Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis

Milan Ganguly,[‡] Nathan Brown,*[,†] Ansgar Schuffenhauer,[†] Peter Ertl,[†] Valerie J. Gillet,[‡] and Paulette A. Greenidge[†]

Novartis Institutes for BioMedical Research, Basel, CH-4002, Switzerland, and The Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received December 8, 2005

An evolutionary statistical learning method was applied to classify drugs according to their biological target and also to discriminate between a compilation of oral and nonoral drugs. The emphasis was placed not only on how well the models predict but also on their interpretability. In an enhancement to previous studies, the consistency of the model weights over several runs of the genetic algorithm was considered with the goal of producing comprehensible models. Via this approach, the descriptors and their ranges that contribute most to class discrimination were identified. Selecting a bin step size that enables the average descriptor properties of the class being trained to be captured improves the interpretability and discriminatory power of a model. The performance, consistency, and robustness of such models were further enhanced by using two novel approaches that reduce the variability between individual solutions: consensus and splice modeling. Finally, the ability of the genetic algorithm to discriminate between activity classes was compared with a similarity searching method, while naïve Bayes classifiers and support vector machines were applied in discriminating the oral and nonoral drugs.

1. INTRODUCTION

The rationale for the requirements of pharmaceutical companies to increase the structural diversity of their databases is well documented. However, interest also exists in enriching such databases with molecules that exhibit a certain type of biological activity. These molecules need not necessarily be structurally related; indeed it may be advantageous if they are not. More recent literature has turned to the subject of scaffold hopping[1] as it is beneficial for a medicinal chemistry project to have alternative scaffolds available to it, in case a potent lead series is identified as having undesirable absorption, distribution, metabolism, excretion, and toxicology (ADMET) properties arising in part from a common building block.[2]

A first step in compound selection strategies is usually to apply computational filters to remove compounds with undesirable ADMET properties from further consideration. The simplest methods are based on simple counting of features with the most well-known of these being Lipinski's rule of 5 (Ro5).[3] More sophisticated approaches have been developed that aim to classify compounds as druglike or nondruglike.[4,5] These methods attempt to derive classification rules based on a training set of known class membership. The algorithms 'learn' classification rules from the input data in the training set with the rules being based on the molecular descriptors used to represent the compounds. Once the algorithms have been trained, then predictions can be made about previously unseen compounds. Similar methods have also been applied to predict about specific activities including binary kernel discrimination[6,7] (BKD), support vector machines[8] (SVM), partial least squares−discriminant analysis[9] (PLS-DA), and naïve Bayes classifiers[10] (NBC).

Classification methods are crucially dependent on the training sets used to represent the compound sets, and these have usually been derived from large publicly available data sets. For example, druglike compounds have been represented by the MDL Drug Data Report[11] (MDDR), the Comprehensive Medicinal Chemistry[12] (CMC) database, and the World Drug Index[13] (WDI), and nondruglike compounds have been represented by databases such as the Available Chemicals Directory[14] (ACD) and the SPeicherung und REcherche Strukturchemischer Information[15] (SPRESI) database. More recently, the use of these databases for deriving such rules has been questioned[16,17] since they are not necessarily representative of the respective classes. For example, MDDR contains a large number of biologically inactive molecules, whereas the ACD includes many oral druglike molecules.

The classification methods applied to the druglike, nondruglike problem have included neural networks, decision trees, and genetic algorithms (GAs). A drawback of neural networks (including SVMs) is that the models derived are not easily interpretable. Decision trees, on the other hand, can produce rules that are easy to interpret; however, they are prone to overtraining with the rules being based on chance correlations.[18,19] A variety of different descriptors has been investigated from simple physicochemical properties through substructural fragments.

Gillet et al.[20] described a GA for distinguishing between druglike and nondruglike molecules. The study was based

* Corresponding author phone: +41 (0) 61 324 90 29; fax: +41 (0) 61 324 33 57; e-mail: nathan.brown@novartis.com.
[†] Novartis Institutes for BioMedical Research.
[‡] The University of Sheffield.

10.1021/ci050529l CCC: $33.50
Published on Web 07/21/2006

on a small number of descriptors including counts of structural features (aromatic rings (AR), rotatable bonds (RTB), hydrogen bond donors (HBD), and hydrogen bond acceptors (HBA)) and the simple physicochemical properties molecular weight (MW) and the ${}^2\kappa_\alpha$ shape index. Each descriptor was represented by a fixed number of fixed sized bins. In the case of the structural features, each bin represents an integer count value, whereas for the physicochemical properties each bin represents a range of values. The GA was designed to evolve a set of weights, one for each bin, which is best able to lead to discrimination between the two classes. Each molecule in the training set is assigned a score which is the sum of weights corresponding to its property values; the molecules are ranked on score; and distributions of molecules in the ranked list are examined. The method is essentially a ranking method with molecules being ordered in descending probability of activity as judged by the fitness function of the GA. However it could also be considered as a classification method by applying a suitable threshold to the scores.

The initial method was shown to be surprisingly effective at distinguishing between active and inactive molecules despite being based on such simple easy-to-calculate properties, and it offers the considerable advantage over the simple counting methods such as Ro5, in that it is able to take account of the co-occurrence of features. Furthermore, unlike neural networks, a GA is not a 'black-box', and the models are therefore amenable to interpretation, although this aspect was not exploited in the original study due to the requirement for additional processing. For example, while high weights assigned in the GA could be indicative of features with high significance, they could also represent spurious results that arise simply due to the feature having low incidence in the data set. Furthermore, the nondeterministic nature of a GA means that different results can be identified each time the GA is run, due to the stochastic processes in population initialization and subsequent perturbation of those chromosomes through genetic operators, further complicating the interpretability issue.

In the work presented here, we have improved on the earlier study in a number of ways. We have developed two different approaches to extracting interpretable models from a GA. These methods also result in increased robustness of the models. The first approach takes into account a range of feasible solutions to generate a consensus solution that exhibits more consistent characteristics. Consensus scoring (also called data fusion) methods are applied widely in a variety of challenges in chemoinformatics where multiple evaluations are available that cover alternate solutions. For example, the combination of multiple docking scores using consensus methods has been demonstrated to be of great value in generating enhanced results.[21] As data fusion, consensus methods have been used effectively in ligand-based similarity searching where multiple bioactive reference structures are available.[22] The approach has also been applied successfully to the generation of robust quantitative structure−activity relationships (QSARs).[23] In the second approach developed here, a splice model is generated by exchanging the set of bins for a particular descriptor in instances when the new bins are seen to be more optimal, thereby providing a more information-rich model.

We have applied our consensus approaches to an enhanced GA as compared with the original published method. Here, we consider a wider variety of descriptors, and we undertake a bin analysis in which the optimum number of bins for each descriptor is determined using different bin step size selections. The resulting GA permits consideration of ranges for integer descriptors as opposed to discrete values.

We have tested the approach on several data sets, including two activity classes extracted from MDDR. The renin inhibitor data set consisting of molecules with a high level of structural homogeneity was initially used to investigate model interpretability. We have also compared our results with similarity searches based on conventional 2D descriptors. Here, the GA is not expected to improve on enrichments seen in similarity searches due to the much simpler descriptors employed; however, it provides a complementary view of a data set based on easy-to-interpret descriptors.

We have also trained the GA on a set of oral and nonoral druglike compounds identified by Vieth et al.[24] This data set represents a significant challenge, since Vieth et al. demonstrate significant overlap between the distributions of individual properties and conclude that any differences between the two classes of compounds are subtle. By considering the co-occurrences of features, the GA is able to achieve significant discrimination between these classes with the resulting model providing useful information for subsequent compound selection tasks.

2. MATERIAL AND METHODS

2.1. Genetic Algorithm for Class Discrimination. The GA proposed by Gillet et al.[20] proceeds by evolving a solution model (or models) based on descriptor range weights to separate a given set of molecules into actives and inactives. The GA is an analogue of natural evolution that is very effective at optimizing in large search spaces and generally finds an optimal or near-optimal solution very quickly. The typical GA proceeds by determining a suitable encoding strategy that encodes the solution space into a chromosome representation that may then be decoded to provide the score for a particular chromosome, called the fitness. A population of these chromosomes is perturbed over a number of generations using crossover and mutation operators. Crossover (an analogue of recombination) exchanges genetic subcomponents of two chromosomes to generate two new chromosomes that are reflective of both progenitors. Mutation proceeds by randomly altering the genes of a single chromosome with some probability, thereby introducing new genetic material into the population.

The chromosome representation that Gillet et al. applied encapsulates a set of descriptors by binning those descriptors into defined ranges. Each bin (or gene) may be assigned a bin weight (or allele) in a user-specified range $\{0...R\}$, where R is the maximum permissible value and is user-defined. Initially, each of the chromosomes in a population is assigned random integer weights for the bins in the specified allele range. As the evolution process proceeds, the bin weights are evolved to maximize discrimination according to our fitness function.

The evaluation of a chromosome is performed by a fitness function that considers how well the chromosome discriminates the two classes under consideration. Each molecule is

scored in the following way: the descriptor vector of the molecule is used to determine which bin weight to extract for each descriptor. The extracted weights are summed over all descriptors to give a score for each molecule. The molecules are then sorted into descending order of score. A number of fitness measures are applied in this study, and these are considered in more detail below. The fitness functions in each case consider the distribution of the two molecular classes under consideration (whether this is the molecular class membership or simply bioactivity) within the ranked list.

The GA was reimplemented in Java for the study published herein. The modifications made to the previous method are discussed below.

2.2. Measures of Performance. Previously, Gillet et al.[20] considered two separate performance measures of a given molecular ranking: the *initial enhancement* (IE) and *global enhancement* (GE). IE is calculated by, first, counting the number of active molecules that occur in the first *NACT* molecules of the ranked list, where there are *NACT* active molecules in the data set. This number is divided by the number of molecules that would be expected to occur in the top *NACT* by chance. The number of molecules expected by chance can be calculated by dividing *NACT* by the total number of molecules in the data set (*NACT* + *NINACT*) and then multiplying the result by *NACT*. GE is calculated in the following manner. The number of molecules that would have to be tested in order to retrieve half of the active molecules if the data set were randomly ordered (half the data set) is divided by the number of molecules required to retrieve half of the actives in the ranked data set. IE considers the quality of ranking in terms of the enrichment of the number of actives in a defined first portion of the ranking only, compared with an idealized random ranking. GE, however, considers the enhancement of the ranking in terms of the actual ranking of the first half of the actives compared with the expected random ranking of this first half.

While these scores, when applied in concert, are useful for enrichment studies, for the challenge of classification it is more important to score a given model based on maximizing the separation of the two classes under consideration. Therefore, we propose the use of the *maximized difference enhancement* (MDE) where the explicit score of each molecule in the data set under investigation is applied to calculate the average score of each class. The absolute difference between these two scores then provides a more accurate indication of the quality of separation. MDE is calculated by direct application of the scores of each of the molecules in the training set. The average score of the molecules in the two classes is first calculated, and then the MDE is given by the absolute difference of these average scores.

While the fitness function is based on the MDE, IE and GE values are reported for the data sets for several reasons. First, these values give an indication of whether the fitness function has been successfully implemented or not. That is, it is to be expected that the IE and GE values for a particular run of the GA should be similar if the fitness function is effective. Second, MDE values are not directly comparable between runs of a GA (experiment) with different numbers of descriptors. Third, IE and GE values allow the performance of the GA for class discrimination to be directly compared with that of the similarity method in the second part of the study. The maximum values that can be returned for IE and GE depend on the composition of the data set, and hence values are reported as a percentage of the "maximum obtainable" value, when comparing IE and GE values between data sets with different compositions.

2.3. Determining Optimum Bin Sizes. For each data set, an initial set of descriptors was determined, as described below, and several runs of the GA were carried out in which bin ranges, population size, and GA operators were varied and the consistency of the bin weights was analyzed. A bin that has a consistently high weight for a given descriptor may provide an indication of the importance of that descriptor for class discrimination and also the optimal value of that property that is required in order to maximize class discrimination. Conversely, descriptors that are represented mainly by bins with fluctuating weights are unlikely to contribute to the predictive power of the GA model and therefore may be removed to provide a more parsimonious model. Not all descriptors were assigned the same number of bins in the GA chromosome representation. The number of bins and their individual step sizes were iteratively optimized based on the analysis of bin weights. Ranges for integer descriptors were investigated. Our strategy was to calculate the minimum (bin start) and maximum (bin stop) value for each selected descriptor (in the class being trained) and then to apply these values as the lower and upper limit values for the bin ranges. Thus, depending on these values, it may be that a particular descriptor is represented by many more bins than the other descriptors. In such a case, the relevance of the descriptor for class discrimination may be inadvertently overemphasized; however, this may be avoided by using a larger bin step size. A workflow of the model development process used for each case study is provided in Figure 1.

2.4. Consensus Models. In addition to the use of individual models, consensus models were constructed. The generation of a consensus model proceeds as follows. The average bin weights from five runs of a particular GA experiment were taken and rounded to the nearest integer. These values were then used to construct a new model which was run with the test sets (Figure 2).

The reason for building the consensus models is to reduce the variability between individual candidate solutions due to inconsistent bin weights. It was anticipated that the consensus model resulting from the ensemble would emphasize bins with high weights but at the same time de-emphasize bins with fluctuating and generally minimal weights therefore leading to a more interpretable model with increased discriminatory power. The usefulness of a model should be judged not only on its accuracy but also on its interpretability, as a chemist's ability to provide a solution to a problem is often hampered by the lack of transparency of a model.[25]

2.5. Splice Models. A complementary extension of the consensus model is the splice model. A splice model may be constructed by combining consensus models which result from the use of the same or different descriptors. The generation of a splice model is illustrated in Figure 3. The weight of each bin is optimized with respect to that of all other bin weights in a particular model; however, a subcomponent of a consensus model may be transferable between consensus models. In consensus models, the effects

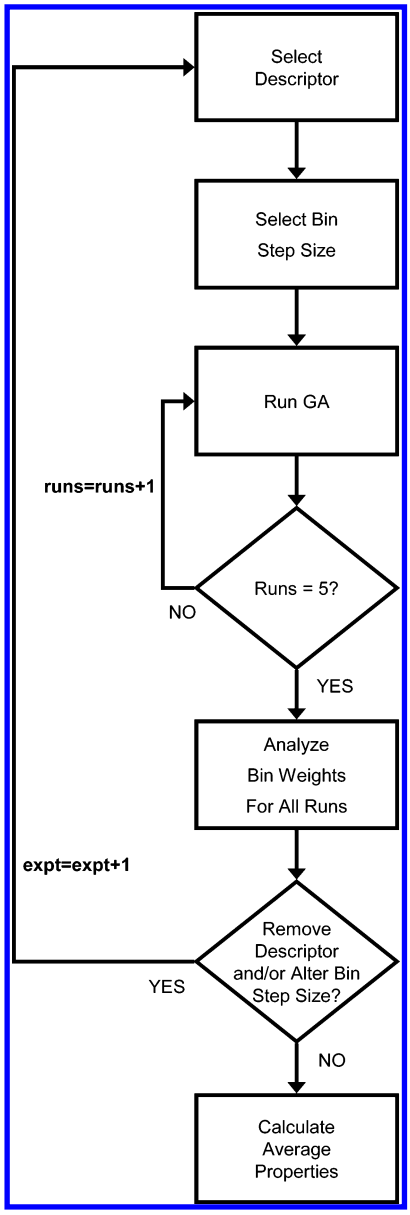

Figure 1. The workflow of the iterative model development method applied in this study. For each of the runs of the GA, the population size and number of generations were both constant at 500. The probability of crossover and mutation were 0.7 and 0.01, respectively. The weight range was set to 7 based on the trials of the GA with another data set.

of fluctuating bin weights are dampened, and thus the effects of the variability of bin weights arising from their optimization with respect to other bin weights has already been alleviated. The desirability of a splice model stems directly from this fact that the bin weights are all optimized with respect to each other, so that a consensus model with an overall different total number of bins and/or descriptor composition may find alternative solutions for a common descriptor.

2.6. Data Sets. The data sets used in this study are the 11 activity classes (8284 molecules) profiled from the MDL Drug Data Report[11] (MDDR) database by Hert et al.[26] (summarized in Table 1) and the data set of Vieth et al. consisting of 1729 marketed drugs.[24] A brief description of each descriptor calculated for our data sets using the Novartis in silico profiling tool[27] is given in Table 2.

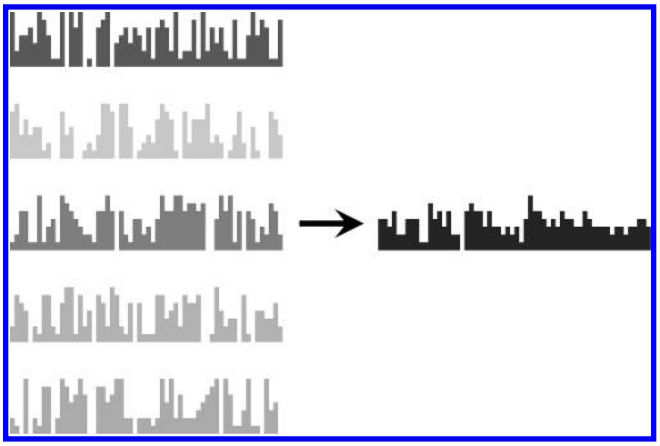

Figure 2. Diagram of consensus model construction. Average bin weights from five runs of an experiment are combined to produce a consensus model.

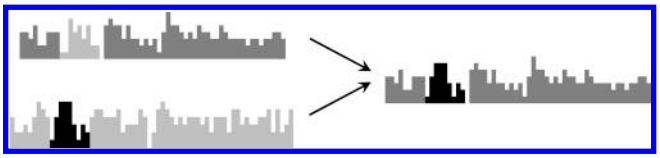

Figure 3. Diagram of splice model construction. A particular descriptor block is exchanged between the consensus models.

Table 1. Summary of Data Sets Profiled from MDDR

activity class	no. of molecules
5HT3 antagonists	745
5HT1A agonists	827
5HT reuptake inhibitors	359
D2 antagonists	395
renin inhibitors	1,130
angiotensin II AT1 antagonists	943
thrombin inhibitors	802
substance P antagonists	1,246
HIV protease inhibitors	750
cyclooxygenase inhibitors	636
protein kinase C inhibitors	451

Vieth et al. compiled data for marketed drugs which contain both structural and route of administration information. The oral data set was taken to be representative of drugs with good pharmacokinetic and pharmacodynamic (PK/PD) properties, while those with no oral formulation (absorbent, topical, and injectable) were taken to be representative of drugs with poor PK/PD properties. Eight descriptors thought to be relevant for differentiating between these categories were calculated: MW, ON (number of oxygen and nitrogen atoms), OHNH (number of OH− and NH− groups), HBA, NRING (number of rings) RTB, HALOGEN (number of halogen atoms), and ClogP. Based on the mean values of the descriptors, injectable drugs are least like oral drugs, and topical (and absorbent) drugs are the most like oral drugs. However, because of the substantial overlap in the ranges of the calculated properties, Vieth et al. concluded that it was not possible with any confidence to classify a drug as being either oral or injectable. We have restricted ourselves to this size of data set because of the detailed analyses that were performed.

2.7. Generation of Training and Test Sets. *Renin and COX Data Sets*. A protocol was created in PipelinePilot[32] to calculate descriptor values for the data sets given in Table 1. Each of the 11 data set files was also given a label of zero for the field indicating whether a molecule belongs to

Table 2. Summary of Structural and Physicochemical Molecular Descriptors Applied in This Study

abbreviation	descriptor	description
ClogP	octanol/water partition coefficient	The octanol/water distribution coefficient logP is calculated by the CLOGP program from BioByte (v. 4.71).[28]
SA	polar surface area	The polar surface area is calculated as a sum of the surfaces contributed by nitrogen and oxygen atoms and their attached hydrogen atoms.[29]
MV	molecular volume	The molecular volume is the volume inside the molecule's Connolly surface. The calculated surface is based on a single molecule conformation obtained from the 3D structure generating program CORINA.[30]
CMR	calculated molar refractivity	The molar refractivity CMR is calculated by the CMR program from BioByte (v.4.71). Molar refractivity characterizes the molecular polarizability.[28]
MW	molecular weight	The mass of a molecule in Daltons.
HBA	hydrogen bond acceptors	Counts the number of hydrogen bond acceptors. Following the Lipinski rule-of-five definition all nitrogen and oxygen atoms are counted.
HBD	hydrogen bond donors	Counts the number of hydrogen bond donors. Following the Lipinski rule-of-five definition all OH and NH groups in the molecule are counted.
AB	amide bonds	Counts the number of amide bonds.
RTB	rotatable bonds	Rotatable bonds were defined as any single bond, not in a ring, bound to a nonterminal non-hydrogen atom. Excluded from the count were amide C−N bonds because of their high energy barrier.
FI	flexibility index	The ratio of rotatable bonds to molecular weight.
logBB	blood/brain partition coefficient	The prediction of the partition coefficient of blood/brain distribution is based on published data.[31] Positive values indicate a good penetration into the central nervous system (CNS), formula: $\log BB = 0.106 - 0.0139PSA + 0.1642ClogP$.
WS	aqueous solubility	The prediction of thermodynamic solubility from a linear regression model based on molecular fragments and physicochemical properties.
logPM	permeation coefficient	The permeation coefficient for a molecule is predicted from the calculation of the sum of its fragment contributions. The contributions of 100 atom-centered fragments were determined by least-squares analysis of a molecule set with experimental values for Caco-2 permeabilities.

the sought after class. Additionally two more files (copies of the renin inhibitors and COX inhibitors, respectively) were profiled and labeled one to indicate they were the sought after molecules. One training set and two test sets were created using the random percent filter in PipelinePilot to divide the 1130 renin inhibitors into three approximately equally sized subsets. The remaining classes were combined in one file which was also subsequently divided into three approximately equally sized subsets. Initially, the COX inhibitor set that was labeled zero was included in the mixed class data set. The first of the three subsets from each master set were combined to create the training set. This consisted of 358 renin inhibitor molecules and 2341 mixed class molecules to give a total of 2699. This number was reduced to 2693 after removing molecules for which descriptors values could not be calculated. The two test sets were combined by merging the second and third subsets of each master set, respectively. The training and test sets were constructed in the same way for the COX inhibitors as for the renin inhibitors, except that the renin inhibitors labeled as zero were included in the mixed class set and the COX inhibitors labeled one were now the molecules being sought.

Oral and Nonoral Data Sets. After removing molecules with missing data, there were 1182 oral, 109 topical, 115 absorbent, and 294 injectable drugs, respectively. The nonoral (topical, absorbent and injectable) drugs were combined to create one data set. The oral and nonoral data sets were each split into 3 approximately equal parts with the random percent filter component of PipelinePilot. These were used to create one training set and two test sets in the same manner as detailed for the renin and COX data sets above.

2.8. Comparative Studies. To compare the GA results with the similarity method for drug class discrimination, as described by Hert et al.,[25] it was necessary to carry out experiments that would allow direct comparison of performance measures. For both the renin inhibitor and COX inhibitor data sets 50 molecules were selected randomly. These 50 molecules were divided into 5 sets of 10 to be used as reference structures, and 5 searches were conducted—one with each set of reference structures. The same protocol was used for each search as in the paper by Hert et al. (i.e. the Tanimoto similarity coefficient was used in combination with ECFP_4 fingerprint descriptors). The molecules of the remaining classes along with the remaining renin and COX inhibitors (plus the 90 COX and renin inhibitors not immediately being used as reference structures) were combined to create the test set consisting of 8274 molecules in total. This test set was ranked according to similarity to the reference structures. The ranking was then evaluated with the GE and IE. The same protocol was used when the number of reference structures was increased to 50, this time giving a test set size of 8234 structures.

The ability of the GA to discriminate between the oral and nonoral drugs is compared to the well-established NBC and SVM methods both in terms of the discrimination achieved and also in the interpretability of the models that are generated.

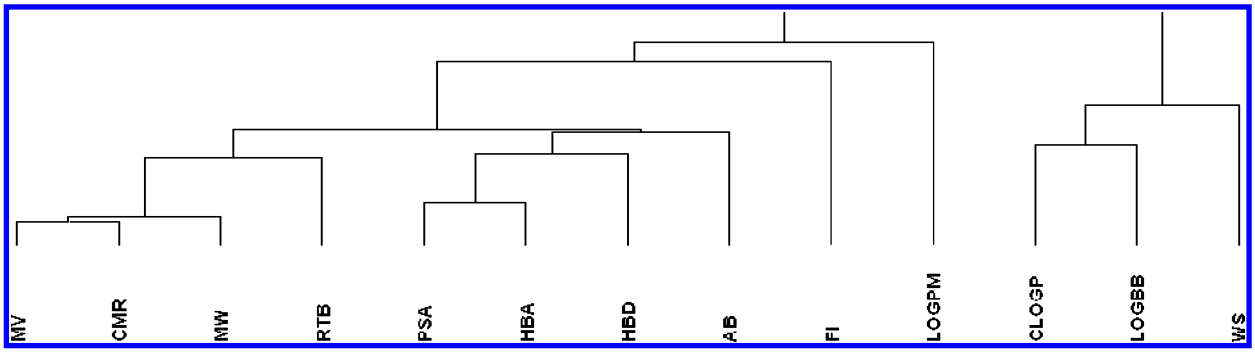

Figure 4. Dendrogram of the hierarchical clustering results of physicochemical and structural descriptors for the renin inhibitors.

Table 3. Summary of Descriptors and MW Step Size Used in Each of the Six Renin Inhibitor Experiments

	descriptors					
experiment	logPM	AB	HBD	ClogP	MW	MW step size
1	●	●	●	●	●	49.99
2		●		●	●	49.99
3	●	●	●	●	●	99.98
4		●		●	●	150.0
5	●	●	●			
6	●	●	●	●	●	150.0

Table 4. Bin Range Values for Selected Descriptors in Renin Inhibitor GA Training Runs[a]

descriptor	bin start	bin stop	bin step
logPM	−6.0	−1.5	0.5
AB	1	10	1
HBD	3	16	1
MW	350.0	1,350.0	{49.99, 99.98, 150.0}
ClogP	−2.0	9.5	0.5

[a] Start and stop values correspond to minimum and maximum values of the renin inhibitors data set. These bin sizes values were used for all experiments with the exception of the MW descriptor.

3. RESULTS

3.1. Case Study One: Renin Inhibitors. Descriptors were calculated for the renin inhibitors, and the profile was then analyzed by means of hierarchical clustering. The results from the cluster analysis assisted in determining descriptors that are orthogonal while also ensuring that those selected are also intuitive. The results of the clustering analysis (Figure 4) show the relationship between the different descriptors, and from inspection of the plot an informed decision was made on which descriptors would be most appropriate to use. The descriptors selected were MW, HBD, AB, logPM, and ClogP. They were chosen as they belong to distinct clusters with the exception of AB and HBD which belong to the same cluster.

The GA was run using the descriptors selected from the clustering analysis. For six experiments with a given set of descriptors (Table 3) and bin step sizes (Table 4), the GA was run five times, and average and standard deviation values were calculated for the bin weights. From consideration of the results, descriptors were removed (experiments 2, 4, and 5) and/or the bin step size was altered (experiments 3, 4, and 6) in a process of iterative optimization of the number of descriptors and bin ranges. IE, GE, and MDE values were also calculated for each run in an experiment.

Maximum Difference Fitness Function. Since there was no marked difference between the values for *GE* and *IE* for the six experiments (mean difference of 0.2±0.2), we concluded that the newly introduced *maximum difference* fitness function successfully combines the best features of the *top rank* and *average rank* fitness functions previously used by Gillet et al.[20]

Table 5. Mean and Standard Deviation of the Descriptor Values for the 358 Renin Inhibitors and for the 2335 Nonrenin Inhibitors of the Training Set

	renin inhibitors		nonrenin inhibitors	
descriptor	mean	SD	mean	SD
logPM	−3.2	0.7	−3.6	0.9
AB	3.5	1.2	1.1	1.3
HBD	5.4	1.9	2.0	2.0
MW	706.7	109.7	454.4	438.1
ClogP	4.6	1.8	4.1	2.0

Bin Step Size, Weight Consistency, and Average Descriptor Properties. Descriptors with bins having an average weight of ≥5 over five runs and a standard deviation of ≤1 were considered to be most important for class discrimination, whereas descriptors containing bins with lower weights were not considered to provide significant discrimination. Figure 5 summarizes the iterative optimization of bin sizes and descriptors, starting from experiment 1, which includes all five descriptors and uses bin step sizes that fit the range for the descriptor. To aid the interpretation of the significance of bin weight consistency, the average and standard deviation values were generated for the selected descriptors for renin and nonrenin inhibitors, respectively, in the training set (Table 5). It can be observed from Table 5 that for logPM and ClogP, respectively, there is a significant overlap between these values for the renin inhibitors and nonrenin inhibitors, respectively. Hence, these descriptors are not expected to contribute to discriminating between these classes. However, for the values of the other three descriptors (AB, HBD, and MW) some separation does exist between the data sets. Only experiment 6 solely highlights the three descriptors (AB, HBD, and MW) whose mean values (and standard deviations) are expected to be able to provide any discrimination between renin and nonrenin inhibitors (Figure 5 and Table 5). The AB descriptor, bin range 3−4 is highlighted which is in keeping with the average AB descriptor value of 3.5 for renin inhibitors. Also highlighted are the HBD descriptor, bin range 5−6 and the MW descriptor bin range 650−t800. The average values for these descriptors are 5.4 and 706.7, respectively, for renin inhibitors. Significantly, experiments 1, 3, and 6 use the same descriptors but have different step sizes for the MW descriptor (Figure 5). In contrast to experiment 6, experiments 1 and 3 find solutions for several descriptors that are

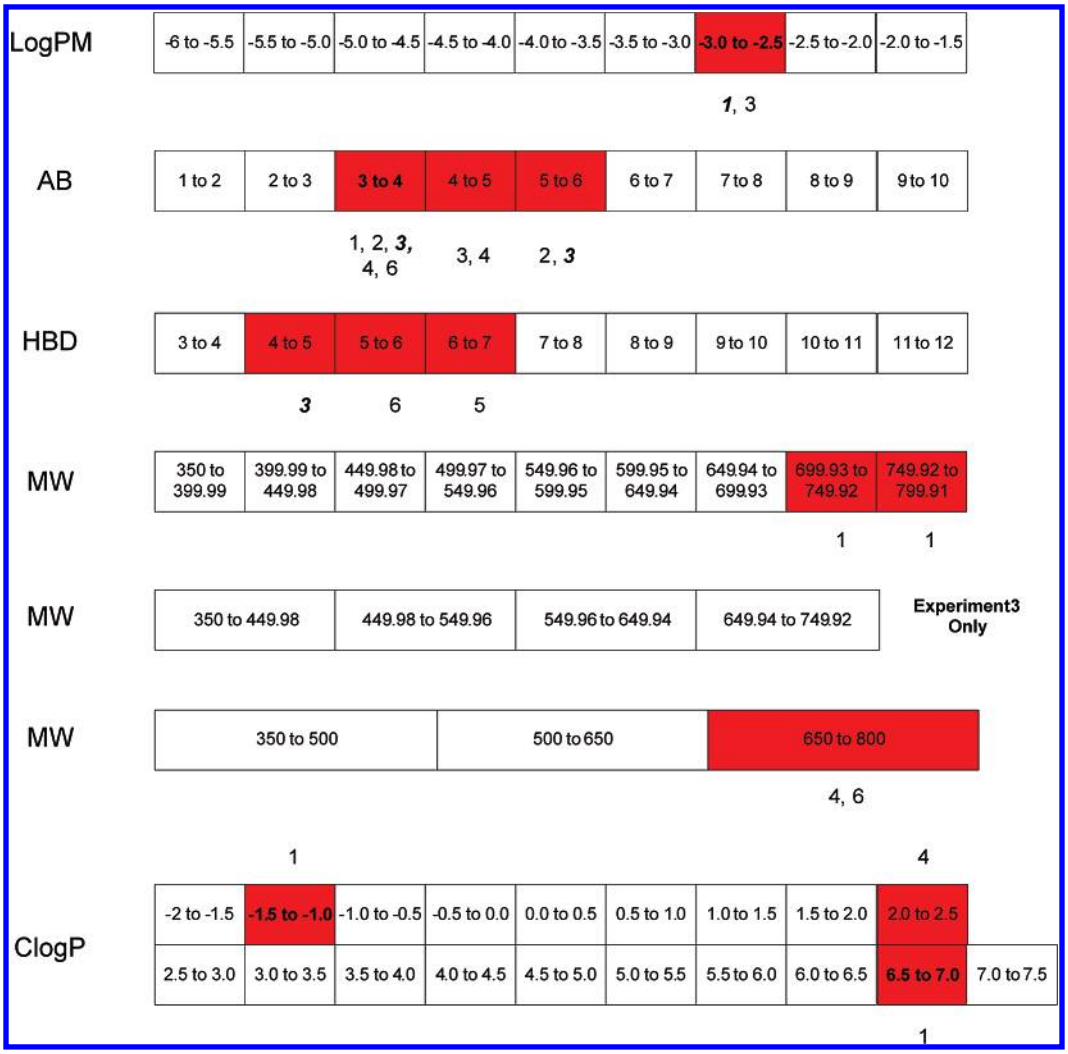

Figure 5. Results of analysis for consistency of weights corresponding to experiments 1 to 6. Bins having an average weight of ≥5 over five runs and standard deviation ≤1 are highlighted. Bins with lower weights were not considered to provide significant discrimination. Numbers in bold did not quite meet the cutoff criteria.

Table 6. Mean and Standard Deviation of the Descriptor Values for the 636 COX Inhibitors and for the 2505 Non-COX Inhibitors of the Training Set

	COX inhibitors		non-COX inhibitors	
descriptor	mean	SD	mean	SD
logPM	−3.4	0.7	−3.5	0.9
RTB	4.8	2.8	10.0	6.9
HBD	1.4	1.0	2.7	2.8
MW	350.5	73.6	510.0	188.3
ClogP	4.2	1.7	4.2	2.0

capable of distinguishing between renin and nonrenin inhibitors that have values other than the mean values for the renin data set. Instead the highlighted bins capture the upper or lower bounds of the standard deviation of the mean values (Figure 5). Since the candidate solutions from experiment 6 exactly mirror the average properties of the renin inhibitors, they may therefore be preferred for discriminating between renin inhibitors and other classes of drugs. The average of the 5 models for each of experiments 1, 5, and 6 are provided in Figure 6 (parts (a)−(c), respectively).

3.2. Case Study Two: COX Inhibitors. As for the renin data set clustering analysis was also conducted to select appropriate descriptors. The results are given in Figure 7.

Table 7. Bin Range Values for Selected Descriptors in COX Inhibitor GA Training Runs[a]

descriptor	bin start	bin stop	bin step
logPM	−6.5	−1.5	0.5,[b] 1.5[c]
RTB	0	22	3
HBD	0	6	1
MW	170.0	620.0	75
ClogP	−0.5	9.5	2.5

[a] Start and stop values correspond to minimum and maximum values of the COX inhibitors data set. [b] Step size used in experiment 1 only. [c] Step size used in experiment 2 only.

The descriptors selected were logPM, RTB, HBD, MW, and ClogP. The mean and standard deviation for the descriptor values for two sets of molecules composing the training set are given in Table 6. The bin step sizes (Table 7) were selected so as to capture the average property ranges of the COX inhibitors but to, at the same time, exclude any overlap with the properties of the non-COX inhibitors.

In the first of two experiments, three bins with consistently high weights were observed from the bin weight analysis. One corresponding to the RTB descriptor bin 2 (range 3−6) and the average value for the RTB descriptor for the COX inhibitors is 4.8 (Table 6), and the other two to the MW descriptor bins 2 (range 245−320) and 3 (range 320−395).

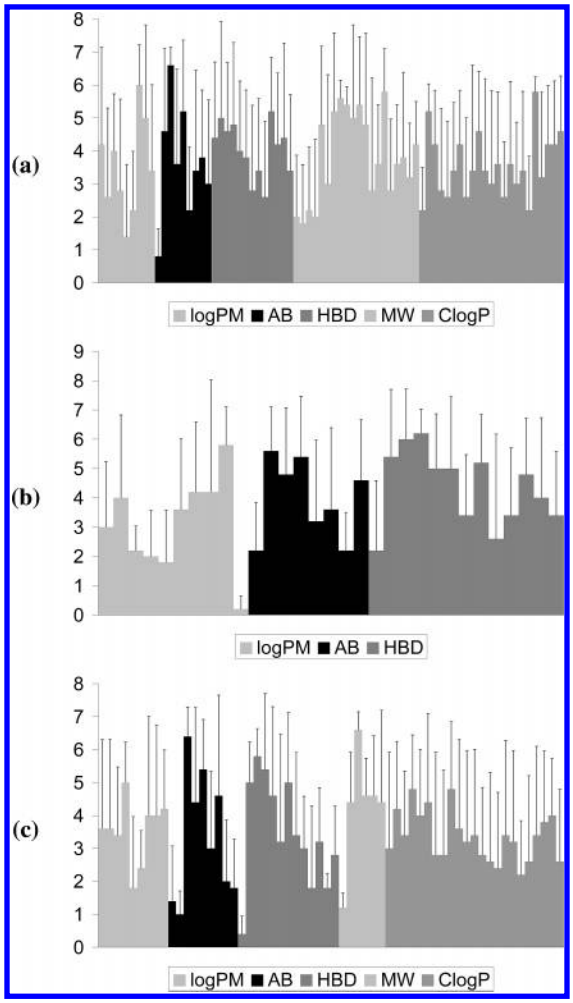

Figure 6. The 5 models for experiments (a) 1, (b) 5, and (c) 6, respectively. The average of the bin weights for each descriptor is given together with the standard deviations.

Bin 3 captures the average MW of the COX inhibitors (351), and bin 2 falls in the range of the standard deviation of the MW for the COX inhibitors which significantly excludes the non-COX inhibitors. In the second experiment we sought to optimize further the GA model by increasing the bin step size for logPM from 0.5 to 1.5; the mean property values of logPM overlap between COX and non-COX inhibitors; therefore, this descriptor is unlikely to contribute to discrimination between the classes (Table 6). As in experiment 1, the MW descriptor bins 2 and 3 plus the RTB descriptor bin 2 had consistently high weights with a low standard deviation. Additionally for the RTB descriptor, bin 3 (range 6−9) was identified as being important, and this value is within the standard deviation for the COX inhibitors.

Table 8. Mean and Standard Deviation of the Descriptor Values for the Oral and Nonoral Data Set

	drug category							
	oral		absorbent		injectable		topical	
descriptor	mean	SD	mean	SD	mean	SD	mean	SD
MW	342.3	142.2	394.0	251.6	491.2	305.2	372.0	173.3
ClogP	2.3	2.7	1.6	3.2	0.7	3.3	2.9	2.9
ON	5.5	3.4	6.6	6.3	9.6	7.6	5.1	4.5
OHNH	1.8	1.8	3.1	3.9	3.7	4.1	1.9	2.8
NRING	2.6	1.3	2.5	1.9	3.2	2.2	3.0	1.5
RTB	5.4	4.0	8.0	9.3	9.8	9.9	5.3	4.9
HBA	3.2	2.4	3.7	3.2	5.4	4.7	3.3	3.1
HALOGEN	0.5	0.9	0.6	1.4	0.3	1.0	0.9	1.2

Table 9. Bin Range Values for Selected Descriptors in Oral versus Nonoral Drugs GA Training Runs

descriptor	bin start	bin stop	bin step
MW	74	1274	150
ClogP	−7.6	23.6	3.0
ON	0	31	6
OHNH	0	20	2
NRING	0	10	2
RTB	0	41	4
HBA	0	18	4
HALOGEN	0	7	1

3.3. Case Study Three: Oral Drugs. Average descriptor values for the Vieth et al. data set are shown in Table 8, and the bin range values used by the GA are shown in Table 9.

Based on the analysis of 1082 FDA approved oral drugs, it was concluded that the mean of the computed molecular properties of the drugs show no significant variation with respect to their launch date (20 year time span).[24] This consistency was taken to be indicative of the necessity of oral drugs to maintain these physical properties within a certain range, especially as these findings were also target independent. Thus, once trained, the GA can be applied for recognizing new generations of oral drug candidates. In contrast, new structural information about a target and an unexpected mode of interaction can dramatically influence the design of molecules for a drug class as was the case for renin.[33−35] The reliance on historical data makes the GA unwieldy in responding to new trends, thus, its application for the prediction of oral bioavailability is likely to be more suitable than its use for classification of drugs according to their therapeutic activity.

The results for only one experiment are reported. A total of five bins with consistently high weights and a low standard deviation (Figure 8) were observed. These bins capture the average properties of the MW, ON, NRING, HBA, and OHNH descriptors for oral drugs (Table 8 and Figure 8).

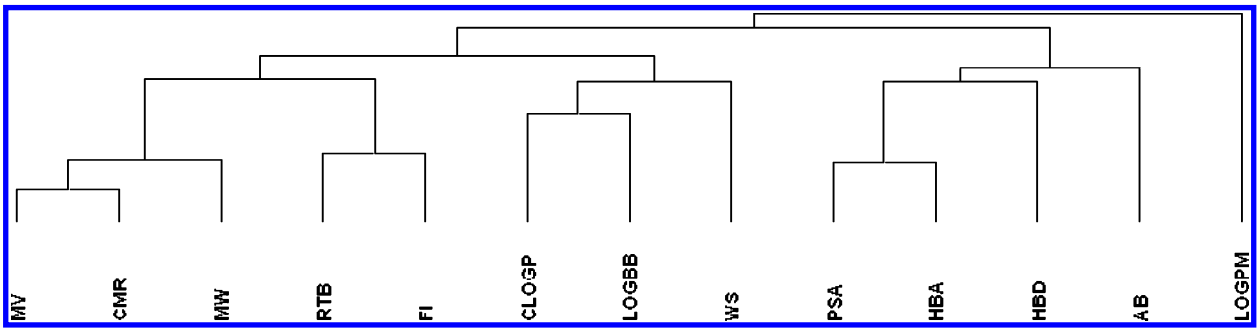

Figure 7. The results of clustering the descriptors for the COX inhibitors data set.

Figure 8. Results of analysis for consistency of weights. Bins having an average weight of ≥ 5 over five runs and a standard deviation of ≤ 1 are highlighted. Bins with lower weights were not considered to provide significant discrimination.

These ranges selected to be of significance by the GA simultaneously exclude several categories of nonoral drugs when their mean values are considered. For four out of five of these descriptors, injectable drugs have higher mean values than the ranges of the highlighted bins. The mean values of absorbent compounds also fall outside of the range of three out of the five descriptors with consistently high bin weights. Based on the descriptors and their ranges identified as being significant from the weight analysis, it is not expected that the GA can differentiate between oral and topical drugs. However, Vieth et al. note that many topical drugs probably have PK/PD profiles similar to those of oral drugs; their administration is topical to limit distribution to selected regions of the body.

The descriptors selected by the GA to be important for class discrimination (with the exception of NRING) combined with their low ranges are consistent with the conclusions of Vieth et al. who examined the means and percentiles of physical properties for marketed drugs. They concluded that with respect to other routes of administration, oral drugs tend to be lighter and have fewer hydrogen bond donors and acceptors and rotatable bonds than drugs with other routes of administration. Specifically, the average descriptor values for oral drugs found by Lipinski et al.,[3] Vieth et al.,[24] and Wenlock et al.[36] are in the range 300−344 for MW (range found to be significant by the GA is 224−374) and 4.5−5.5 (GA range 0−6) for ON, respectively. The average OHNH for marketed drugs is 1.8 (GA range 0−2).

3.4. Discussion: Average Descriptor Properties. A general observation that comes from analyzing the GA results with all three data sets (renin and COX inhibitors, plus oral and nonoral data compilation) is that bins with consistently high weights tend to occur for ranges of descriptors that reflect the mean values and their standard deviation, of the class being trained. For a particular descriptor, a molecule which has a property value that coincides with a bin that has a consistently high weight will obtain a higher score than a molecule that does not. However, its final score is the sum of all relevant descriptor contributions. There may be substantial overlap between the property ranges of the class being trained and the other data set such that a high score for a particular descriptor value may also become assigned to a portion of the molecules being discriminated against. Therefore, to achieve the optimal separation between the classes, the GA has the complex task of finding the correct balance between awarding higher scores for a particular descriptor and range of descriptor but lower ones for other ranges and other descriptors. Thus, the selection of descriptor bin step size impacts on the solutions that are generated.

In summary, it has been shown that judicious choice of bin step size is important for capturing the average properties of the class being trained, and this is expected to contribute to the discrimination between classes. This hypothesis will be validated using representative models from the GA training experiments with two external test sets. Also, a model that somehow reflects the properties of the training

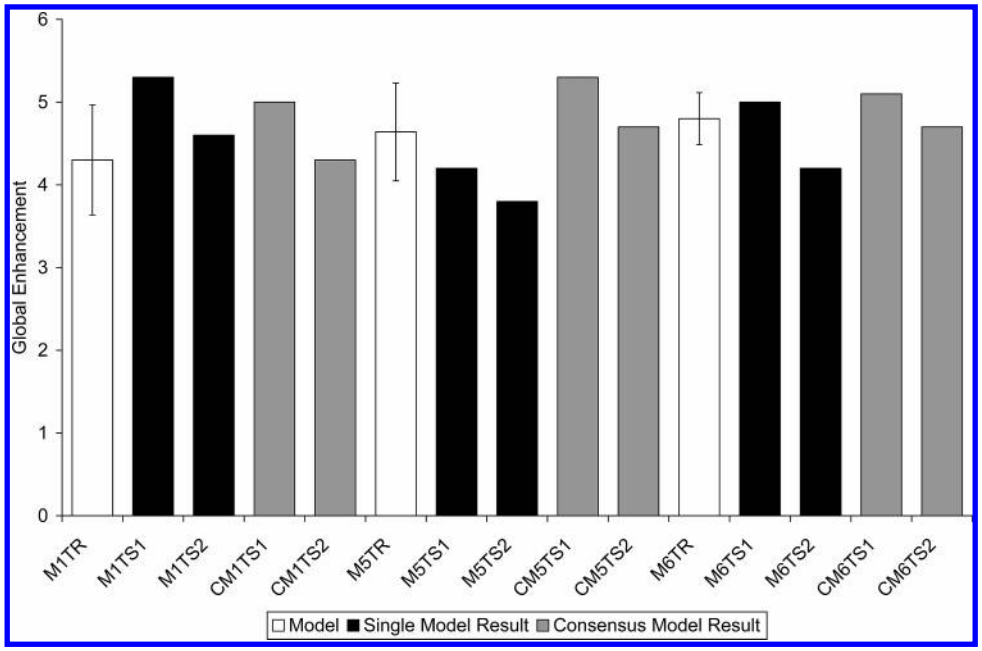

Figure 9. Graph showing the GE results for renin experiments 1, 5, and 6 resulting from the training models (M*m*TR), single best models with the test sets (M*m*TS*x*), and consensus models with the test sets (CM*m*TS*x*). $m = 1$, 5, or 6 and $x = 1$ or 2.

set is deemed more informative. The dangers of applying sequential filters have been addressed in the literature, and instead the consideration of the overall balance of properties has been advocated.[33] This is characterized by the GA models with bins with high weights for multiple descriptors that have markedly different values between molecule classes.

4. CONSENSUS AND SPLICE MODELS

4.1. Case Study One: Renin Inhibitors. The runs with the best GE values were selected from experiments 1, 5, and 6 for renin inhibitors. Experiments 1 and 6 both had the same 5 descriptors and with respect to experiment 1 the bin step size for MW in experiment 6 was optimized. Experiment 5 uses only 3 descriptors; it was designed to act as a control, and 2 descriptors were randomly removed. The reason for selecting these experiments was to determine if solutions with more descriptors were more effective in class discrimination, and if there was any benefit in bin step size optimization.

From the graph of GE results (Figure 9) it can be seen that the standard deviation of the training models for experiment 6 is low as compared to the other two experiments. Training models with a low standard of deviation for GE are indicative of the GA having identified a consistent set and range of descriptors as being important for discrimination. Additionally, experiment 6 is the only experiment to show an increase in all three performance measures for both test sets on going from the single model to the consensus model (Tables 10 and 11). This indicates that the consensus model of experiment 6 does discriminate between classes better than its best single model when the entire data set is considered and not just for a defined portion.

In contrast, the training models for experiment 1 have the highest standard deviation of the 3 selected experiments. The solutions of training models that have a high standard deviation for GE are liable to be more distinct than those that arise from training models with a lower standard of deviation. It is the only experiment not to show any improvement for any of the performance measures, for either test set, on going from the single model to the consensus model (Tables 10 and 11). The individual model has a number of high bin weights, particularly for the HBD and AB descriptors, but these high weights are dampened by the consensus model. Though its best single model has a very high performance in relation to the other experiments when applied to test set 1, this result is not favorable as it is not within the standard deviation of the five runs of the training model. This raises fears about the general applicability of

Table 10. Results of Applying Models Created in Renin Inhibitors Training Runs to Test Set 1[a]

model	MDE	GE	IE
experiment 1 run 4	11.3	5.3	4.8
experiment 1 consensus model	9.9	5.0	4.9
experiment 5 run 2	7.7	4.6	4.4
experiment 5 consensus model	7.6	5.3	4.7
experiment 6 run 5	9.5	4.7	4.5
experiment 6 consensus model	10.4	5.1	5.2
splice model	10.9	6.1	5.2

[a] For a consensus model, average bin weights from 5 five runs of an experiment were rounded to the nearest integer. These values were then used to construct a new model. The splice model is the consensus model of experiment 6 but with the AB descriptor bin weights replaced by those of the consensus model of experiment 3.

Table 11. Results of Applying Models Created in Renin Inhibitors Training Runs to Test Set 2[a]

model	MDE	GE	IE
experiment 1 run 4	11.0	4.6	4.2
experiment 1 consensus model	9.9	4.3	4.1
experiment 5 run 2	7.3	3.8	4.1
experiment 5 consensus model	7.7	4.7	4.4
experiment 6 run 5	9.3	3.9	4.1
experiment 6 consensus model	10.4	4.7	4.8
splice model	11.0	5.3	4.7

[a] For a consensus model, average bin weights from 5 five runs of an experiment were rounded to the nearest integer. These values were then used to construct a new model. The splice model is the consensus model of experiment 6 but with the AB descriptor bin weights replaced by those of the consensus model of experiment 3.

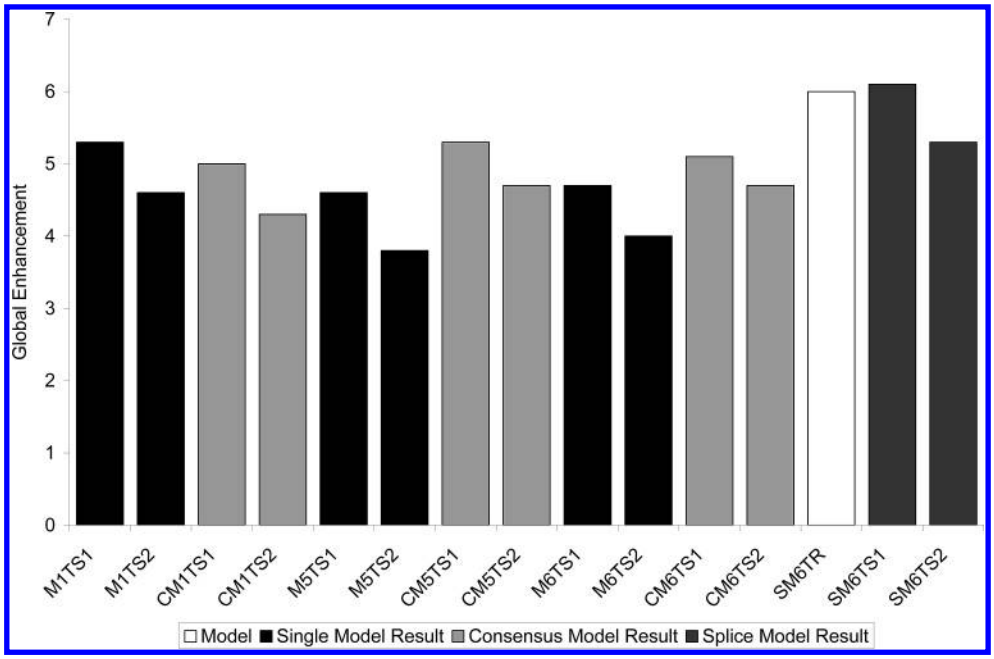

Figure 10. Graph showing the GE results for renin experiments 1, 5, and 6 resulting from the single best models with the test sets (M*m*TS*x*), consensus models with the test sets (CM*m*TS*x*), and splice model (SM*m*TR, SM*m*TS*x*). $m = 1$, 5, or 6 and $x = 1$ or 2.

such a solution. Also the best single model is less interpretable than the consensus model to which it contributes.

With respect to the GE and IE values (Tables 10 and 11) which result from applying the consensus models to the test sets, the three descriptor model (experiment 5) is equivalent to the five descriptor models (experiments 1 and 6). Though experiment 5 uses only three descriptors, two of these (HBD and AB) were identified as being important for the discrimination between renin and nonrenin inhibitors. By analyzing the bin weights of the training models, it was observed that a total of 8 bins in the HBD and AB descriptors have an average bin weight of 5, but these weights fluctuate between the models. However, this variability is reduced by construction of the consensus model, hence the competitive performance of the three descriptor models versus the five descriptor models. The high bin weights coincide with the average values and the upper limits of their standard deviation, for the HBD and AB descriptors for which a clear separation exists between renin and nonrenin inhibitors. The potential usage of information-rich subcomponents of models is discussed below in the section on splice models.

So in response to the initially posed questions, a model that has more descriptors does not necessarily provide better discrimination between classes than a model with fewer descriptors. Optimizing the step size of the MW descriptor on going from experiment 1 to experiment 6 has resulted in more consistent training models.

Splice Model. Experiments 3 and 6 used the same five descriptors but had different bin step sizes for the MW descriptors. It was observed that experiment 3 captured the extreme property ranges for the renin inhibitors but experiment 6 the average values. Hence, it was decided to combine the consensus models of experiments 3 and 6. The splice model was created by replacing the AB descriptor weights of the experiment 6 consensus model with those of the experiment 3 consensus model (Figure 10).

With respect to experiment 6, the creation of the consensus model impacts greatly upon the IE values obtained with test sets 1 and 2; there is a sharp increase as compared to the single model values (Tables 10 and 11, respectively). This effect is reinforced by the splice model, leading to a further increase in the MDE and GE values for both test sets. This latter increase in MDE values was anticipated. As compared to the consensus model, the AB descriptor of the splice model has high weights for bins which extend beyond the mean to the upper limit of the value for renin inhibitors. Since there is no overlap with nonrenin inhibitors for this range of descriptor, the score of some renin inhibitors should increase, but the score of the nonrenin inhibitors should remain relatively unaffected. This leads to an overall increase in the number of renin inhibitors being scored higher than nonrenin inhibitors. The final GE values represent between 74 and 85% of the maximum obtainable values for the test sets. More investigation is however necessary to determine whether the splice modeling concept is generally applicable or advantageous.

4.2. Case Study Two: COX Inhibitors. The runs with the best GE values for each of the COX inhibitor experiments were selected for validation with the test sets (Figure 11). The training models for experiment 1 give a lower standard deviation for GE values than those for experiment 2, but both of the consensus models perform better on the test sets than their best single models. As compared to experiment 1, experiment 2 identified an extra bin of the RTB descriptor as being important for discrimination. However, there is no clear separation between the classes for this value, hence the more variable performance of experiment 2's training models. As compared to the renin data set where three descriptors contribute to class discrimination, for the COX inhibitors only two do so, perhaps reflecting the fact that the COX data set is more heterogeneous than that of renin. The use of alternative descriptors with less overlap in physicochemical property space between the classes being discriminated against would likely lead to improved results.

Neither the COX inhibitors nor the oral drug data sets are suitable for application of the splice model method, due to

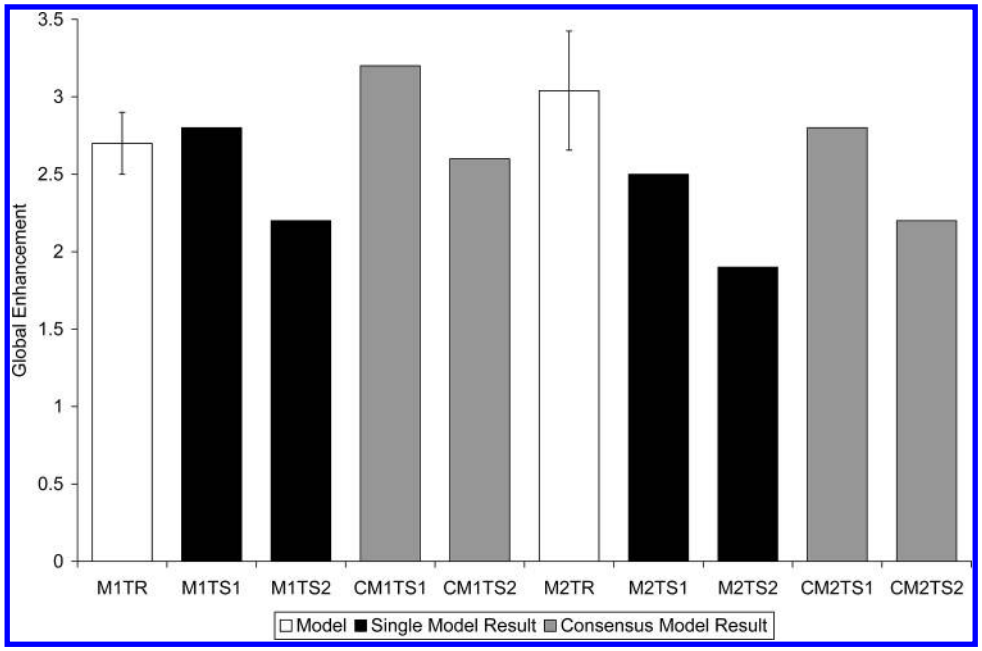

Figure 11. Graph showing the GE results for COX experiments 1 and 2 resulting from: the training models (M*m*TR), the single best models with the test sets (M*m*TS*x*), and consensus models with the test sets (CM*m*TS*x*). $m = 1$ or 2, and $x = 1$ or 2.

Table 12. Mean and Standard Deviation Values for the Results of Five Runs of the Tanimoto Similarity Method on Renin Inhibitors

reference structures	GE		IE	
	mean	SD	mean	SD
10	7.4	0.04	6.6	0.2
50	7.6	0.04	7.0	0.05

Table 13. Mean and Standard Deviation Values for the Results of Five Runs of the Tanimoto Similarity Method on COX Inhibitors

reference structures	GE		IE	
	mean	SD	mean	SD
10	2.6	0.9	4.9	0.7
50	8.1	2.1	7.3	0.7

the significant overlap of the property ranges of their descriptors with those of the class being discriminated against.

4.3. Comparative Studies. It is important to put the results of the GA modeling approach in context by comparing them with the results from alternative methods. The performance of the GA method with respect to the classification of the renin and COX inhibitors is compared with similarity searching. For the oral data set, the GA is compared with the alternative NBC and SVM classification methods.

Renin and COX Inhibitors. The results of the evaluation of each similarity search with varied numbers of reference structures are presented for renin and COX inhibitors in Tables 12 and 13, respectively. The renin inhibitor search consistently and selectively recognizes this class versus the other classes (Table 12). GE values increase from 97 to 100% of the maximum obtainable value, and IE values increase from 87 to 92% of the maximum obtainable value on going from 10 to 50 reference structures. For the COX inhibitors, on going from 10 to 50 reference structures, IE values increase from 35% to 51% of the maximum obtainable, and for GE values the increase is from 18% to 57% of the maximum obtainable. Hence, more diverse series benefit from having additional molecules in the reference set in order to better capture their diversity. Using the GA method, the GE values approach 74−85% of the maximum value obtainable for the renin inhibitors (splice model). For the COX inhibitors only 20−25% of the maximum GE value is obtained (consensus model from experiment 1).

The relative difference between the results of the renin and COX experiments is due to the homogeneity and heterogeneity, respectively, of these data sets. Indeed, Hert et al.[26] report the mean pairwise Tanimoto similarity of the renin inhibitors as 0.573 (SD 0.106), while the similarity is only 0.268 (SD 0.093) for the COX inhibitors. This demonstrates that the COX data set contains much more structurally diverse molecules, and therefore it will be difficult using structural similarity methods alone to develop effective models of discrimination.

More COX inhibitors are misclassified by the GA method than are correctly classified. In an attempt to gain a better understanding about the poorer performance of the GA method with the COX data set relative to that of the similarity method, we enumerated the number of reduced ring system (RRS) present in the top *NACT* positions of the GA (consensus model experiment 1) and similarity (50 reference structures) output data. A RRS is an unordered ring system containing information about the number of rings and internal edges but not the order of connectivity.[37] It was found that 67% of COX inhibitors in the top *NACT* positions misclassified by the GA (consensus model experiment 1) had RRS in common with the entire data set of COX inhibitors. Thus, the GA is adept at being able to recognize COX-like RRS, but the similarity method is more effective in identifying true COX inhibitors. This is expected since the molecular fingerprint descriptors are more information-rich than the limited set of descriptors used in the GA studies. If two molecules have the same RRS, the similarity method will assign the higher score to the molecule with a connectivity that is most similar to any of the reference molecules. In the case of the GA, molecules with identical RRS are judged solely on the balance of their physiochemical properties and

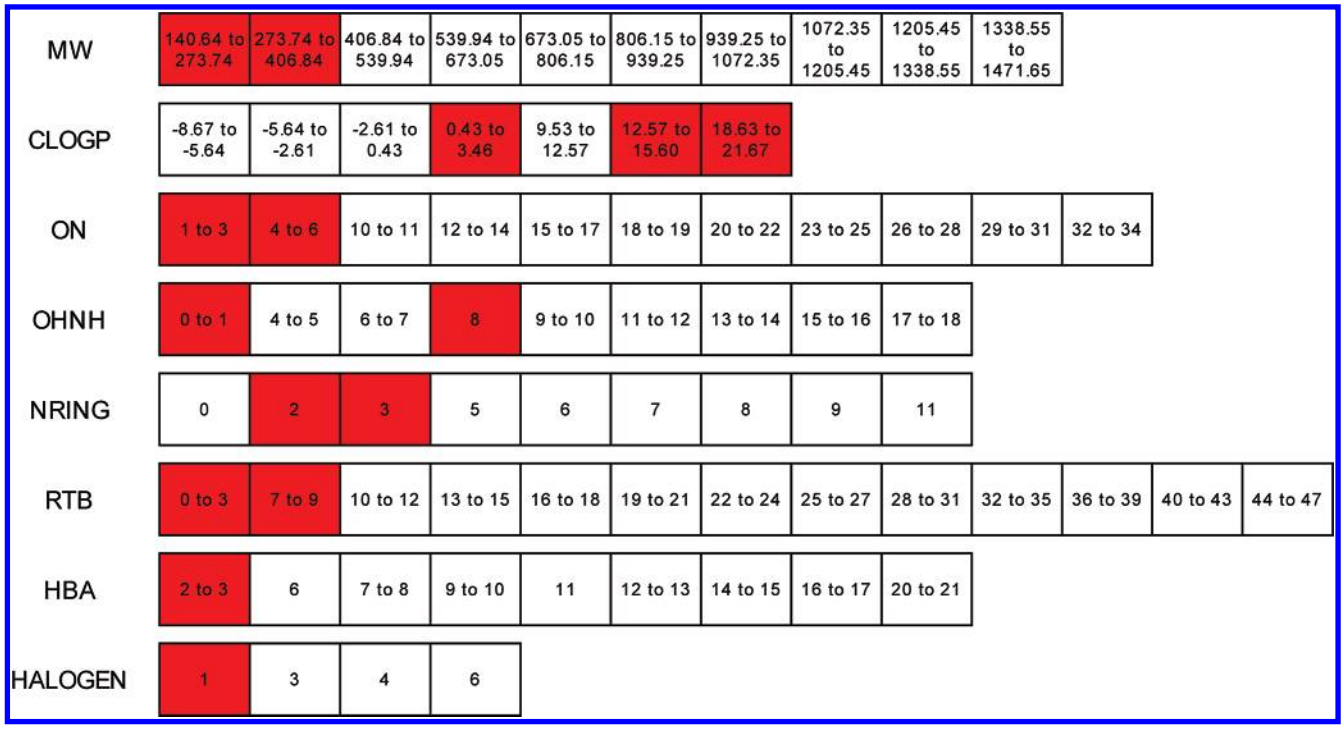

Figure 12. Results of analysis for consistency of weights with the NBC method. The bins that are most significant for discrimination are highlighted. Bins with lower weights were not considered to provide significant discrimination.

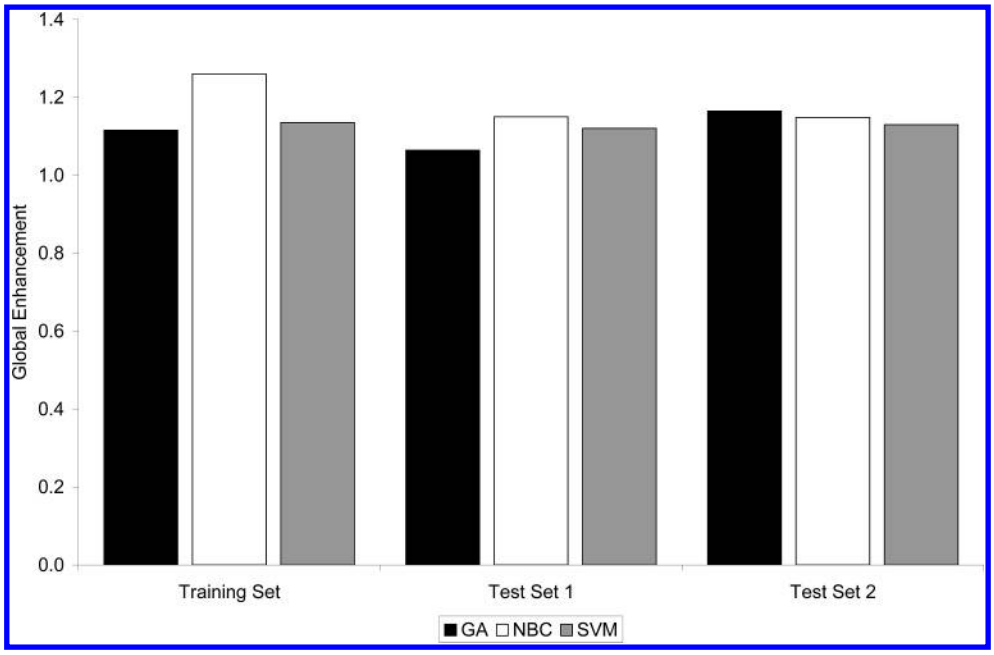

Figure 13. The global enhancement scores for the training set and test sets 1 and 2 with the GA, NBC, and SVM modeling methods, respectively. The maximum possible global enhancement value for this data set is 1.44.

not on the order of connectivity of the rings. Thus, the use of only physicochemical descriptors by the GA is not only the source of its weakness but also its strength as is discussed below with respect to its performance with the oral data set.

Oral Data Set. To gauge an accurate comparison with regard to the results for the oral data set we applied two of the most often used DA methods in the literature: the NBC and SVM, respectively.

The NBC statistical learning method uses both the prior probability of class membership together with the local likelihood according to descriptor vectors to classify new data points. The prior probability is calculated based on the global probability of a data point belonging to a particular class according to the number of data points that represent each of the classes. The local neighborhood of the new data point is considered in terms of the descriptor vector and assigning an overall probability of that data point belonging to a particular class. From these two probabilities, a single probability of membership to each class can be calculated.

SVMs use a kernel-based learning method to locate a hyperplane that best separates the data under consideration into two classes with a maximum margin. The application

of kernels in SVMs permits a mapping from the input space to the feature space where they are linearly separable. The SVM implementation applied in this study is SVM*light*.[38,39]

Using the same training and test set partitions together with the same descriptor sets as for the GA models, an accurate indication of the relative benefits of each of the methods is expected. The GA and NBC methods provide a level of transparency that permits their interpretation; however, as mentioned earlier, the SVM is somewhat more difficult to apply in the interpretive sense.

With respect to GE and IE values for the training and both test sets, the GA, NBC. and SVM methods perform equally well (Figure 13); however the results from the GA are more comprehensible. The descriptor ranges with consistently high weights as identified by the GA mirror the mean values of the oral drugs better than the descriptors ranges with high probabilities arising from the NBC method. Although the mean OHNH descriptor value is 1.8, the NBC descriptor range with the highest probability is 8 to 8. Thus, it is not apparent why this descriptor value is important for discrimination between the data sets. However, even though the average ClogP descriptor value for the oral data set is 2.3, the NBC ranges are 12.57−15.60 and 18.63−21.67. These ranges would be expected to exclude all nonoral drugs based on the minimum and maximum values of the 0−100% percentiles. With the GA method, the ClogP descriptor is not associated with any consistently high weights. The ClogP range with the third highest NBC probability is 0.4−3.5 which even though it captures the average value for the oral data set does not allow for any discrimination against the nonoral data set. Similarly, the significance of the RTB descriptor bins with high NBC probabilities is not clear. The GA method more precisely captures the upper bound of a descriptor range. For instance the mean MW for oral drugs in this data set is 342. The GA upper bound for this descriptor is 374 which excludes both the absorbent and injectable data sets which have higher MW mean values. However, the NBC upper bound for the MW descriptor is 406.8, only allowing for the discrimination against the injectable data set. Based on the above analyses and comparisons, the results from the GA method are more readily interpretable than those from the NBC method. In addition, the NBC model shows a tendency to overfit with regard to these data sets, with the discrepancy in GE being more distinct between the training and test sets with the NBC as opposed to either the GA or the SVM. The summary of the bin weights for the NBC method are provided in Figure 12.

5. CONCLUSIONS

We have applied a GA to the problem of differentiating between drug classes. In two instances the class being trained acted at a specific drug target, and in the third case it possessed particular PK/PD properties. This reflects the flexibility of GAs in being able to distinguish between categories of molecules based on descriptors that are not purely structure related. By analysis of bin weights over several runs of a GA with fixed parameters and descriptors, it was observed that bins with consistently high weights tended to occur in ranges of a descriptor where there are differences around the mean value of the class being trained and the rest of the data set. Solutions which reflect the average properties of the class being trained are deemed to be more consistent than those that do not when used with external test sets. The selection of appropriate bin step sizes is essential for the generation of such solutions.

Two novel approaches for exploiting and enhancing the information contained in multiple good solutions have shown marked improvements on the results obtained from single optimal solutions. The consensus weighting method demonstrates a way of reducing the variability that can be apparent in individual solutions that may have abnormally high weights for particular bins. The consensus model also emphasizes the importance of bins that have consistently high weights over a number of runs leading to a more consistent and interpretable model. This latter facet is particularly important, as the results of such discriminant analyses are likely to be presented to medicinal chemists. The splice models allow for the replacement of bin weights for particular descriptors that have been observed to be beneficial in the discrimination of molecular classes. A splice model is generated from multiple complementary models giving a single model that emphasizes the most important subcomponents of each of the models. Each descriptor could be used as a discriminator in isolation; however, we expect that this may lead to overfitting in some cases and due caution should be expressed.

Here we have reported a novel predictive modeling protocol that is both designed to be interpretable and results in models that are highly competitive when compared with less interpretable methods. In summary, the similarity searching method as applied here is superior in discriminating power to the models evolved with the GA for the COX (diverse) inhibitor data set. Although molecular fingerprints are highly informative molecular descriptors, the information they encode is generally inaccessible resulting in the methods that apply them being largely uninterpretable. However, by using a small set of descriptors that are intuitive to the medicinal chemist, we can discover ranges of molecular properties that are directly interpretable by the chemist. The issue here remains a tradeoff between highly effective descriptors versus highly interpretable descriptors. The application of these methods can be summarized as belonging to two distinct styles of predictive modeling: models for use in predictions against models to be used for diagnostics. In the predictive modeling mode, one often requires a method that is highly predictive with little or no regard made to the interpretability of the resultant model other than indirectly through the molecular predictions. However, in the low-throughput case, it is important to be able to present transparent information regarding the models generated to encourage confidence in our methods due to their interpretability. There was less difference in the performance between the similarity and consensus GA methods for the renin inhibitor (homogeneous) data set. In terms of class discrimination, the consensus GA method performed as well as two alternative well-established DA methods using the oral data set, but the GA gives results that are easier to interpret and comprehend.

We envision that this modeling protocol will be highly effective when working in close contact with medicinal chemists to inspire confidence in our modeling techniques while also retaining high predictive power.

ACKNOWLEDGMENT

The authors would like to thank Drs. Claus Ehrhardt, Edgar Jacoby, Stephen P. Jelfs, Richard A. Lewis, Nikolaus Stiefl (all Novartis Institutes for BioMedical Research), Dr. Jérôme Hert (University of California, San Francisco), and Prof. Peter Willett (University of Sheffield) for various discussions and support in this research. The authors would also like to thank the anonymous reviewers for constructive comments that led to the improvement of this manuscript. M.G. was funded by the Engineering and Physical Sciences Research Council and the Novartis Institutes for BioMedical Research.

REFERENCES AND NOTES

(1) Böhm, H.-J.; Flohr, A.; Stahl, M. Scaffold Hopping. *Drug Discovery Today: Technol.* **2004**, *1*, 217−224.
(2) van de Waterbeemd, H.; Gifford, E. ADMET *in silico* Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.
(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and developmental settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.
(4) Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug Delivery Rev.* **2002**, *54*, 255−271.
(5) Lajiness, M. S.; Vieth, M.; Erickson, E. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 470−477.
(6) Aitchison, J.; Aitken, C. G. G. Multivariate Binary Discrimination by the Kernel Method. *Biometrika* **1976**, *63*, 413−420.
(7) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(8) Wilton, D. J.; Willett, P.; Lawson, K.; Mullier, G. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.
(9) Pirard, B.; Pickett, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(10) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of Extremely Noisy High-Throughput Screening Data Using a Naïve Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.
(11) The MDL Drug Data Report (MDDR) database is available from Elsevier MDL at http://www.mdl.com/products/knowledge/drug_data_report/.
(12) The MDL Comprehensive Medicinal Chemistry (CMC) database is available from Elsevier MDL at http://mdl.com/products/knowledge/medicinal_chem/.
(13) The World Drug Index (WDI) database is available from Thomson Scientific at http://scientific.thomson.com/products/wdi/.
(14) The Available Chemicals Directory (ACD) is available from Elsevier MDL at http://www.mdli.com/products/experiment/available_chem_dir.
(15) The SPeicherung und REcherche Strukturchemischer Information (SPRESI) database is available from InfoChem, GmbH at http://www.spresi.com/.
(16) Kubinyi, H. Drug research: myths, hype and reality. *Nat. Rev. Drug Discovery* **2003**, *2*, 151−154.
(17) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular Properties that Influence Oral Drug-like Behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 470−477.
(18) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish Between "Drug-like" And "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(19) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(20) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(21) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates From Docking Databases of Three-dimensional Structures Into Proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.
(22) Ginn, C. M. R.; Turner, D. B.; Willet, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 23−37.
(23) Ajay. On Better Generalization by Combining Two Or More Models: A Quantitative Structure−activity Relationship Example Using Neural Networks. *Chemom. Intell. Lab. Styst.* **1994**, *24*, 19−30.
(24) Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224−232.
(25) Beresford, A. P.; Segall, M.; Tarbit, M. H. *In Silico* Prediction of ADME Properties: Are We Making Progress? *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 36−42.
(26) Hert, J.; Willet, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A.; Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(27) Ertl, P.; Mühlbacher, J.; Rohde, B.; Selzer, P. Web-based Cheminformatics and Molecular Property Prediction Tools Supporting Drug Design and Development at Novartis. *SAR QSAR Environ. Res.* **2003**, *14*, 321−328.
(28) CLOGP and CMR are available from BioByte at http://www.biobyte.com/.
(29) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(30) CORINA is available from Molecular Networks, GmbH at http://www.mol-net.com/.
(31) Clark, D. E. Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 2. Prediction of Blood-Brain Barrier Penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.
(32) PipelinePilot is available from SciTegic, Inc. at http://www.scitegic.com/.
(33) Oefner, C.; Binggeli, A.; Breu, V.; Bur, D.; Clozel, J.-P.; D'Arcy, A.; Dorn, A.; Fischli, W.; Grüninger, F.; Güller, R.; Hirth, G.; Märki, H. P.; Mathews, S.; Müller, M.; Ridley, R. G.; Stadler, H.; Vieira, E.; Wilhelm, M.; Winkler, F. K.; Wostl, W. Renin inhibition by substituted piperidines: a novel paradigm for the inhibition of monomeric aspartic proteinases? *Chem. Biol.* **1999**, *6*, 127−131.
(34) Vieira, E.; Binggeli, A.; Breu, V.; Bur, D.; Fischli, W.; Güller, R.; Hirth, G.; Märki, H. P.; Müller, M.; Oefner, C.; Scalone, M.; Stadler, H.; Wilhelm, M.; Wostl, W. Highly potent renin inhibitors due to induced fit adaptation of the active site. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1397−1402.
(35) Güller, R.; Binggeli, A.; Breu, V.; Bur, D.; Fischli, W.; Hirth, G.; Jenny, C.; Kansy, M.; Montavon, F.; Müller, M.; Oefner, C.; Stadler, H.; Vieira, E.; Wostl, W.; Märki, H. P. Piperidine-renin inhibitors. Compounds with improved physiochemical properties. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1403−1408.
(36) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.
(37) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144−6159.
(38) SVM^{light}, version 6.01. http://svmlight.joachims.org/.
(39) Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*; Kluwer Academic Publishers: Boston, MA, 2001.

CI050529L