaluated by assessing performance of the developed models against a portion of the data randomly selected for holdout. It has been our experience that such assessments, while widely practiced, result in an optimistic assessment. This paper describes the development of a series of ensemble-based decision tree models, shares our experience at various stages in the model development process, and presents the impact of such models when they are applied to vendor offerings and the forecasted compounds are acquired and screened in the relevant assays. We have seen that well developed models can significantly increase the hit-rates observed in HTS campaigns.

INTRODUCTION

High-Throughput Screening (HTS) is an integral component of many pharmaceutical, animal health, and crop-protection discovery operations. However, the implementation of HTS can differ significantly between these industries in that crop-protection screening libraries can be evaluated directly against the pest species of interest at very early stages in the discovery process. This is often accomplished using 96-well plate-based assays evaluating the affect of a compound directly on the viability of the intact insect, weed, or fungal pathogen. A potentially significant value of in-vivo HTS is that completely unexpected or novel mechanisms of action may be discovered as a result of a screening campaign. However, HTS requires testing large numbers of compounds in order to produce the required number of hit and lead compounds for a discovery effort, a common theme in all of these industries. Therefore, historical HTS data are a rich starting point from which to learn rules for classifying candidate compounds into activity categories with the ultimate goal of improving HTS by applying the resulting classifiers to preselect compounds for screening in the next HTS campaign.

Over the last few years ensemble modeling has become widely accepted as a significant advance in predictive modeling and is being applied to a broad range of machine learning methods.[1−18] Significant improvements in predictive performance can be realized when multiple models are combined into one classifier by averaging or polling of the results of the individual models. It is important that the individual models are diverse in their predictive performance with regards to the classes being forecast so that the weakness of one model is offset by the strength of another model. Consistent with these observations, our earlier work[19] demonstrated that significantly stronger predictive models could be achieved by developing a series of recursive partitioning classifiers from a data set and averaging the predictions from each classifier to form the final classification. Ensemble modeling has been applied to predict various endpoints such as mutagenicity,[1] docking scores,[2] aqueous solubility,[3,16] drug likeness,[8] aquatic toxicity,[12] and numerous in-vitro biological activities.[4−7,9−11,13−15,17,18] These published modeling efforts have utilized a number of different chemical descriptor types and various machine learning methods, such as Decision Trees, Nearest-Neighbor Analyses, Support Vector Machines, Artificial Neural Networks, Linear Discriminant Analyses, and Genetic Algorithms. Essentially all of the published models were developed on small data sets (typically a few hundred compounds) with the largest reported data set being 15,000 compounds.[13] In every example model development was performed by first randomly splitting the data set into training and testing portions, and the reported model performance was assessed solely on the prediction quality for the testing set. In only one reference was the resulting classifier used to actually score externally available chemistry and the success observed for that scoring exercise reported.[13]

During the course of expanding the ensemble models from our earlier studies[19] we have noted that the traditional means of evaluating classifier performance, namely randomly split-

* Corresponding author e-mail: KirkASimmons@gmail.com.
† Simmons Consulting.
‡ DuPont Stine Haskell Research Laboratories.
§ DuPont Engineering Research and Technology.
|| Drexel University.
⊥ Sun Edge, LLC.
Quantum Leap Innovations.

10.1021/ci800164u CCC: $40.75
Published on Web 11/05/2008

Table 1. Bioassay Applied Doses and Activity Criteria

domain	primary	advanced	active designation
fungicide	17 μM	≤200 ppm	≥80% control
herbicide	1000 g/Ha	≤500 g/Ha	≥80% control
insecticide	1050 μM	≤250 ppm	≥80% control

Table 2. Chemical Descriptor Types Used To Train Classifiers

descriptor	source	size
ISIS 960 fingerprint	Cheshire	960
ISIS 166 fingerprint	Cheshire	166
2D AtomPairs	in-house code	825
3D AtomPairs	in-house code	825
AlogP Atom Types	Cerius2	120
MolconnZ	MolconnZ	221
Diverse Solutions BCUTS	Diverse Solutions	8
Cerius2 combichem	Cerius2	44

ting the data set into training and testing subsets, can lead to an optimistic assessment of classifier performance. This occurs because corporate compound collections are often populated with families of chemistry that arise during historical synthetic optimization efforts. Random splitting of such collections into training and testing subsets results in chemical structures in the testing set that can be closely related to those in the training set. We encountered this phenomenon during the project outlined in this paper, and we will present and discuss strategies designed to minimize this historical bias.

This paper also provides an overview of this multiyear project in which classifiers were researched and developed using historical in-house in-vivo HTS data. The primary goal in developing predictive models from historical HTS data was to use them to identify, acquire, and screen structures more likely to be active in our various agrochemical assays. Specifically, our desired outcome was to achieve at least a 500% increase in HTS hit-rate. During the course of this project several machine learning methods were explored as well as a number of different chemical descriptor types commonly used in the cheminformatics community.[20,21] As optimal classifiers were developed throughout this project they were used to score current offerings from chemistry vendors, and those structures scoring well were purchased and evaluated in our biological assays as the basis for proof of concept experiments. The biological activity observed for these chemical purchases was used to evaluate the predictive performance of the corresponding classifiers, a significantly more rigorous and practical assessment of their performance than the simple random split method. Key lessons learned from these experiments were then incorporated into the development of the next generation of classifiers.

METHODS

Biology. In the agrochemical industry, early primary biological screens are often based upon a few key species that are especially sensitive to chemical control. Screening first on these sensitive species essentially eliminates inactive or weakly active compounds from advancing into more labor intensive biological assays which are comprised of a broad spectrum of species which represent major market areas for crop protection products. In order to develop activity classifiers with predictive capabilities that fit those market areas, we utilized as many of the species endpoints as possible for model development. For each of the crop protection domains in Table 1, we trained two classifiers, each focused on a different but relevant major market area. The market focus was control of ascomycete and oomycete pathogens in fungicides, control of monocot (grass) and dicot (broadleaf) weeds in herbicides and control of homoptera (sucking) and lepidoptera (chewing) insects in insecticides. Compounds were initially screened in the primary bioassays in each domain (fungicide, herbicide, and insecticide) at the concentrations listed in Table 1. Compounds that were active in the primary bioassays were advanced into higher level foliar-based screens to further explore and define the biological activity that was initially observed. For our studies compounds marked as "inactive" produced less than 10% inhibition or mortality in the corresponding primary bioassay, and compounds marked as "active" were confirmed active against any relevant specie within a given market area in advanced foliar-based screens at application rates listed in Table 1. The HTS results were subjected to a triage in which duplicate structures and those of unknown or questionable structure were removed.

It should be emphasized that the endpoint in each of the agrochemical assays represents death of the respective organism, a process than can be induced by inhibiting or interfering with a variety of differing biochemical mechanisms. As such, in-vivo HTS data become especially challenging to successfully train useful classifiers given the diversity of potential biochemical targets that are represented by the biologically active examples in the data set. Typically mechanisms of action studies are not initiated until much later in a project, so the mechanisms of action are unknown for the biologically active compounds on which the classifiers are being trained.

Chemistry. All of the structures in the HTS data sets were preprocessed using Pipeline Pilot[22] to standardize structural representations (e.g. -[N+](=O)[O-] => -N(=O)=O) and to remove counter ions for structures represented as salts. The standardized clean structures were output as MDL connection tables.[23] As needed, 2D structures were converted to 3D using Concord, and the structures were output as Sybyl mol2 connection tables.[24] All calculated chemical descriptors were generated from the standardized clean chemical structures.

Chemical Descriptors. A large number of widely available descriptors can be computed from a connection table. We selected a subset of them for consideration in model development. The chemical descriptor types that were evaluated in model development are listed in Table 2. Our choices focused on exploring a diversity of descriptor types that could be computed efficiently since we envisioned ultimately applying our final classifiers to score millions of chemical structures each and every year.

The ISIS fingerprint descriptors were generated using the Cheshire software package (version 3.0).[25] We computed both the "public" keys (166) whose definitions are available as well as the full private 960 key set.

The atom pair descriptors[26] which represent all of the <atom type - distance - atom type> combinations in the structure were computed from the Sybyl mol2 connection table. Our implementation of the atom-pair descriptors, coded

in the C programming language,[27] collapses the ca. 25 Sybyl mol2 atom types into 10 atom types and maps the Cartesian distances between them into 1-Angstrom distance bins (distances ranging from 1 to 15 Angstroms) for the 3D variant. The 2D atom pair variant maps the bond spacing between the atom types (distances ranging from 1 to 15 bonds). The final descriptors consisted of 825 numerical values representing 55 possible atom-type pairs mapped onto 15 distance ranges. Concord was used to generate a reasonable 3D-geometry for the structure before computing the 3D atom pair descriptors.

The AlogP atom type descriptors (counts of the occurrences of each atom type used in calculating AlogP) were computed using the Cerius[2] software package[28] (version ccO) and exported as a tab-delimited text file. Additionally, the software allowed for calculating the "combichem" subset of descriptors which consisted of 44 descriptors often used as the basis for diversity analysis of combinatorial chemistry libraries within the software. The interested reader should reference Accelrys documentation.[29]

The MolconnZ software[30] (version 3.5) was used to convert structures into a set of 221 descriptors representing molecular connectivity, shape, and charge distribution in the underlying structure. The interested reader should reference the MolconnZ manual[31] for a more in depth discussion.

The BCUT descriptors[32] were computed using DiverseSolutions[33] software (version 4). From the myriad of possible BCUT descriptors an 8-member subset ('Const_0.250_H', 'Gast_0.031_H', 'Gast_0.219_L', 'Ha_0.500_H', 'Hd_0.562_H', 'Pol_0.208_L', 'Pol_1.250_H', 'MolVol') was selected as providing a significant dispersion of the chemical structures in the DuPont corporate chemistry collection as well as representing features that made chemical sense with regards to small molecule-protein interactions.

Model Development/Assessment. Model development using Neural Networks analysis[34] and InfoEvolve[35] were conducted using software proprietary to DuPont. Oblique decision tree-based analyses were completed using OC1[36] (Oblique Classifier). EnsembleFIRM modeling was implemented using a proprietary routine coded in FORTRAN interfacing to FIRM[37] (formal inference-based recursive modeling).

Initially data sets were randomly divided into training and testing subsets for the purposes of model development. Models and the associated machine learning techniques were evaluated on the performance of predicting the testing set, following model development on the training set. Predictive performance was assessed by developing ROC curves for the predictions of the testing data sets by sorting the test sets descending on the prediction probabilities which run from zero to one. These ordered lists were then traversed from the top most ranked compound to the least, and the number of actives retrieved were counted as a function of list depth. When classifiers were also used to score externally available chemical structures for compound acquisition campaigns, the hit-rate from biological evaluation of the acquired chemistry was compared to our historical HTS baseline as an additional factor to evaluate classifier performances.

Neural Network Models. Neural network models were developed stepwise by first randomly selecting 10% of the training data set as a validation set. During training of the neural network the validation set was used to decide the optimal number of hidden units and to determine an error limit for terminating training. Using these stopping conditions the neural network was then retrained using the entire training data set.

Oblique Decision Tree Models. Oblique classifiers search linear combinations of descriptors for splitting the data set and so offer the advantage of potentially simpler yet more accurate decision trees than parallel axis methods such as CART, C4.5, and FIRM. However, they tend to be significantly more computationally intensive. The OC1 procedure was used to induce a set of oblique decision trees using 10-fold cross-validation during training to select the final tree.

EnsembleFIRM Decision Tree Models. FIRM treats continuous descriptors by binning them initially into 10 approximately equally populated intervals. The algorithm analyzes the data set deciding which descriptor and which split cardinality is optimal at each node in the tree by systematically and recursively collapsing adjacent descriptor bins when they are not significantly different with respect to the predicted class. FIRM uses a chi-squared test (if class endpoint is categorical) or a T-test (if the class endpoint is continuous) to decide the cardinality of the optimal split. The user can specify the minimum significance required during these tests as well as the overall significance of the variable choice. In all runs the minimum significances were set to a p-tail $= 0.05$. Statistical test results are corrected using the Bonferroni adjustment[38] to reflect the number of descriptors considered at each split. In addition FIRM allows the user to specify the smallest size node that is allowed to be further split. Tree induction ceases when statistical tests for descriptor selection exceed statistical thresholds, when the node size is below the splitting threshold specified by the user or when the node is too homogenous to split further.

Our approach to ensemble decision tree modeling involved wrapping FIRM with a FORTRAN procedure which managed data set sampling and model averaging while basically invoking FIRM for decision tree induction. Our sampling strategy consistently included all of the actives from the training set and appended a number of inactive structures randomly selected from the training set. Across resampling iterations the active structures are always reused while the inactive structures are not. For example, an HTS training set containing 1% actives would produce 100 smaller training data sets consisting of a 1:1 inactive/active ratio, when resampled using our strategy. The optimal ratio of inactive to active structures in the training subsets was empirically determined. FIRM models were developed for each of these smaller data sets, and the final classifier predictions were derived from averaging the individual classifier predictions across the models.

InfoEvolve Models. The InfoEvolve method uses the same sampling strategy described for the ensemble decision tree method. InfoEvolve first analyzes each of the data subsets separately in order to identify the most information rich combinations of possible input variables for that data subset. All possible combinations of inputs are represented in a gene pool in which each bit in the gene represents each possible input. The global information content for the data set is used to drive the genetic algorithm based optimization. Once the evolution is completed there exists a pool of genes that represents optimal combinations of inputs which are

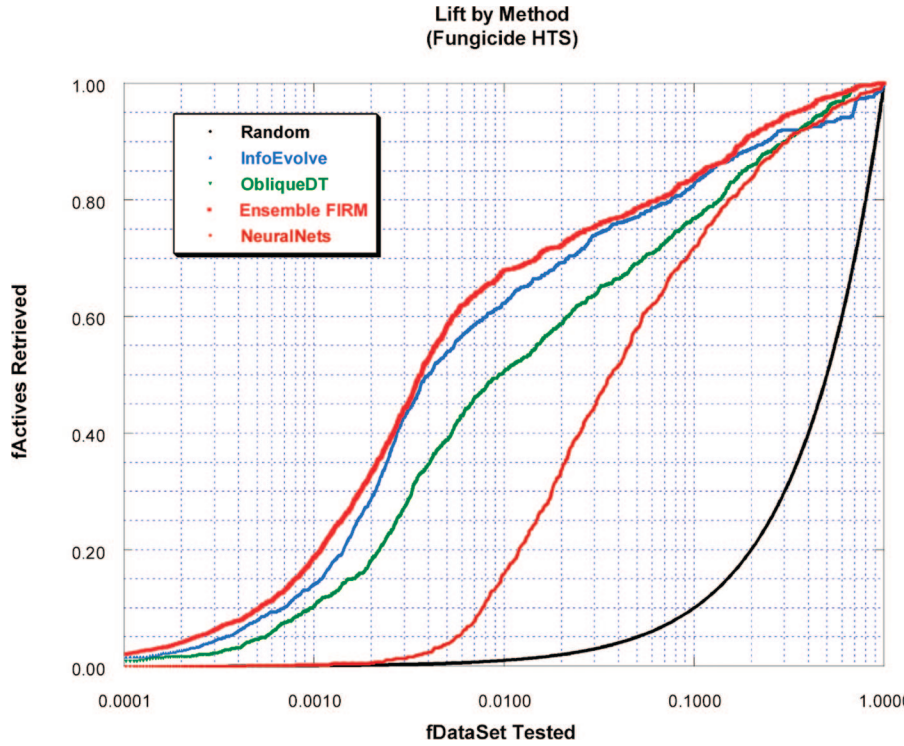

Figure 1. ROC curves for several early fungicide classifiers using MolconnZ descriptors. In order to better contrast performance of the activity classifiers at the early stages of retrieval, the fraction of actives retrieved is plotted against the fraction of the data set tested on a log scale rather than the more typical linear scale. Note that the "random" curve corresponds to retrieving x% of the actives when x% of the data set has been tested.

most information rich for that data subset. Analysis of the frequency of occurrence of the inputs across this gene pool is used to select the final inputs considered in modeling each data subset. Once the optimal inputs are selected a second genetic algorithm-based optimization step occurs in which the root mean square (RMS) error between predicted and actual is minimized for the training data set, while monitoring a small validation set RMS error in order to avoid overfitting. The single models are then combined into a single classifier by polling. The interested reader is referred to the work of Vaidyanathan[35] for additional discussion.

RESULTS

During the course of this project, we assembled HTS data sets for each of two major market areas within the agrochemical domains of herbicides, fungicides, and insecticides. Data mining models were developed for each of these HTS data sets using eight different chemical descriptor types. Our modeling efforts initially focused on a fungicide HTS data set since it was the first to be assembled and used the selected machine learning routines to develop ensemble-based classifiers starting with the 3D Atom-Pair and MolconnZ descriptors. While models were being developed in the fungicide area, HTS data sets were assembled for the other biological areas. Each of the HTS data sets was randomly divided into training (67%) and testing (33%) subsets. The final classifiers developed on the training portion were then applied to the testing subset, and the forecasted probabilities were used to construct ROC curves in order to evaluate and compare their performance. We were quite pleased with our initial fungicide results[19] in which all four machine learning methods produced classifiers with lifts significantly better than random (Figure 1) For example, only the top 6% of the EnsembleFIRM ranked compounds need to be retrieved in order to obtain over 80% of the active compounds. This corresponds to an enrichment factor of thirteen. The other classifiers derived from MolconnZ descriptors were InfoEvolve (enrichment=11), Oblique Classifiers (enrichment=6), and Neural Networks (enrichment=5).

During further model development it quickly became apparent that oblique decision tree classifiers, while offering the possibility of simpler tree models, were too compute intensive for on-going development and deployment. With some of the larger HTS data sets derived from the atom-pair descriptors we were encountering compute times for training measured in cpu-weeks. Neural networks can have problems converging with wide data sets as well (i.e., data sets with a large number of descriptors relative to the number of data points), and our implementation is no exception. We encountered problems developing Neural Network models with the atom-pair descriptors (825 inputs). We have prior experience in engineering projects where InfoEvolve could be successfully used to perform variable selection prior to performing Neural Network analysis (data not shown), but since our previous studies[19] indicated that models derived using Oblique Decision Tree and Neural Networks were generally weaker we narrowed the field for further study to InfoEvolve and EnsembleFIRM.

As work continued at building and evaluating classifiers from a variety of chemical descriptor types (Table 2), several questions arose.

1. Is the rank order of actives the same across different descriptor types using the same machine learning method?

2. Is the rank order of actives the same across different machine learning methods using the same descriptor type?

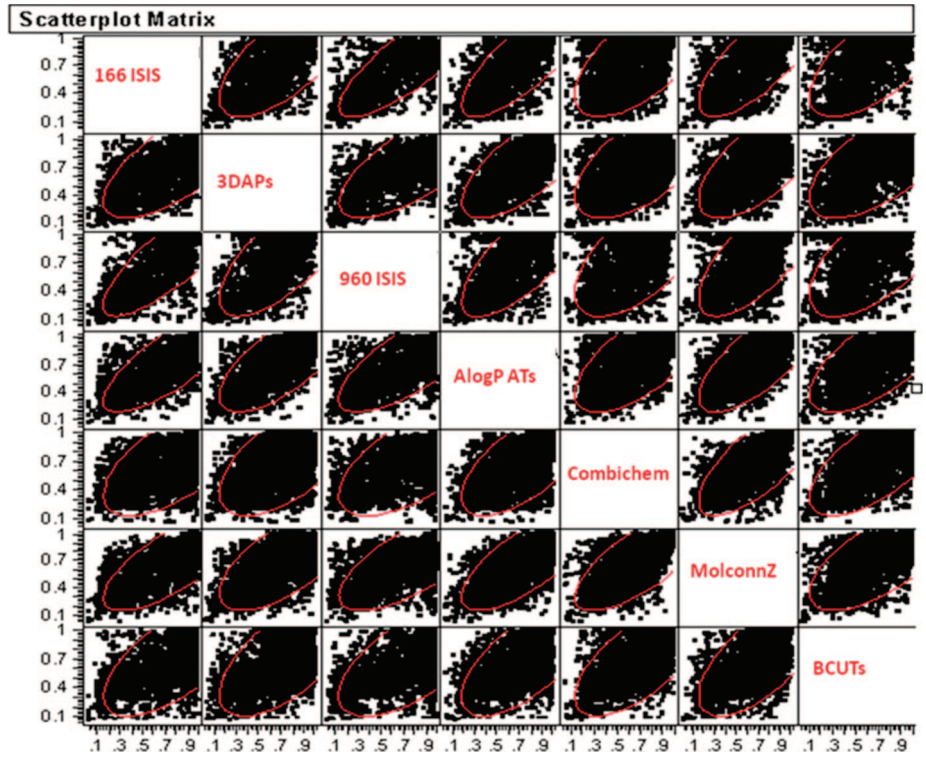

Figure 2. Pair wise plots of prediction probabilities for HTS actives from EnsembleFIRM classifiers built using different descriptor types.

Table 3. Lift vs Descriptor-Herbicide HTS Data and EnsembleFIRM Classifiers[a]

	166 sBITs	3DAPs	960 sBITs	AlogPATs	C^2 Combichem	MolconnZ	BCUTs	2DAPs
166 sBITs	**16.0**							
3DAPs	24.4	**18.2**						
960 sBITs	25.4	30.3	**24.2**					
AlogPATs	24.9	27.2	30.0	**17.2**				
C^2CombiChem	15.9	15.5	19.4	17.5	**5.4**			
MolconnZ	22.4	26.8	29.3	27.4	14.0	**14.9**		
BCUTs	17.7	18.0	21.5	19.5	8.9	16.8	**6.2**	
2DAPs	27.4	28.5	32.1	29.8	22.6	30.1	24.1	**23.2**

[a] Lifts were calculated at 80% recovery of actives.

In order to address the first question we extracted the known actives from the testing set of one of the HTS data sets and plotted the classifier probabilities from EnsembleFIRM classifiers built using each of the descriptor sets. Two way plots of the prediction probabilities (Figure 2) reveal that for a large proportion of the active compounds, the rank ordering is quite similar across descriptors with the bulk of the compounds falling on or near the diagonal. The red ellipses represent the 95% confidence intervals for the plots, and the scales of the individual axes run from zero to one. However, there were also a significant number of occurrences in which an active compound was well predicted by a classifier developed on one descriptor type (i.e. classifier probability approaching one) but was poorly predicted by classifiers built using another descriptor type (i.e. classifier probability approaching zero). Examples of these cases reside in the upper left or lower right portions of the individual plots. This behavior is seen across all descriptor combinations and revealed that each descriptor type is capturing significantly different information around a chemical structure. These findings suggested that averaging the prediction probabilities from several classifiers built from different descriptor types might lead to stronger models. This proved to be the case, and a significant enhancement was seen by averaging the results from two classifiers (Table 3) The diagonal elements in Table 3 represent the lift observed at 80% recall for EnsembleFIRM classifiers built from one of the herbicide HTS data sets. The off diagonal elements represent the observed lift when the prediction probabilities for the two classifiers are first averaged and the lift recomputed using the averaged probabilities. In a majority of the cases the averaged predictions afforded lifts that were numerically larger than the lift seen with the best of either classifier. In a smaller number of cases the enhancements in lift approached 125−160% of that seen with the best of either classifier. Although we have no definitive explanation, we suggest this synergy is due to differences in the structural information coded by each descriptor type. Enhanced lifts are observed even when weaker classifiers are averaged (e.g. BCUTs lift = 6.2 and Cerius^2 Combichem lift = 5.4, averaged lift = 8.9). Clearly there is a wide ranging performance for EnsembleFIRM classifiers built from different descriptors with lifts ranging from 5.4 to 24.2, depending upon the descriptor type used in building the models. The dependence of classifier performance on the

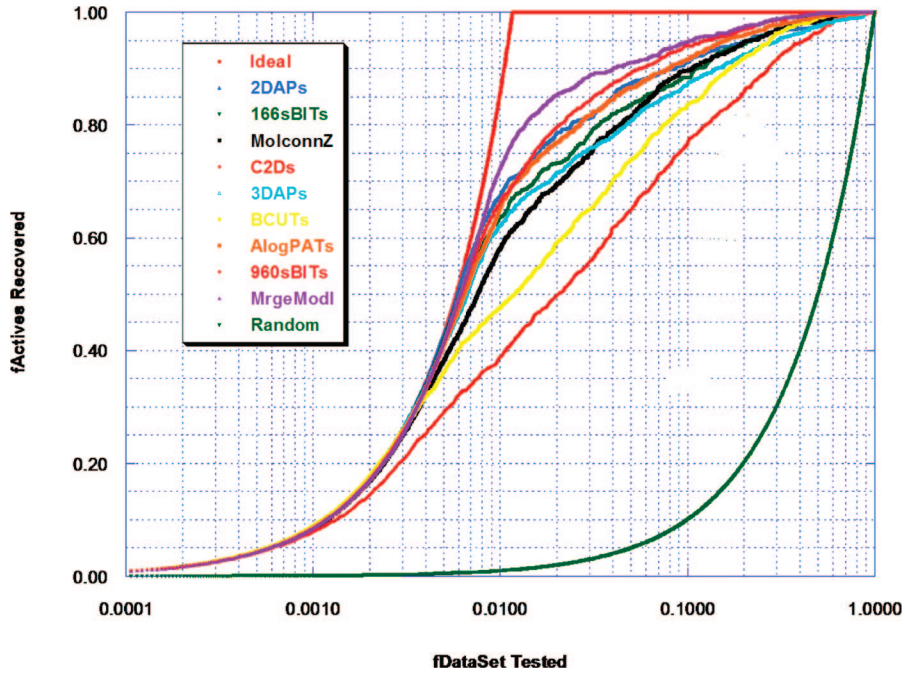

Figure 3. Comparative performance of EnsembleFIRM insecticide classifiers trained using different descriptor types. The fraction of actives retrieved is plotted against the fraction of the data set tested on a log scale. Note that the "random" curve corresponds to retrieving x% of the actives when x% of the data set has been tested.

chosen descriptor was observed across all machine learning methods and all HTS data sets (data not shown).

Lift enhancements of 125−160% over that of a single classifier by combining two classifiers are significant because they effectively translate into a lower false positive error rate for the same level of recall of the actives for the combined classifiers, a very important consideration when classifiers are being used to forecast several million structures. Figure 3 shows the ROC curve for each EnsembleFIRM classifier that was trained on an insecticide data set using each of the eight descriptor types. The purple curve labeled "MrgeModl" is the resulting classifier performance when the classifier probabilities of two of the individual classifiers (2D Atom Pairs and 960 ISIS fingerprints) were averaged and the ROC curve recomputed using the averaged results. Visual inspection reveals that the resulting classifier is capable of retrieving 80% of the active compounds in the testing set in the top 1.4% of the ranked list, compared to 2% to 14% for the single classifiers.

One might ask why not simply build a classifier using the combined set of descriptor types? We considered this scenario but rejected it in favor of developing activity classifiers using each descriptor type and then averaging the classification results. Given that many of these descriptor types code similar information from a chemical structure and thus are statistically correlated there is a strong likelihood that the resulting classifier would have selected, as a worst case scenario, a few descriptors from each of the possible types. Methods development and ultimate deployment of the final classifiers into a compound acquisition strategy would then have required computing all of the descriptor types for all of the structures in order to apply the classifiers. Quite useful results were obtained by averaging predictions using only two descriptor types thus avoiding substantial overhead in licensing and maintaining eight separate descriptor software packages.

Table 4. Lift vs Method-Fungicide HTS Data and MolconnZ Descriptors

	IE	OC	NN	EF
InfoEvolve	**10.6**			
OC1	9.7	**5.8**		
NeuralNets	9.8	63	**5.2**	
EnsembleFIRM	15.1	8.8	8.9	**12.6**

In order to address the second question, we have also explored the case in which the classifiers were built using different machine learning methods but using the same descriptor type (Table 4). Lift enhancement is also observed when averaging the classifier probabilities across machine learning methods. However, the magnitude of the enhancements is not as strong as when averaging classifier probabilities across descriptor types. For the most part we see that the averaged performance across classifiers lies somewhere between the performance of the two individual classifiers, revealing no synergy, the exception being the blended models between InfoEvolve and EnsembleFIRM. This same behavior was seen in modeling all of the HTS domains (data not shown) as well as using other descriptors besides MolconnZ and suggested that the machine learning methods are all reasonably capable of extracting useful information from a data set.

As models were developed throughout this project, they were used to classify externally available chemistry offerings to identify compounds for purchase as part of a series of proof of concept experiments. Compounds that were ultimately acquired were still subject to our usual diversity assessment in which their structures were compared to our corporate chemistry collection to avoid those which were too similar to heavily worked historical areas or contained structural elements deemed unworthy, i.e. inappropriate functional groups, unstable substructures, poor physical properties, and other considerations. Compounds were tagged

Table 5. Proof of Concept Experiments

		compounds					biology results	
strategy	year	analyzed	ordered	received	screened	active	hit-rate	improvement
HTS benchmark					212,730	421	0.20%	
data mining	2001	533,000	12,287	12,168	12,168	67	0.55%	**2.8**
	2002	2,370,675	9,432	3,733	3,733	66	1.77%	**8.9**
	2003	1,959,508	16,573	9,163	9,163	92	1.00%	**5.1**
	2004	2,323,869	8,512	5,330	5,330	66	1.24%	**6.3**
totals		7,187,052	46,804	30,394	30,394	291		

Table 6. Cluster Summary for HTS Data Set Actives at 90% Similarity Using the ISIS 960 Fingerprints

	fungicide		herbicide		insecticide	
	ascomycete	oomycete	monocots	dicots	homopteran	lepidopteran
NumCompds	9176	4114	17,823	15,659	6155	7999
MaxClusterSize	336	214	1164	1358	376	618
AveClusterSize	3.2	3.8	11.3	13.3	4.1	8.7
NumClusters	4769	1853	4173	3265	2587	2262
NumSingletons	3498	1299	2485	1826	1598	1278

upon arrival from the vendors to track them through the screening process. Once all of the compounds were screened through the various domain bioassays, the screening results were analyzed in order to compute a hit-rate within the selected sets. The hit-rates were compared to the baseline hit-rate for our HTS process which was computed based on screening results several years just prior to these experiments. The proof of concept experiments was conducted over several years, and the results are summarized in Table 5.

During the course of these experiments a large number of classifiers were explored from a variety of machine learning methods based upon eight commonly encountered chemical descriptor types. In all years the application of data mining models to identify compounds with an enhanced likelihood of bioactivity proved to be successful. Even our earliest models, developed in 2001 and which represented only a portion of the agrochemical domains, afforded screening hit-rates that were nearly 300% over our HTS benchmark. Models developed later, which benefited from the key lessons learned in the project, were providing hit-rates that were 630% over benchmark. Ultimately the analysis of over 7 million chemical structures identified less than 47,000 that were deemed worthy of purchase. This illustrates why classifiers are required to possess a low false positive rate as well as a high true positive rate if they are to be useful. Classifiers developed later in the project were sufficiently discriminating that typically less than 1% of the structures being forecast would score above the required thresholds for predicted activity. Screening externally available chemistry compounds which were forecast as 'active' by our classifiers provided hit-rates in the range of 0.6−1.8% (280−890% over the HTS benchmark).

Our experiences revealed that external vendor compounds may only be available for a relatively short period of time following the initial offering, and so speed in model development and implementation become important factors in an effective compound acquisition strategy. For example, in 2002 only about 40% of the compounds that scored well in our models were still available when the order was placed. This reinforces the need to be able to build, retrain, and apply models quickly, one nice attribute of decision tree modeling. For this same reason we also chose not to consider chemical descriptors that required an optimal energy-minimized 3D structure (i.e. 3D pharmacophore feature-based descriptors) because their computation for several million candidate structures would result in slower turnaround for forecasting.

At the time of their development, the activity classifiers demonstrated lifts for predictions from the randomly selected testing subsets significantly higher than 5 to 6 with lifts averaging 50−55 across all of the models. We explored the individual HTS data sets and the corresponding training and testing subsets more carefully so as to understand why the models performed more poorly in compound acquisitions than the original validation results would have suggested. Sheridan[39] has noted that the level of similarity between a molecule and those in the training set is a good discriminator as to how well that molecule will be predicted by a model built on that training data. We analyzed the overall diversity of the chemistry comprising the actives in each our HTS data sets. What we found (Table 6) were islands of similar compounds clustered around molecular scaffolds that represented synthetic efforts around tightly focused historical lead areas. Of course, there were diverse actives that had arisen from effective compound acquisitions, but these were dominated by analogs derived from lead optimization efforts. Randomly dividing the HTS data into training and testing subsets, a process widely practiced for model validation, would essentially place structurally related analogs into both the training and testing subsets. Inspection of Table 6 reveals the herbicide HTS data set actives were the least structurally diverse, displaying the largest maximum cluster size, the largest average cluster size, and the lowest proportion of singleton clusters at 90% similarity. The fungicide HTS data set actives were the most structurally diverse, while the insecticide HTS data set actives were in between these two. Clearly there is a bias within the HTS data sets toward actives derived from historical synthetic optimization efforts. This bias had to be mitigated if trained models were to be more robust.

To understand and quantify the effect of including similar molecular analogs on classifier performance we quantified the similarity between molecules in the testing and training subsets. Each active molecule in the testing subsets was compared to the compounds in the respective training subset

Table 7. Similarity Adjusted Classifier Lifts for Two HTS Insecticide Data Sets

		similarity adjusted lift			sim lift/lift (%)		
descriptor	lift	0.950 Sim	0.900 Sim	0.850 Sim	0.950 Sim	0.900 Sim	0.850 Sim
			Homopteran				
166 sBITs	23.6	23.3	17.8	9.9	99%	75%	42%
2D AtomPairs	33.7	33.4	25.0	21.1	99%	74%	63%
3D AtomPairs	17.1	16.6	14.5	9.1	97%	85%	53%
960 sBITs	38.4	37.1	32.8	23.3	97%	86%	61%
AlogPATs	30.0	30.0	24.4	16.2	100%	81%	54%
BCUTs	10.5	10.3	8.7	6.4	98%	83%	61%
C^2Combichem	6.4	6.4	5.5	4.2	100%	86%	66%
MolconnZ	18.3	17.9	15.3	12.1	98%	84%	66%
average	22.2	21.9	18.0	12.8	98%	82%	58%
			Lepidopteran				
166 sBITs	37.0	24.9	8.9	4.1	67%	24%	11%
2D AtomPairs	63.9	52.9	23.7	6.5	83%	37%	10%
3D AtomPairs	50.1	35.3	17.0	4.8	70%	34%	10%
960 sBITs	65.8	57.7	22.6	6.1	88%	34%	9%
AlogPATs	37.5	26.3	13.7	5.9	70%	37%	16%
BCUTs	14.3	10.0	5.7	3.0	70%	40%	21%
C^2Combichem	9.3	6.7	4.2	2.9	72%	45%	31%
MolconnZ	38.0	23.3	11.1	4.8	61%	29%	13%
average	39.5	29.6	13.4	4.8	73%	35%	15%

and the number of nearest neighbors counted at three levels of similarity. Similarity scores were computed using the 960 ISIS fingerprints and the Tanimoto similarity coefficient. If an active in the test set was found to have 12 or more neighbors in the training set at or above the selected similarity threshold, then its prediction was disregarded during the construction of the ROC curves used to compute the enrichment. Our reasoning was that a molecule with 12 or more neighbors in the training set was likely a member of a closely knit family of analogs from an historical synthesis optimization program and so would be expected, based on Sheridan's work,[39] to be trivially well predicted because of its obvious close structural relationship to other family members in the training set. These similarity analyses were conducted for all of the HTS data sets and at several levels of similarity defining a nearest neighbor and the results for two analyses summarized in Table 7. A significant drop in classifier performance (Lift vs Similarity Adjusted Lift) was observed as the maximum allowed similarity level was lowered from 95% to 90% to 85%. The performance of the Lepidoptera models degraded more significantly than the Homoptera models at the same level of allowed similarity. We attribute this to the higher degree of similarity amongst the actives in the Lepidoptera HTS data set (Table 6). Thus the Lepidoptera models, having been developed on a more clustered sampling of chemistry, suffered the most when required to predict more novel structures. Across both data sets we see a dramatic drop in model performance when immediate family members are ruled ineligible for inclusion in selection lists.

These simulations provided insights into how well classifiers might perform when forecasting structures relatively novel compared to those on which the classifier was trained. As the test set contains more diverse chemical structures compared to those in the training set the performance of classifiers will suffer. The magnitude of drop in classifier performance in these simulations was, in fact, comparable to the differences that had been seen between the classifier performance during development (lifts of 50−55) and what was observed in screening vendor chemistry forecast as active by the classifiers (lifts of 5.9−6.3). Clearly the choice of similarity threshold, the similarity metric, the allowed number of neighbors, the choice of fingerprints used in the similarity calculations, and the composition of the data sets themselves would all affect the similarity adjusted outcome. Nevertheless the analysis revealed that models are much stronger at predicting compounds the more structurally similar they are to those molecules in the training set. These observations reinforce the dictum that an ideal training set would have high diversity amongst its members and would uniformly and broadly sample chemistry and biological activity space. Our goal for enhancing the hit-rate in HTS was to more effectively identify potentially novel active compounds from screening. Simply finding more of what you already know, while perhaps useful, would not have been very satisfying. Guided exploration of new areas of structural space, while more risky, opens the opportunity for pleasant surprises. Chance favors the well-prepared classifier.

These observations lead us to adopt the following clustering strategy for preprocessing HTS data sets before they were used in developing training and testing sets. An in-house version of Stochastic Cluster Analysis[40] was used to cluster the actives from each HTS data set at a specified level of similarity using the 960 ISIS Key fingerprints. The average cluster size is noted. Compounds were then chosen for inclusion in the modeling data set by randomly selecting up to the "average cluster size" number of members from each cluster. All members from clusters whose size is smaller than the "average cluster size" were allowed into the modeling data set. The net result of this strategy was to ensure under represented classes in a corporate collection were always included in the training process thus enhancing the overall diversity while balancing the contributions from historical analog synthesis optimization programs. These clustering scenarios were repeated at various levels of similarity. The clustering results for a typical HTS data set are reported (Table 8) and indicate the outcome (number of unique active structures allowed into modeling data set) is not strongly dependent upon the chosen similarity level. For example, clustering the actives from one HTS data set at 85%

Table 8. Typical Clustering Results for one HTS Data Set[a]

Tanimoto	AveSize	MaxSize	MinSize	NumSingle	TotalClusters	NumUnique	% of total
0.800	5.6	758	1	784	1265	2312	56%
0.825	5.7	648	1	890	1377	2370	58%
0.850	5.0	591	1	1012	1514	2505	61%
0.875	4.6	368	1	1135	1658	2650	64%
0.900	3.8	214	1	1299	1853	2773	67%
0.925	2.9	130	1	1546	2125	2859	69%
0.950	2.1	90	1	1898	2456	2931	71%

[a] All actives: 4114.

similarity produced 1012 singleton clusters and 502 non-singleton clusters. The largest cluster that formed contained 591 compounds. The average cluster size was five. Randomly selecting up to five members from each cluster provided a data set of 2505 unique active compounds for inclusion in the modeling data set, 61% of the total actives. After clustering a number of HTS data sets using this strategy and comparing the results we settled on a Tanimoto similarity threshold of 85%. This clustering strategy assures that data sets on which models are developed are not dominated by large numbers of compounds from a few chemical classes but contain a more uniform sampling of representatives from all of the classes.

Our clustering strategy was now addressing the need for a uniform and unbiased sampling of locally available compounds for model development, but the chemistry space of active pest control compounds is broader than our local collection. Our strategy did not yet address the issue of breadth of diversity of the chemistry on which a model is developed nor did it address the issue that random splitting of a data set potentially places closely related analogs into both the training and testing subsets.

Up to this point all models had been developed using HTS results from within the DuPont enterprise. It logically followed that training activity classifiers using screening results from the entire crop protection industry would provide more robust classifiers since they would be trained on a much broader sampling of active compounds. To this end a project was initiated[41] to curate the published agrochemical patents for companies which comprised the crop protection industry from 1970 to 2005. Information collected from each patent consisted of the structure of the chemical analogs that were disclosed and the level of biological activity against various agrochemically relevant species reported in that patent. Curation was limited to compounds for which some form of physical data was also reported. In order to prevent large patents and their associated chemistry from dominating the resulting database no more than 30 representative structures were abstracted from any one patent. If two patents were deemed equivalent only one was curated. The resulting database (AgPatentdb) contained over 93,000 chemical structures from over 7000 patents spanning nearly 35 years of reported discoveries. A nearest neighbor analysis was completed in which each AgPatentdb active was compared to our in-house active structures using the 960 ISIS fingerprint and a Tanimoto similarity threshold of 90% to define a near neighbor. The vast majority of patent actives (75 to 90% depending upon market area) returned zero near neighbors (at $\geq$ 90% similarity), and a smaller proportion (10 to 20%) returned fewer than 10 near neighbors from our corporate chemistry collection. The appropriate active structures were used to augment our in-house HTS training data sets and the respective classifiers retrained. This strategy addresses the breadth of diversity of chemistry for model development by sampling examples from a much broader source of crop protection actives.

One final strategy was employed that addressed the issue related to creating training and testing sets by random splitting of data sets. When classifiers are used to preselect structures from a vendor's chemistry collection, the underlying assumption is that models trained on yesterday's screening results can reliably predict tomorrow's novel active chemistry. Our earlier similarity adjusted lift discussion sheds some light on how true this assumption might be and clearly implies that there is a limit to how distant an analog can be from those in the training set and still be reliably forecast. However, our desire was to mitigate any excessive myopic bias in the training subsets that might ultimately impair a classifier's ability to extrapolate to novel structures. In order to reduce the likelihood that analogs of the same chemistry end up in both the training and testing subsets, we chronologically split the HTS data to create the training and testing subsets. Data used in training were selected from screening results up to a specific point in time, and screening results for all years after this point were placed into the testing set. Chemistry projects, even successful ones, have a finite lifetime after which the synthesis teams are redeployed to the optimization of more recently discovered leads. One component of our on-going compound acquisition strategy was to purchase compounds each year that were structurally dissimilar to compounds in our present collection, so that leads arising from screening over time tend to be structurally unrelated. For the purpose of testing the far-sightedness of our models, a chronological split strategy offers a significant advantage since classifiers can now be evaluated against test sets that are less likely to be comprised of or dominated by structurally related analogs. The performance estimates based upon forecasting the chronologically derived test sets should more closely mirror how the classifiers are subsequently used in the "real world". The classifier's inherent performance during training may not necessarily improve because of this strategy. However, the resulting assessment of that performance is likely to be a more realistic reflection of what will be seen when classifiers are used to forecast structurally dissimilar vendor chemistry. We wanted to train our classifiers to think out of the box.

Chronological splitting has another significant advantage when testing the real-world performance of our classifiers. When vendor chemistry was forecast using our classifiers, those compounds that scored well in the models were purchased. Ideally one would like to assess a classifier's performance for predicting both the active and inactive

Table 9. Hit-rate of Compounds by Classifier Forecast[a]

	fungicide models		herbicide models		insecticide models	
	ascomycete	oomycete	monocot	dicot	homopteran	lepidopteran
Model Says "Active"	3.92%	2.68%	2.94%	4.12%	4.13%	3.62%
Model Says "Inactive"	0.13%	0.13%	0.25%	0.29%	0.12%	0.18%
discrimination	30	21	12	14	34	20
actives recovered-new models	84%	81%	64%	65%	89%	85%
actives recovered-old models	78%	64%	6%	19%	39%	43%

[a] Model Says "Active" entries tally the actual biological screening hit-rate for those compounds forecast to be active. Likewise, Model Says "Inactive" entries tally the actual biological screening hit-rate for those compounds forecast to be inactive. Discrimination is the ratio of Model Says Active/Model Says Inactive hit-rates. The "Actives Recovered" entries tally the % of the HTS actives that were actually correctly predicted to be active by the classifiers.

classes, but this would have required the purchase of compounds that the models forecast to be inactive. This was not seen as a wise use of resources. Chronological splitting actually has the advantage that one can now use the classifiers to predict a large collection of chemistry (potentially several years of screening and tens or hundreds of thousands of compounds) whose biological outcome is already known and which comprise both active and inactive structures. Comparing the actual biological activity of compounds forecast as "active" and those forecast as "inactive" by the classifiers would provide a measure of discrimination for the classifier across both classes. We chose the time point for the chronological split such that nearly 30 months of screening results were placed into the testing subsets and more than 20 years of screening results were placed into the training subsets for each classifier domain.

Classifiers were retrained following implementation of all three strategies (pre-cluster in-house active chemical structures, include relevant active chemical structures from the AgPatentdb, and split the data sets chronologically into train and test sets) and the resulting test sets forecast by the new classifiers. The resulting classifiers demonstrated significant improvements in their predictive capabilities (Table 9) compared to those developed earlier.

Our implemented strategies resulted in classifiers that could achieve significant discriminations between active and inactive structures. When our newest classifiers predicted structures to be "active", the actual resulting hit-rates for these structures were 12 to 34 times higher than that seen for the structures predicted to be "inactive". The new and improved classifiers were capable of correctly identifying 81−84% of the actives in the fungicide domain, 85−89% of the actives in the insecticide domain, and 64−65% of the actives in the herbicide domain. These results were significantly better than those achieved forecasting the respective test sets using older models developed prior to implementation of these final strategies. The area in which the actives in the original HTS data sets were least diverse (Table 6 - herbicides) saw the greatest improvement in classifier performance (6 to 19% recovery increases to 64 to 65% recovery), reinforcing the benefit of these preprocessing strategies. The classifier performance seen for the older models vs that seen with the new models emphasizes that training classifiers on too narrow of a sampling of chemistry results in significantly poorer performance. These new classifiers are now an integral part of the compound acquisition process for HTS.

CONCLUSIONS

We have provided an overview of a multiyear project with the goal of developing activity classifiers that when applied to preselect compounds for HTS would provide significant enhancements to the observed screening hit-rates. We have seen that such classifiers can be developed if they are trained on historical HTS results that are balanced with respect to the sampling of chemistry, that are appropriately diverse, and the classifiers are validated in ways that accurately reflect how the models will be ultimately used. Significantly lower false positive error rates are achieved with ensemble-based classifiers that are a composite of models derived from multiple chemical descriptors or fingerprints, making them especially useful when applied to large chemistry collections.

Supporting Information Available: Listing of all of the descriptors that were considered in modeling for each of the descriptor types. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Zhang, Q.-Y.; Aires-de-Sousa, J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *J. Chem. Inf. Model.* **2007**, *47* (1), 1–8.
(2) Teramoto, R.; Fukunishi, H. Supervised Consensus Scoring for Docking and Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (2), 526–534.
(3) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47* (1), 150–158.
(4) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques. *J. Chem. Inf. Model.* **2007**, *47* (1), 92–103.
(5) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual Screening of Chinese Herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, *47* (2), 264–278.
(6) Dutta, D.; Guha, R.; Wild, D.; Chen, T. Ensemble Feature Selection: Consistent Descriptor Subsets for Multiple QSAR Models. *J. Chem. Inf. Model* **2007**, *47* (3), 989–997.
(7) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47(1*, 219–227.
(8) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis. *J. Chem. Inf. Model.* **2006**, *46* (5), 2110–2124.
(9) Arodz, T.; Yuen, D. A.; Dudek, A. Z. Ensemble of Linear Models for Predicting Drug Properties. *J. Chem. Inf. Model.* **2006**, *46* (1), 416–423.
(10) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Model.* **2005**, *45* (3), 786–799.
(11) Li, S.; Fedorowicz, A.; Singh, H.; Soderholm, S. C. Application of the Random Forest Method in Studies of Local Lymph Node Assay

Based Skin Sensitization Data. *J. Chem. Inf. Model.* **2005**, *45* (4), 952–964.
(12) Gini, G.; Craciun, M. V.; Konig, C. Combining Unsupervised and Supervised Artificial Neural Networks to Predict Aquatic Toxicity. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1897–1902.
(13) van Rhee, A. M. Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Comput. Sci* **2003**, *43* (3), 941–948.
(14) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forest: Combining the Predictions of Multiple Independent Decision Tree Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 525–531.
(15) 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* (6), 1947–1958.
(16) Manallack, D. T.; Tehan, B. G.; Gancia, E.; Hudson, B. D.; Ford, M. G.; Livingstone, D. J.; Whitley, D. C.; Pitt, W. R. A Consensus Neural Network-Based Technique for Discriminating Soluble and Poorly Soluble Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 674–679.
(17) Lanctot, J. K.; Putta, S.; Lemmen, C.; Greene, J. Using Ensembles to Classify Compounds for Drug Discovery. *J. Chem. Inf. Comput. Sci* **2003**, *43* (6), 2163–2169.
(18) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci* **2002**, *42* (4), 903–911.
(19) Simmons, K.; Kinney, J.; Owens, A.; Kleier, D.; Bloch, K.; Argentar, D.; WalshA.; Vaidyanathan, G. A , Comparative Study of Machine-Learning and Chemometric Tools for Analysis of In-vivo High-Throughput Screening Data. J. Chem. Inf. Model. **2008**, 48 (8), 1663–1668.
(20) Livingston, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
(21) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.
(22) Pipeline Pilot, version 5, Scitegic Inc., San Diego, CA. http://www.scitegic.com (accessed July 19, 2008).
(23) MDL connection table specifications available at Symyx, Inc., San Ramon, CA. http://www.mdli.com (accessed July 20, 2008).
(24) Concord is available from Tripos, Inc., St. Louis, MO. http://www.tripos.com/data/SYBYL/Concord_072505.pdf (accessed July 19, 2008). For the specifications of the mol2 format, see: http://tripos.com/data/support/mol2.pdf (accessed July 19, 2008).
(25) Cheshire, version 3, distributed by Symyx, Inc., San Ramon, CA. http://www.mdl.com/products/pdfs/mdl_cheshire_ds.pdf (accessed July 20, 2008).
(26) Carhart, R.; Smith, D. H.; Venkataraghavan, R. J. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(27) Many thanks to Prof. Alex Tropsha, University of North Carolina.
(28) Cerius2, version ccO, Accelrys, Inc., San Diego, CA. http://accelrys.com/products/cerius2/ (accessed July 20, 2008).
(29) Accelrys, Inc., San Diego, CA. http://accelrys.com/products/datasheets/ds_qsar_0907.pdf (accessed July 23, 2008).
(30) MolconnZ, version 3.50, EduSoft, Ashland, VA. http://www.edusoft-lc.com/molconn (accessed July 19, 2008).
(31) MolconnZmanual, version 4, EduSoft, Ashland, VA. http://www.edusoft-lc.com/molconn/manuals/400/ (accessed July 20, 2008).
(32) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. In *Perspective in Drug Discovery and Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; KLUWER/ESCOM: Dordrecht, 1998; Vols. *7/8*, p 31.
(33) DiverseSolutions, version 4, University of Texas, Austin, distributed by Tripos, Inc., St. Louis, MO. http://www.tripos.com (accessed July 20, 2008).
(34) Owens, A. J. D. L.; Filkin, D. L. Efficient training of the Back Propagation Network by solving a system of stiff ordinary differential equations. International Joint Conference on Neural Networks, II, Washington, DC, 1989; pp 381−386.
(35) Vaidyanathan, G. InfoEvolve Moving from Data to Knowledge Using Information Theory and Genetic Algorithms. *Ann. N.Y. Acad. Sci.* **2004**, *1020*, 227–238.
(36) Murphy, S. K.; Kasif, K.; Salzberg, S. A System for Induction of Oblique Decision Trees. *J. Artif. Intell. Re.* **1994**, *2*, 1–32.
(37) Hawkins, D. Formal Inference-Based Recursive Modeling, version 2.3, Univ. of Minnesota, Duluth, MN, 1999.
(38) For a discussion of the Bonferroni statistical correction, see: http://en.wikipedia.org/wiki/Bonferroni_correction (accessed Mar 31, 2008) and http://www.itl.nist.gov/div898/handbook/prc/section4/prc473.htm (accessed Mar 31, 2008).
(39) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (6), 1912–1928.
(40) Reynolds, C. H.; Druker, P.; Pfahler, L. B. Lead Discovery Using Stochastic Cluster Analysis (SCA): A New Method for Clustering Structurally Similar Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305–312.
(41) GVK Biosciences, Hyderabad, India. http://www.gvkbio.com (accessed July 20, 2008).

CI800164U