Lead Hopping Using SVM and 3D Pharmacophore Fingerprints

Jamal C. Saeh,*,† Paul D. Lyne,*,† Bryan K. Takasaki,‡ and David A. Cosgrove§

Cancer Discovery, AstraZeneca R&D Boston, 35 Gatehouse Drive, Waltham, Massachusetts 02451, Global Science and Information, AstraZeneca R&D Boston, 35 Gatehouse Drive, Waltham, Massachusetts 02451, and Cancer Discovery, AstraZeneca, Alderley Park, Cheshire SK10 4TG, United Kingdom

Received August 30, 2004

The combination of 3D pharmacophore fingerprints and the support vector machine classification algorithm has been used to generate robust models that are able to classify compounds as active or inactive in a number of G-protein-coupled receptor assays. The models have been tested against progressively more challenging validation sets where steps are taken to ensure that compounds in the validation set are chemically and structurally distinct from the training set. In the most challenging example, we simulate a lead-hopping experiment by excluding an entire class of compounds (defined by a core substructure) from the training set. The left-out active compounds comprised approximately 40% of the actives. The model trained on the remaining compounds is able to recall 75% of the actives from the "new" lead series while correctly classifying >99% of the 5000 inactives included in the validation set.

1. INTRODUCTION

Drug discovery in the pharmaceutical and biotechnology industries is seeing increased integration of computational chemistry and cheminformatic approaches. The integration of these methods has resulted in an increase in the number of reports of successful applications of these approaches to drug discovery and suggests that early stage drug discovery is benefiting greatly from the integration of experimental and in silico approaches.[1,2]

One particular aspect of drug discovery that can benefit from the application of computational approaches is lead hopping.[3−5] Often in a drug discovery program, the project is faced with many challenges relating to the biological activity, selectivity, intellectual property, pharmacokinetic profile, and toxicity of a specific lead series.[6,7] Overcoming these issues is often only achieved by identifying a new chemical series that retains the desirable properties of the original series but lacks the liabilities associated with that series.

There are a number of computational approaches that have been applied successfully in the past to lead hopping. These include structure-based virtual screening,[8] similarity searching,[9−11] pharmacophore screening,[12,13] and the shape-based screening of chemical databases.[14−16]

There have been several recent reports of successful structure-based virtual screens for generating novel hits for specific biological targets;[6,7] however, this approach suffers from the limitation of requiring a high-resolution three-dimensional structure of the target. As a consequence, this approach may only be applied to a subset of the pharmacologically relevant targets.

A more common approach to lead hopping involves the similarity searching (2D or 3D) of compound databases, relying on the premise that molecules that are chemically similar will exhibit similar biological profiles. The usefulness of 2D approaches for lead hopping are often limited by the typical reliance of the algorithms on chemical connectivity to define the similarity metrics employed,[17] whereas recent studies have not found 3D methods to offer an advantage over 2D methods for identifying active compounds.[18,19]

Methods that focus on the shapes of molecules have received attention recently for the potential application to lead hopping. These approaches are based, usually, on the shapes of molecules with known biological activities, or fragments of these molecules, but they may also be used in combination with pharmacophore descriptions (such as the "shrink-wrap" approach).[20]

In this paper, a rapid and powerful method for lead hopping that utilizes support vector machines (SVMs)[21−23] to develop a biological activity model for a specific target or class of targets is presented. The information used by the SVMs includes the three-point pharmacophore fingerprints of compounds with known activities and inactivities against the target or target-class of interest. The pharmacophores are generated from conformationally enumerated (to account for ligand flexibility) 3D structures of the compounds.

A number of applications of SVM to computational chemistry have been published.[24,25,45] The purpose of the present study is to investigate the use of SVMs to produce general models that will enable "lead hopping", that is, moving from one chemical series to another while retaining activity in the assay of interest. The ability of the model to lead hop is determined primarily by the choice of descriptors that represent the compounds. Daylight fingerprints, for instance, have been shown over the years to be excellent at quantifying chemical similarity but are not suitable, in this case, where one is hoping to find chemically different

* To whom correspondence should be addressed. Tel.: (781) 839-4000. Fax: (781) 839-4650. E-mail: jamal.saeh@astrazeneca.com (J.C.S.), paul.lyne@astrazeneca.com (P.D.L.)
† Cancer Discovery, AstraZeneca R&D Boston.
‡ Global Science and Information, AstraZeneca R&D Boston.
§ Cancer Discovery, AstraZeneca, Alderley Park.

10.1021/ci049732r CCC: $30.25 © 2005 American Chemical Society
Published on Web 07/06/2005

molecules within the same activity class. Fingerprints based on 3D pharmacophores were chosen as they describe the geometrical relationship between the pharmacophore points in a molecule without directly referring to the underlying chemical structure. This paper focuses on the application of these methods to various G-protein-coupled receptor (GPCR) targets with a particular emphasis on lead hopping. The models have been tested against progressively more challenging validation sets where steps are taken to ensure that compounds in the validation set are chemically and structurally distinct from those of the training set. In the most challenging example, an entire class of compounds (defined by a core substructure) is held out of the training set. The model trained on the remaining compounds is able to recall 75% of the actives from the "new" lead series while correctly classifying >99% of the 5000 inactives included in the validation set.

The method described here has utility both in early and late stages of drug discovery projects. In the early phases, the models could be used as in silico screens for hit identification or adapted to the analysis of high-throughput screening data. In the latter stages of drug discovery, the models could be used to drive lead hopping.

2. METHODS

2.1. Computational Methods. 2.1.1. Data Set. Data in this study were taken from two sources, namely, the MDL Drug Data Report (MDDR) database and in-house screening data. Compounds were extracted from the MDDR database using the DDR activity index corresponding to GPCRs of interest. There are over 130 000 unique compounds in the MDDR, of which 1500 are listed as having some form of activity against 5HT1A, making it the largest single receptor set represented in the database.

For binary classification purposes, it is also necessary to include information about the inactive compounds. For the in-house screens, this was directly available, but for the MDDR data, it is somewhat more problematic. The assumption was made that if a particular DDR activity index was not quoted for a compound, the compound was not active for that activity type. This is not necessarily correct, however, because in fact the only indication in the database is that activity was not tested for that compound in that screen. It may well be the case that, were the compound to be tested, it might be active. This assumption could introduce some false negatives into the data at the model-building phase.

2.1.2. Data Preparation. The original source of all the compounds considered was in SMILES format.[26] These compounds were initially profiled using Leatherface (an internal application) to assign appropriate protonation states to the molecules and to generate common tautomers where necessary. Leatherface is a molecular editor that employs the Daylight programming toolkits. Leatherface uses a set of rules specified as SMARTS,[27] derived from in-house medicinal and physical chemistry knowledge, to modify molecular connection tables. Leatherface is also capable of enumerating forms that are appropriate for representing relatively unbiased equilibria. A 3D version of the database was generated using Corina,[28,29] with explicit enumeration of stereocenters (generating a maximum of eight stereoisomers per molecule). A conformational version of the database was then generated using the program Omega.[30] A maximum of 1000 conformers (GP_NUM_OUTPUT_CONFS) were generated for each molecule in the database, with a root-mean-square cutoff of 0.6 Å (GP_RMS_CUTOFF) to define geometrically distinct conformers and an energy threshold of 5.0 kcal/mol (GP_ENERGY_WINDOW), above which conformers were discarded.[31] Three-point pharmacophores were assigned to each molecule using Loob, an in-house program that calculates three- and four-point pharmacophore fingerprints. The pharmacophore points are defined using combinations of SMARTS definitions specified at runtime, allowing great flexibility in the types of points represented in the fingerprints. Six chemical features and six binned distances were used to define the triplets. The features were hydrogen-bond donor or acceptor, positive or negative charge, rings, and hydrophobic regions. The distances used in all the examples in this work were at fixed bins 0−4.5, 4.5−7, 7−10, 10−14, 14−19, and 19−24 Å. The structures are read from a conformationally expanded database, such that each molecule is treated as an ensemble of rigid conformations. The pharmacophore triplets and quadruplets are uniquely encoded using the algorithm of Abrahamian et al.[32] It must be emphasized that all the models in this work were built using only triplets. The fingerprint length for the three-point pharmacophore was 10 152 bits.

2.2. Support Vector Machines. The SVM is a binary classification technique developed by Vapnik and his group at AT&T Bell Laboratories.[33−35] SVM has become very popularbecause of its excellent generalization capacity. It is based on structural risk minimization and aims to balance the tradeoff between bias and variance. SVM is different than empirical risk minimization algorithms, such as neural nets. Whereas neural nets seek to minimize the errors over the entire training set, SVM attempts to place a boundary using support vertors (example in the training) and ignores those examples in the training that are outside the boundary. These are often a small fraction of the training data and, as such, allow a SVM model to be less prone to over training while maintaining an excellent degree of generalizability. However, like neural nets, SVM is a "black-box" approach, and compared to other approaches (hierarchal clustering and other QSAR approaches), it is more difficult to interpret the resulting models.

The generalization of a model can be thought of as the measure of a learner, in this case, a computer algorithm, to abstract an object to its proper class. Whereas chairs may be made of wood, a good learner can recognize a chair even if it was made of metal, while refraining from calling a tree a chair just because it contains wood. A learner that is unable to generalize will require that each chair be included in its training set before it can correctly identify any chair as a chair. The success of the algorithm to learn and generalize is a function of the inherent robustness of the algorithm and the descriptors used to abstract the object. In this application, the descriptors that we have decided to use are three-point pharmacophore fingerprints that abstract a molecule in terms of the functional features it contains and their orientation in three-dimensional space.

For SVM, the input requirements are a set of descriptors, represented as a vector. Feature selection using term weighting of the descriptors was initially considered. We also

considered feature frequency to be included in the vectors. In the end, a simple binary representation of the three-point pharmacophores present in the compound was used. It was not clear that term weighting or reducing the size of the descriptors yielded better classifiers. In contrast, classifiers whose input vectors are normalized to length 1 are significantly better. In the learning phase, each compound is assigned to a class, namely, active or inactive. For classification purposes, active compounds are assigned a +1 value, whereas inactive compounds are assigned −1. The SVM calculates a hyperplane that defines a surface with active examples on one side and inactive examples on the other. In some cases, the training examples cannot be separated linearly, and nonlinear transformations, via kernel functions, may be employed to render the examples linearly separable.

There are numerous hyperplanes that may separate the data in this manner. The SVM uses a dual optimization algorithm that simultaneously minimizes the training error, the correct classification of the actives and inactives, while maximizing the margin, the separation between the two classes and the hyperplane. The placement of the hyperplane defines the zero threshold. It is assumed that a hyperplane with a larger margin gives a more general classification than one with a smaller margin. However, it should be noted that the general nature of the model depends not only on the performance of the classification algorithm but also on the diversity and relevance of the descriptors chosen to represent the problem.

Once the hyperplane and support vectors have been developed, the SVM may be used in classification mode. The set of test vectors is input, and the SVM assigns a positive or negative value to each one, depending on which side of the hyperplane it falls on and how far it is from the hyperplane. By varying the absolute position of the threshold, the size of the list of predicted actives may be varied. Sorting the classified list by descending SVM output essentially prioritizes the list in decreasing confidence of classification.

In the library design problem addressed in this study, the descriptors are some representation of molecular properties and could be binned physical properties, molecular fingerprints such as those produced by the Daylight fingerprint toolkit[36] or Tripos' Unity software,[37] MDL MACCS keys,[38] or, as in our case, a normalized representation of a molecule's 3D pharmacophore fingerprints. The model classifies the molecule represented by the descriptor vector as active or inactive in a particular biological assay.

2.3. Model Generation. The models were built using SVM*light* [39,40] and a variety of Perl and C-shell scripts that automate the process of model generation, testing, and validation.

Model generation involved three steps. First, the active compounds were randomly split into training, testing, and validation sets in the ratio 80:15:5. The number of inactives included in each set was determined by the number of actives available (see Figure 1). Because of the overabundance of inactives compared to the actives, an unequal ratio of actives to inactives was used. For the testing and training sets, enough inactives were included in each set to give an active/inactive ratio of 1:10, whereas the ratio was increased to 1:300 for the validation set. The rationale behind the 1:10 ratio for the training and test sets was based on the assumption that inactive compounds carry important 3D pharmacophoric information about the ligand−receptor interaction. A 1:1 ratio of actives to inactives essentially throws away experimentally measured activity. The more information the algorithm is given about what promotes inactivity, the better the model is at identifying the inactives and the greater the enrichment. This is, of course, provided that the SVM is capable of offsetting this bias in building the model. In SVM*light*, this bias is controlled by the cost factor.

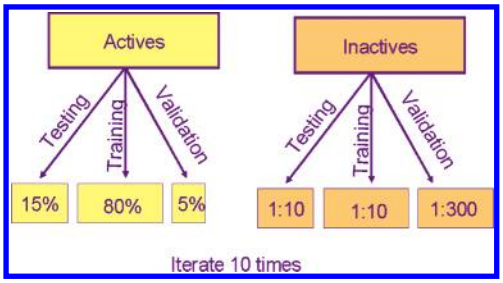


Figure 1. Splitting of the data into training, testing, and validation sets by taking random subsets of the original data. The number of inactives in the original data is usually 3−4 orders of magnitude greater than the number of actives. All of the actives are split between the three sets. Once the number of actives in each set is determined, the appropriate number of inactives is added to each set to give the active/inactive ratios shown.

The optimization procedure in SVM*light* is driven by a variety of parameters adjustable by the user. The parameter space of SVM was explored by changing the kernels[23] between linear, polynomial, and radial basis functions (RBF). In the polynomial case, the order (-d option in SVM*light*) was adjusted between second and fourth order, and in the case of RBF, a variety of kernel widths (γ) were used, $\gamma \in \{0.01, 0.03, 0.1, 0.3, 1.0, 2.0\}$. In text classification, Joachims observed that better classification was achieved with smaller values of γ.[40] It was not clear what value of γ to pick in a chemical application. In a few instances, using the default value ($\gamma = 1$), none of the active compounds were recalled. All the RBF models built for 5HT2c failed when tested against an external validation set obtained from the MDDR. These models had previously shown promise against the left-out test and validation sets. The same problem with the other kernels used was not encountered, namely, the linear and polynomial kernels.

The parameter C adjusts the level of training errors tolerated. Whereas some guidelines have been suggested in Joachims' work on text mining,[40] there is no guidance as to what is an appropriate value for classification for a chemical application. SVM*light* computes a default value according to $C = n/[\sum_{i=1}^{n}(\vec{x}_i\vec{x}_i)]$,[39] where n is the number of compounds in the dataset and $\vec{x}_i$ is the vector of the ith compound. In this work, models are generated using the default value of C as well as other values, which were chosen by trial and error. The C values were default, 0.1, 0.5, 0.75, 1, 2, 3, 5, 7, 9, 10, 20, 40, 50, and 100.

Because of the unequal distribution of actives and inactives in the training set, the cost factor (-j option in SVM*light*) was adjusted to penalize the misclassification of actives 10 times more than the inactives.

For each combination of parameters (options -c, -j, -t, and -gamma), 10 models were generated for 10 random splittings of the data. The performance of a set of parameters was

Table 1. Summary of the Performance of the Various Models on the Held-Out Validation Sets[a]

	TP	FP	TN	FN	sensitivity (%)	specificity (%)	new hit rate	old hit rate	EF
β1Adren	45	2791	57 209	23	66.18	95.35	1.59	.11	14.02
5HT2c	156	618	19 382	72	68.42	96.91	20.16	1.13	18.89
D2	34	1539	128 424	7	82.93	98.82	2.16	0.03	68.54

[a] The statistics were extracted from the usual confusion matrix, where sensitivity = TP/(TP + FN) and specificity = TN/(FP + TN). The hit rate is defined as precision(A)/prevalene(A) = TP/(TP + FP)/[(TP + FN)/*N*], where *N* is the total number of compounds in the validation set. The enrichment factor (EF) is defined as the ratio of the new library hit rate to the old library hit rate.

characterized by the model average performance in the 10 random splittings of the data. After each trial, the training and testing actives and inactives are joined and then randomly split into new training and test sets. In contrast, the same validation actives were used in each of the 10 cross validation sets. Because, typically, there is a large number of inactive compounds not used in the training or testing, we randomly select a new set of inactive compounds for each validation while maintaining the 1:300 split between the actives and inactives. Therefore, the validation sets, in maintaining the same list of actives across all 10 validation sets, provide us with a good measure of the capacity of the models to recall the same actives and the overall ability of the models to generalize to a large number of unknown inactives.

2.4. Automatic Model Selection. It is an understatement to say that models can be easily generated to fit training data. In fact, hundreds of models are generated for each receptor. It is much more difficult to decide on a criterion that measures the goodness of fit. The output from the SVM training data is a measure of the model in terms of units of margin, but this is merely an indication of how well the hyperplane classifies the training data. Table 1 shows some of the ways in which the performance of the model against a test set might be measured. True positives (i.e., active compounds correctly predicted as active by the model) are denoted by TP, false positives by FP, and analogously, TN and FN are used for true negatives and false negatives, respectively. For some applications, the best model is the one that optimizes the number of true positives; in others, it optimizes the number of true negatives. Frequently, the different measures are in disagreement. For instance, the sensitivity, which measures the fraction of actives correctly predicted as such, and the specificity, which is the fraction of inactives correctly predicted, may be at odds. One can optimize the correct prediction of active compounds, for instance, but generally only at the expense of specificity; predicting all compounds to be active will correctly predict all active compounds, but the model will be of limited use in practice. The enrichment factor, EF, as defined in Table 1, suffers from similar problems. For instance, in a collection of 10 000 compounds containing 10 actives, a model that correctly classifies 1 of the active compounds and predicts the remaining 9999 as inactive will have an EF of 1000, whereas one that correctly classifies 9 of the 10 actives and 9090 of the 9990 inactives has an EF of 9.9. It could be argued that the second model is more attractive because it is able to recover 9 out of the 10 actives even though it has a much lower EF.

Several statistical metrics have been suggested that can simultaneously measure sensitivity and specificity. Lewis and Gale[41] suggested the F_β measure as a metric that takes into consideration a weighted measure of precision and sensitivity. The widely used Kappa statistics[42] compare the predicted values with what may be expected by chance. It is, however, a measure that is highly influenced by prevalence. Güner[43] suggested a goodness of hit score, GH. Similar to the F_β measure, the GH score is a linear combination of sensitivity and precision with a fixed weighting. He also defines a normalized GH score (GH_n), which introduces a normalization constant to account for the size of the dataset. Because the prevalence is fixed in the training and testing sets, these three measures are comparable. In this work, a modified version of the GH_n score was used, a linear combination of the sensitivity and specificity. The advantage of this score over the Kappa statistics, which is sensitive to low prevalence, is that the modified GH_n score allows for the easy comparison of the score generated for the test and validation sets despite the low prevalence of actives in the validation set.

Additionally, for a model to be considered, at some threshold, it must satisfy the following two criteria. (1) It must recall 50% of the actives, and (2) it must provide hit rates at least twice as good as the original hit rate; that is, EF $\geq$ 2.

Of all the models that meet these criteria, the model with the highest average score on the test set is selected. To assess the suitability of a model, a receiver-operating-characteristic (ROC) curve[44] is generated for each model using the predictions on the test set. The ROC plots ensure that models are not discarded just because they do not meet the criteria at the default threshold (the placement of the hyperplane).[45,46]

To illustrate the automatic model selection criteria, Figure 2 shows a plot of EF versus sensitivity for the various models built for 5HT2c. Each point represents a model with different SVM parameters. For each model, a ROC plot is generated, and each threshold is ranked using the GH_n score. The threshold with the best possible sensitivity and enrichment is kept as representative of the highest capacity of the model. The model that was autoselected (highlighted) recalls more than 60% of the actives and more than 95% of the inactives, giving an EF close to 5. The model with the highest score is retained and is used for subsequent classification of new compounds. The performance of the model on a second left-out set (the validation set) is used to report the robustness of the models.

3. RESULTS AND DISCUSSION

3.1. Method Validation. In addition to testing the model against a dataset where the biological response (active/inactive) is scrambled, it is standard practice to test the models on a second "external" validation set. In the absence of a truly independent dataset, one is forced to divide randomly the same dataset into training, test, and validation

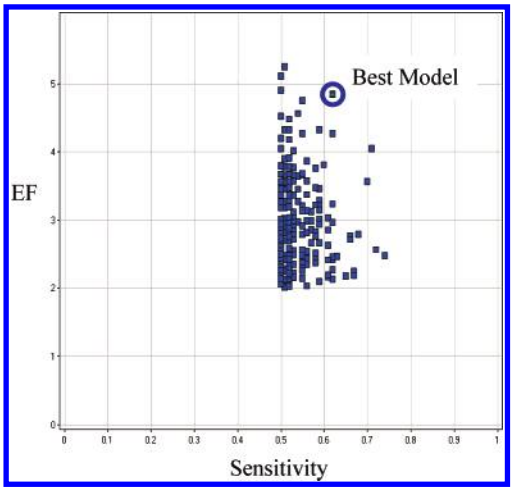


Figure 2. Performance of the various models built for 5HT2c on the held-out test set. For each model, each point on the ROC curve is scored and the threshold with the maximum score is kept. The best model (highlighted) is automatically selected. Note that all the models represented achieve a minimum of 50% recall and EF ≥ 2.

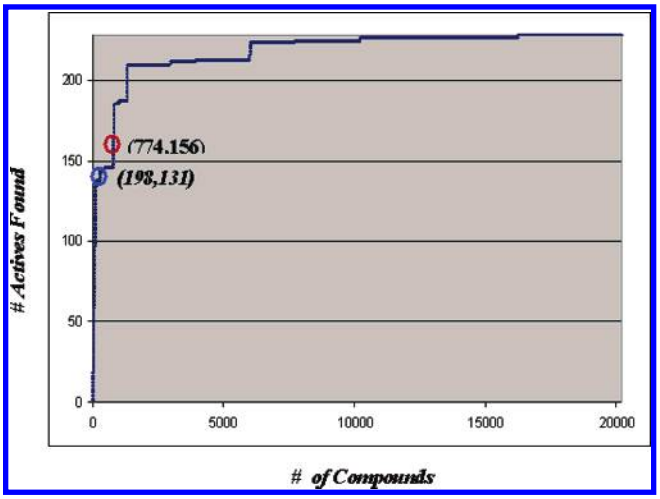


Figure 3. Vertical lines show the placement of the SVM hyperplane. The default threshold, red circle, results in a 17-fold enrichment in the hit rate. Moving the threshold to 0.2, blue circle, results in better enrichment (EF = 58), see Table 4.

subsets (Figure 1). One obvious criticism of this approach is that such a splitting scheme that is based on a randomization procedure generates test and validation sets that are similar to the training set because the samples are drawn from the same "chemical space". It is, therefore, not too surprising to find that the models that perform well on the test set perform equally well on the validation set. In this work, an attempt was made to test and validate the models on progressively more difficult datasets. Three approaches are illustrated in this section. In the first approach, models were trained and tested on proprietary datasets, and for validation, the same model was used to classify the MDDR data collection. In the second approach, the model is trained on MDDR data and the validation is done using proprietary datasets. In the absence of a true external dataset, the third approach shows the result of applying a model that was trained on MDDR data and validated on a dataset biased toward GPCR-focused compounds.

3.1.1. Models Built from Proprietary Data and Validated on MDDR Data—β1 Adrenoceptor and 5HT2C. Proprietary screening data was used to build a model for the GPCR β1 adrenoceptor. The model was generated from a dataset of 200 active and 2000 inactive compounds, and once again, the best model was selected automatically. The average sensitivity of the best model against the training data was 91%, and the average specificity was 98% (see Table 1). There are 68 compounds in the MDDR labeled as being active against β1. These compounds were then added to a pool of 60 000 inactive compounds from our proprietary collection, specifically excluding any of the 2000 inactive compounds used in the training phase. At the default threshold, the model correctly predicted 45 of these as being active, out of a total prediction of 2836 active compounds. This represents a 14-fold increase over the original 0.1% hit rate.

For the external validation of the model above, the active compounds were taken from the MDDR but the inactive set was taken from those found inactive in an in-house GPCR screening campaign. This is also the set from which the inactives were taken for the training set. Although there were no compounds in common in the two sets, it could be argued that because they came from the same "chemical space", there might be compounds in one that are very similar to compounds in the other, potentially biasing the results. To some extent, this possibility can be discounted by considering that only 1−3% of the inactives in the full set are being used to produce a model that predicts the remaining 97−99% of the inactive set with great success. To address this issue, a model was built using active and inactive data from an in-house screen testing for activity at the 5HT2c receptor, and subsequently, data solely from the MDDR was used as the external validation set. For this validation set, 228 compounds were listed in the MDDR as active against 5HT2c, and these comprised the active set. For the inactive set, 20 000 compounds were selected at random from those not listed as active against this receptor. It must be noted that although the fact that a compound has not been listed in the MDDR as being active against a particular receptor implies no activity, this is not explicitly known. It may be the case that the compound is indeed active but the test has not been carried out.

This example illustrates the power of SVMs as a classification tool. As discussed in section 2.2, once the vectors are generated, the SVM assigns a positive or negative value for each compound depending on how far and on which side of the hyperplane the compound lies. Sorting the output of the SVM essentially prioritizes the compound list. Depending on how many compounds a chemist is interested in screening, one may move the threshold (the placement of the hyperplane) between the margin, that is, between −1 and +1, thus increasing or decreasing the size of the suggested library. Figure 3 is a plot of the number of active compounds correctly classified versus the total number of compounds in the library (20 228), Table 1 summarized the performance of the model on the left-out validation set. If one chooses the default placement of the hyperplane ($\epsilon = 0$) to separate the actives from the inactives, that is, any compound with an SVM output ≥ 0 as active and an output < 0 as inactive, then 156 out of 228 actives are correctly classified (sensitivity

= 68.4%), and the suggested library contains no more than 774 compounds, compared with 20 228 in the original library. This translates into an enrichment factor ≈ 17. The sensitivity on the external dataset compares well with the previous tests, where the sensitivity was no more than 60%. Using a threshold of +0.2, which is within the margins for this hyperplane, only 198 compounds are predicted active, of which 131 are in fact in the original active set, resulting in an enrichment factor of 58 and a substantial reduction in the number of false positives.

3.1.2. Models Built from MDDR Data and Validated on Proprietary Data—Dopamine D2. This section addresses the ability of the SVMs to build a model using MDDR data that can predict activity for compounds tested in a proprietary screen. A total of 613 compounds were extracted from the MDDR database[47] with reported activity against D2. The remaining compounds in the database (129 963) were considered inactive. The active compounds were split according to our 80:15:5 rule. A total of 490 active compounds were used in training the models, 92 were used for testing, and 31 were used for validation. The inactives were split following the procedure outlined previously; 4900 went into the training set, 920 went into the testing set, and the remaining MDDR compounds with associated fingerprints went into the validation set. The model that performed best on the test and validation sets was picked for further validation.

The MDDR model was used to predict the activity of a dataset containing proprietary active D2 compounds and MDDR inactive compounds not included in training or testing. The results are summarized in Table 1. The model selected a list of only 1573 compounds for screening from a list of 129 994. The hit rate was approximately 69 times better than the hit rate in the original library.

The implication of this experiment is that SVM models built from external sources, like MDDR, can be validated on proprietary data. As a result, 67 SVM models have been built from MDDR data and made available through the web for routine screening of virtual libraries. Essentially, GPCR e-screens have been implemented for 67 GPCRs, spanning 18 GPCR subfamilies, thus allowing a chemist not only to flag compounds that are predicted active against a particular receptor but also to obtain additional predicted selectivity.

3.2. Lead Hopping. The results so far demonstrate that the SVM models can be used with some success. It is customary to have test and validation sets selected at random from a larger pool. Because both sets, thus, sample the same "chemical space", it could be argued that the fact that the models are as successful in predicting activity in the validation set, which was not used during training, as they are in the training test set is due to the presence of similar compounds in both sets. This concern has been addressed by building models using the in-house activity data and using the MDDR database to provide the validation set, and vice versa, as well as altering the chemical space of the validation sets by testing the models on GPCR-focused libraries. A more stringent test would be to attempt to ensure that the test and validation sets contain compounds from different chemical classes. If the models can predict well under these circumstances, they can truly be said to be based on the underlying pharmacophoric requirements of the receptor and will be of some utility in finding new chemical classes active against the receptor, so-called "lead hopping". A significant problem in this approach is creating datasets of this nature, because membership in a chemical class is a somewhat subjective matter. In this section, three such approaches are described, which are referred to as "leave cluster out", "leave nearest-neighbor out" and "leave core out".

Leave cluster out uses a type of sphere-exclusion clustering algorithm, described by Taylor.[48] An implementation of this algorithm using Daylight fingerprints is available as contributed code in the Daylight software distribution (program spherex.c).

Using Daylight fingerprints, it is customary to cluster similar compounds using a high Tanimoto similarity (typically 0.85 T_c). Compounds that are identical have a Tanimoto coefficient (T_c) of 1.00; compounds that are dissimilar have a $T_c = 0$. Given an active compound, we cluster all similar compounds ($T_c > 0.65$) together. Removing one cluster of compounds at a time from the training set is an intuitive and automated approach for simulating lead hopping.

Leave nearest-neighbors out expands the Tanimoto cutoff. In addition to the compounds removed in the first experiment, we remove from the training set compounds that are related to the validation compounds by a Tanimoto coefficient of less than 0.45. This has the disadvantage that the validation compounds do not all belong to the same chemical class, but the approach guarantees that the training set compounds are sufficiently distant from the core of interest.

Leave core out explicitly removes compounds from the training set on the basis of the presence of a core substructure. This is done without regard to fingerprint similarities and is arguably the best simulation of lead hopping.

3.2.1. Leave Cluster Out. The compounds in the MDDR database were clustered using a similarity cutoff of 0.65 T_c. In this section, a cluster that contains an active compound is considered an active cluster. Alternatively, a cluster without any hits is labeled an inactive cluster. By systematically leaving out one active cluster at a time from the training data, the overall performance of the various models on these left-out clusters may provide a more realistic measurement of the generalization of the model to new chemical classes.

The leave-cluster-out experiment was performed on four receptors from four different GPCR subfamilies: D2 (dopamine), 5HT1A (serotonin), NK2 (neurokinin), and α2 (adrenergic). For each receptor, the number of active compounds in each active cluster is listed in Table 2. The active clusters contained 4−90 compounds. Few inactive compounds clustered along with the active compounds.[50,51] To measure enrichment curves, the six largest inactive clusters were removed from the training sets. A compound was considered inactive if it was not listed active in the entire subfamily. For example, there were 1389 compounds that are reported active against one or more dopamine receptors and 613 compounds listed as active against D2.

All compounds not listed as active against a dopamine receptor were treated as inactive. The validation set consisted of one active cluster and the six inactive clusters, a total of 6800 inactive compounds. For D2, removing the third largest cluster (cluster #3 containing 13 active compounds) leaves 600 active compounds for training. Also included in the training set were 6000 compounds that were pulled at random from the inactive list. No model optimization was performed. A linear kernel was used with the default regularization

Table 2. Number of Actives in Each Left-Out Cluster[a]

cluster	D2	5HT1A	NK2	α2
1	13	11	15	12
2	11	17	10	8
3	13	22	9	19
4	4	20	13	9
5	26	18	10	14
6	5	23	11	11
7	15	14	31	8
8	22	27	25	7
9	18	24	10	33
10	10	29	14	7
11	4	29	10	7
12	4	36	14	14
13	12		10	8
14	22			33
15	6			

[a] The same inactives were held out for validation. The number of inactives in each set was 6900, chosen from the inactive MDDR data randomly.

Table 3. Sensitivity for the Left-Out Cluster[a]

cluster	D2	5HT1A	NK2	α2
1	0.85	0.73	0.6	1.00
2	1.00	1.00	0.9	1.00
3	0.69	1.00	0.89	1.00
4	1.00	0.95	1.00	1.00
5	0.27	1.00	1.00	0.93
6	1.00	1.00	0.82	1.00
7	0.2	0.43	0.94	1.00
8	0.95	1.00	0.48	1.00
9	1.00	1.00	1.00	0.61
10	0.9	0.76	0.79	0.43
11	1.00	0.83	1.00	0.14
12	1.00	0.97	0.79	0.00
13	0.83		0.10	1.00
14	1.00			0.55
15	0.67			
average	0.82	0.89	0.79	0.76

[a] The values are at the default threshold, $\epsilon = 0$. The overall average sensitivity for all left-out clusters is listed at the bottom of the table.

parameters, and the model generated was then used to classify the left-out clusters. It is important to emphasize that for each validation set, a new model was built on the remaining list of compounds.

Once the model was generated, the compounds that were left out for validation were classified. The sensitivity and enrichment factors for the various models are reported in Tables 3 and 4. Irrespective of the cluster removed, the model generally performs better than random. The average enrichments for the classification of the D2, 5HT1A, NK2, and α2 adrenergic sets are 12, 67, 10, and 26, respectively (see Table 4). It is perhaps not surprising that the best enrichment is achieved for 5HT1A. This is due, in large part, to the far greater number of active compounds, approximately 1500, that were included in the training set as well as the large number of inactives that were sampled during training (approximately 15 000). For each cluster removed, the enrichment varied from 32 to 121.

The average sensitivities for the D2, 5HT1A, NK2, and α2 adrenergic predictions are 0.82, 0.89, 0.79, and 0.76 respectively (see Table 3). The sensitivity varied from 0.2 to 1.0 for the various models. In the case of α2 adrenergic receptor, some of the models failed to recover any of the actives for a left-out cluster. For example, none of the seven

Table 4. Enrichment Factor for Each Validation Set in the Left-Out-Cluster Experiment[a]

cluster	D2	5HT1A	NK2	α2
1	11.64	121.92	19.14	60.31
2	17.21	87.56	16.11	56.08
3	9.97	87.37	15.71	44.18
4	19.67	82.86	13.95	42.93
5	2.98	81.28	13.66	34.15
6	15.21	63.92	10.96	32.27
7	2.82	63.21	10.55	32.24
8	10.49	51.61	10.42	24.23
9	12.96	49.48	7.20	17.63
10	10.70	47.55	6.63	9.06
11	14.00	39.94	1.84	7.01
12	19.99	30.9	1.75	0.00
13	12.64		1.13	5.53
14	11.71			1.42
15	7.50			
Average	11.97	67.30	9.93	26.22

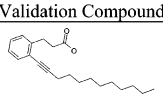
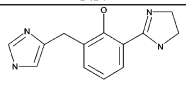
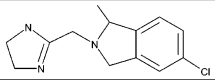
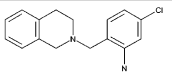
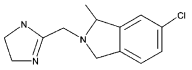
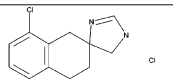
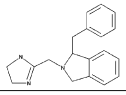
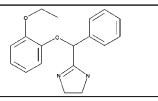
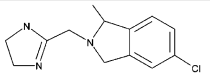
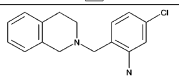
[a] The EF is reported at the default threshold, $\epsilon = 0$. The overall average enrichment for all left-out clusters is listed at the bottom of the table.

actives in cluster 12 were correctly classified at the default threshold of 0.0. However, a threshold of −0.1 results in three of the seven actives and 95% of the inactives being correctly classified.

3.2.2. Leave Nearest Neighbors Out. Removing one cluster at a time from the training set was motivated by lead-hopping considerations. However, for lead hopping, the experiment in the preceding section suffers from an objective criticism. Given the size of the cluster (0.35 T_c), it would not be hard to find compounds in the training set that are similar to the compounds in the left-out cluster. Some clusters in the training set may be near the left-out cluster, and therefore, one must consider the distance of the training compounds to the left-out validation compounds. In this experiment, this issue was addressed by constructing a more stringent validation set.

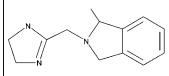
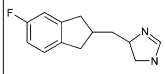
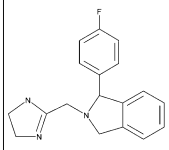
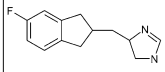
Using MDDRs, 370 active α2 adrenergic compounds were clustered using a small Tanimoto similarity ($T_c = 0.45$), resulting in 35 clusters. This is in contrast to the previous experiment where the compounds were clustered at $T_c = 0.65$, which put the active compounds in 122 clusters. The seventh largest cluster, containing 15 active compounds, was removed from the training set. To ensure that there are no compounds in the training set that are similar to any compound in the validation set, an in-house neighborhood analysis tool was used to generate each compound's nearest neighbor in the training set. Compounds in the training that had a measured Tanimoto similarity greater than 0.45 T_c to any compound in the validation set were removed. Four more compounds were found within that cutoff and were added to the validation set along with the other 15 compounds. As a result, the most similar active in the training set was 0.45 T_c and the least similar was 0.39 T_c. In Tables 5 and 6, examples of actives and their nearest neighbors as well as each pair's Tanimoto similarity are listed. In addition to the left-out validation active set, all the compounds belonging to the largest six inactive clusters (3011 compounds with fingerprints) were removed. The model was trained on the remaining active and inactive compounds. The results are summarized in Table 7. Using the default SVM threshold, 68% of the actives (13 out of 19) are correctly classified.

Table 5. Examples of Compounds Predicted Active in the Validation Set Exhibiting Lead Hopping[a]

Extreg	Validation Compound	Extreg	NN	Tc
146587		219215		.0.39
146580		140906		0.51
146581		222079		0.41
146589		176334		0.39
146580		140906		0.41

[a] The nearest neighbor (NN) in the training set as well as the pair's Tanimoto similarity (T_c) are listed.

Table 6. Two Examples of Compounds Predicted Active in the Left-Out Validation Set Whose Nearest Neighbors Have Low Daylight Similarity ($T_c = 0.4$) but Could Be Considered Relatively Similar

Extreg	Validation	Extreg	NN in Training	Tc
146578		293595		0.4
146584		293595		0.4

The hit rate in the new set is 37.1%, up from 0.6%.

It is customary to use 2D clustering with a high Tanimoto similarity coefficient (0.7−0.85) to investigate an active compound's nearest neighbors. An acquisition or synthetic strategy based on the 2D Tanimoto similarity would have likely missed the 19 actives held out in this validation set. Using our models, compounds that are not similar from a Daylight fingerprint viewpoint, which would potentially be discarded, are correctly identified as active.

The most similar compound in the training set to any compound in the validation is 0.45 T_c, yet the model is able to correctly classify nearly 70% of the left-out validation actives (Table 7). Although the results are encouraging, the question remains: starting with compounds in the training set, are the compounds in the validation set considered novel leads?

It is encouraging to note that compound number 146589 (Table 5) in the validation set is correctly classified despite the fact that its most similar compound in the training set is, at most, 0.39 T_c. Inspection of the structure of the nearest neighbor (compound number 176334) reveals that, indeed, these two compounds belong to different classes. Similarly, compound number 146580, an indoline, is also correctly classified. Its nearest neighbor in the training set is compound number 140906, an isoquinoline derivative, with a low Tanimoto similarity, $T_c = 0.51$.

Upon further examination of the training and validation sets, one may be surprised to find that compound number 293595, an indenyl linked to a dihydro imidazole, is in the training set (see Table 6). In fact, it is the nearest neighbor to two active compounds in the validation set. Some may argue, for example, that compounds number 146578 and number 293595 in Table 6 are very similar molecules and that both should have been removed from the training; however, from a Daylight similarity perspective, the similarity between the compounds is only 0.4 T_c.

These latter examples illustrate the problems of using Daylight fingerprint similarities in characterizing structural similarities. Very few will argue about the similarity of compounds in a cluster with a tight radius. However, using Daylight fingerprints to deduce the *similarity* or *dissimilarity* of compounds outside of that cluster is more difficult. Furthermore, a procedure that involves prioritizing compounds that are similar to a probe molecule will recover highly similar compounds but is less likely to lead to new chemical classes and, for this application, may leave behind similar active compounds. Because Daylight fingerprints are based on the path analysis of atoms in a compound, heteroatom substitutions in the structure of a small molecule will make similar compounds highly dissimilar.[49] This problem is particularly prevalent in small molecules.[50,51]

It could be argued that this experiment demonstrates lead hopping because the nearest neighbors in the training sets are all greater than 0.45 T_c. On the other hand, if one considers compounds 146578 and 293595 to belong to the same chemical class (despite their small Tanimoto similarity), then the next experiment remedies the problem by explicitly removing specific cores from the training set.

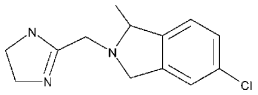
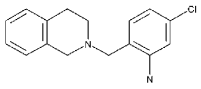
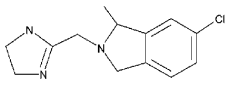
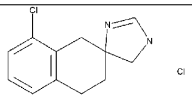
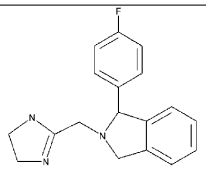
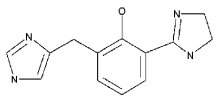
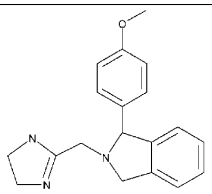
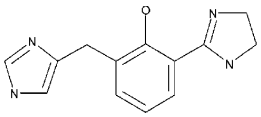
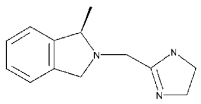
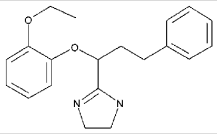
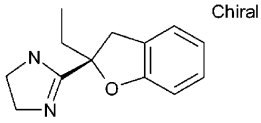
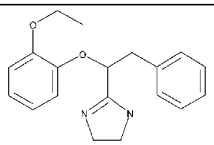
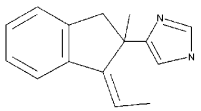
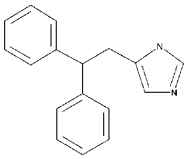
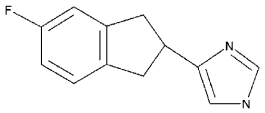
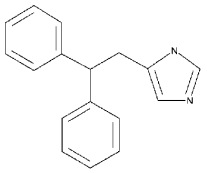
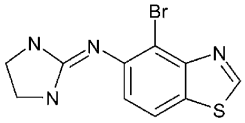
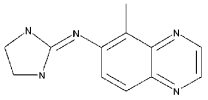
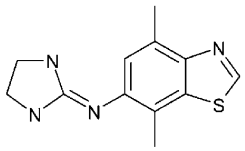
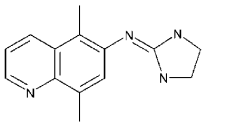
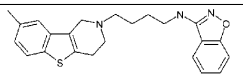
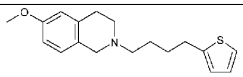
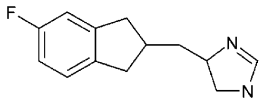
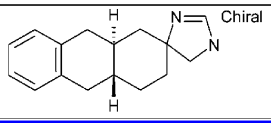
3.2.3. Leave Core Out. Given the limitations of Daylight fingerprints, it was decided to construct another validation set that does not rely on fingerprints but explicitly extracts cores of interest using pattern matching (SMARTS). For this validation, compounds that looked similar to compound number 146587 were of interest, and it was also decided to remove from the training set compounds that looked like compound number 293595. Consequently, all the compounds

Table 7. Summary of the Performance of the Models on the Held-Out Validation Sets for the "Leave-Cluster-and-Nearest-Neighbor-Out" and "Leave-Core-Out" Experiments[a]

	TP	FP	TN	FN	sensitivity	specificity	new hit rate	old hit rate	EF
leave NN out	13	22	2989	6	68.42	99.27	37.14	0.63	59.23
leave core out	106	47	4681	37	74.13	99.01	69.28	2.94	23.6

[a] The performance statistics for the prediction of α2 adrenergic activity were extracted from the usual confusion matrix (confusion matrix defined in Table 1). For the near neighbor analysis, any compound with a similarity > 0.45 Tanimoto was left out of the training.

Table 8. TP Compounds Exhibiting Lead Hopping in the Leave-Core-Out Experiment[a]

Extreg	Validation Compound	Extreg	Nearest Neighbor	Tc
146580		140906		0.41
146581		222079		0.41
146584		219215		0.38
146588		219215		0.42
170879		176339		0.38
256934		172370		.70
138883		126949		0.50
202375		126949		0.58
267079		175348		0.52
267076		228701		0.60
288723		173776		0.50
293595		290990		0.55

[a] T_c is the similarity between the validation compound and its nearest neighbor in the training set. A compound that is identical has a Tc = 1.00, while compounds that are dissimilar have a $T_c = 0.0$.

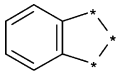


Figure 4. Depiction of the core removed from α2 adrenergic actives. An asterisk indicates any atom, and the dotted line in the five-membered ring indicates a σ or π bond. Compounds that matched the query were left out for validation; the remaining compounds were part of the training set.

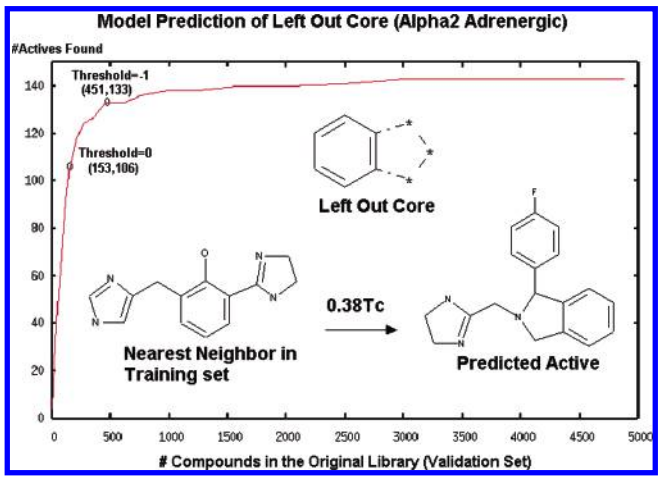


Figure 5. Prediction of the left-out validation set. The active compounds (143 in all) are related by a common core (Figure 4). There were 198 compounds left in the training set. The validation set also included 4728 inactive compounds belonging to the six inactive clusters. The plot shows that in the top 10% of the ranked leave-core-out validation set, 93% of the actives are recalled. The top 10% are at the negative margin ($\epsilon = -1$). At the default threshold, 153 compounds are chosen, of which 106 compounds are active. The default threshold corresponds to a 23.6-fold enrichment. The inset shows that training the model on compounds such as the nearest neighbor (on the left) allows us to hop to compounds such as the predicted active (on the right) where the Tanimoto similarity between the two is less than 0.4.

that contained the core in Figure 4 were left out. Note that the five-membered ring is allowed to contain any heteroatom and any bond order. As a result, the left-out compounds contained a wide range of chemical types including benzofurans, benzothiazoles, benzisoxazoles, indenyls, and indolines.[52]

The new validation set contained nearly 43% of the α2 adrenergic active compounds in the MDDR (143 compounds). This left 192 active compounds in the training set. The six largest clusters containing 4728 "inactive" compounds were removed from the training and held out for validation. A random 1920 inactive compounds were pulled out from the remaining MDDR compounds and were added to the training set. The model was trained on the training dataset using default SVM parameters. A cost factor ($j = 10$) was used to offset the 1:10 ratio of the actives to inactives in the training set. The performance of the model on the held-out (core-centered) validation set is summarized in Table 7. The model correctly predicted 74% of the left-out actives and 99% of the left-out inactive compounds.

In Figure 5, the compounds are ranked in descending order of their SVM score (x axis). On the y axis, the number of true actives found in the new set is shown. Indirectly, the plot illustrates the performance of the SVM model at different thresholds. The negative margin ($\epsilon = -1$) corresponds to roughly 10% of the validation set. The total number of compounds at the negative margin is 451 compounds, of which 133 are true actives. That is, in the top 451 compounds, 93% of the held-out validation actives are found, all related to the core in Figure 4. At the default threshold, 153 compounds are chosen by the model, of which 106 compounds are active. At the default threshold, the model recalls 74% of the actives and provides greater than 23-fold enrichment. The inset shows that, starting with training compounds that looked like the one on the left (nearest-neighbor in training), it is possible to hop to the compound on the right (predicted active) whose Tanimoto similarity is less than 0.40. The new model correctly classifies the core of interest, indolines, represented for example, by compounds such as 146581 and 14658, as well as benzofurans, compounds such as 256934; benzothiazoles, compounds such as 267076; benzisoxazols, compounds such as 288723; and indenyls, compounds such as 293595 (see Table 8), thus providing the most exhaustive and comprehensive illustration of lead hopping.

4. CONCLUSION

This study has demonstrated that SVMs together with 3D pharamacophore descriptors can be successfully used to predict activity against a variety of GPCRs. To show that the models are capable of lead hopping, three experiments were conducted. The first two experiments were based on clustering of the active compounds using Daylight fingerprints and the third by using SMARTS[36] to extract compounds containing certain "cores" from the training set. Both approaches provide convincing evidence of the models' ability to generalize to new compounds and, more importantly, to generalize to new chemical classes. The models were able to correctly classify a significant number of actives (sensitivity = 62−82%) in the held-out validation while providing significant enrichment. This approach has been applied against several in-house GPCR programs and has been found to be as affective as the examples presented in this paper. The approach described in this paper is general in nature and is not restricted to the development of models for GPCRs but may be applied to many gene families.

ACKNOWLEDGMENT

We are grateful to Thorsten Joachims for providing us with helpful insight into the use of SVM*light*. This work benefited from many helpful discussions with Michelle Lamb, Daniel Russel, and William Hayes. The authors also wish to acknowledge the help of Mirek Tomaszewski, V. Santhakumar, Thorsten Nowak, and Garry Pairaudeau for providing us with the data and the initial feedback on the models.

Supporting Information Available: For the leave-core-out experiment, we include a database file of chemical compounds compiled from MDDR for the α2 adrenoceptor active compounds. Included are the structure in Daylight SMILES notation, the compound MDDR ID, the SVM model prediction, the structure and ID of the compounds' nearest neighbor in the training set, the Daylight fingerprint Tanimoto similarity between the validation compound and its nearest neighbor, and a flag indicating whether the compound was used for training or for validation. This material is available via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bajorath, J. Integration of virtual and high throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.

(2) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discovery Today* **2000**, *5*, S61−S69.

(3) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem.* **1999**, *38*, 2894−2896.

(4) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery Toward general methods of targeted library design. Topomer shape similarity searching with diverse structures as queries. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(5) Andrews, K. M.; Cramer, R. D. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(6) Kennedy, T. Managing the drug discovery process. *Drug Discovery Today* **1997**, *2*, 436−444.

(7) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.

(8) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.

(9) Dean, P. M. *Molecular similarity in drug design*; Chapman & Hall: New York, 1995.

(10) Pepperrell, C. *Three-dimensional chemical similarity searching*; Research Studies Press: Baldock, Hertfordshire, U. K., 1994.

(11) Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 220−227.

(12) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.

(13) Van Drie, J. H. Addressing the challenges of combinatorial chemistry: 3D databases, pharmacophore recognition and beyond. *SAR QSAR Environ. Res.* **1998**, *9*, 1−21.

(14) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C. Prospective identification of biologically active structures by topomer shape similarity searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(15) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D. J. Evaluation of a novel shape-based computational filter for lead evolution: Application to thrombin inhibitors. *J. Med. Chem.* **2002**, *45*, 2494−2500.

(16) *ROCS*; OpenEye Science Software: Santa Fe, NM.

(17) Zheng, W. F.; Cho, S. J.; Tropsha, A. Rational combinatorial library design. 1 Focus-2D: A new approach to the design of targetted combinatorial chemical libraries. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 251−258.

(18) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(19) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare clustering methods and descriptors for use in compound selection *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(20) Van Drie, J. H. Shrink-wrap surfaces: A new method for incorporating shape into pharmacophoric 3D database searching. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 38−42.

(21) Vapnik, V. N. *The nature of statistical learning theory*; Springer-Verlag: New York, 1995.

(22) Cortes, C.; Vapnik, V. N. Support vector networks. *Mach. Learning* **1995**, *20*, 1−25.

(23) Schölkopf, B.; Smola, A. J. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(24) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.* **2002**, *9*, 849−863.

(25) Norinder, U. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimizations and variable selection. *Neurocomputing* **2003**, *55*, 337−346.

(26) *SMILES*; Daylight Chemical Information Systems Inc.: Santa Fe, NM.

(27) *SMARTS*; Daylight Chemical Information Systems: Santa Fe, NM.

(28) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of automatic 3-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(29) Sadowski, J. A hybrid approach for ring flexibility in 3D database searching. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 53−60.

(30) *Omega*; OpenEye Science Software: Sante Fe, NM.

(31) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graph Model.* **2003**, *21*, 449−462.

(32) Abrahamian, E.; Fox, P. C.; Narum, L.; Christensen, I. T.; Thogersen, H. Efficient generation, storage and manipulation of fully flexible pharmacophore multiplets and their use in 3-D similarity searching. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 458−468.

(33) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifier. In *Proc. 5th ACM Workshop on Computational Learning Theory*, Pitsburgh, PA, July 1992; pp 144−152.

(34) Burges, C. J. C. *Simplified Support Vector Decision Rules*, 1996.

(35) Cortes, C.; Vapnik, V. Support vector networks. *Mach. Learning* **1995,** *20*, 1− 25.

(36) Daylight Chemical Information Systems, Santa Fe, NM.

(37) Tripos Inc. http://www.tripos.com.

(38) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280. MDL Information Systems, Inc. is at http://www.mdli.com.

(39) SVM^{light}, version 4.00. http://svmlight.joachims.org/.

(40) Joachims, T. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*; Kluwer Academic Publishers: Boston, MA, 2001.

(41) Lewis, D.; Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, London; Springer-Verlag: New York, 1994; pp 3−12.

(42) Carletta, J. Assessing agreement on classification tasks: the *kappa* statistic. *Comput. linguistic*s **1996**, *22*, 249−254. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, 37−46. Cook, R. J. Kappa. In *The Encyclopedia of Biostatistics*; Armitage, P., Colton, T., Eds.; Wiley: New York, 1998; pp 2160−2166.

(43) Güner, O. F. *Pharmacophore Perception, Development, and Use in Drug Design*, IUL Biotechnology Series; International University Line: La Jolla, California, 2000.

(44) Egan, J. P. *Signal Detection Theory and ROC Analysis*; Academic Press: New York, 1975. Begg, C. B. Statistical methods in medical diagnosis. *CRC Crit. Rev. Med. Inf.* **1986**, *1*, 1-22.

(45) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in Drug Discovery Process. *J. Chem. Inf. Comput. Sci*. **2003**, 43 (2), 667−673.

(46) The default threshold, associated with the placement of the hyperplane, does not guarantee that the solution is Bayes optimal. Some compounds will be found within the margin, and it may be advantageous to walk the threshold between −1 and +1. In a recent publication, Warmuth et. al.[45] recommend using the margin at different stages in drug discovery. At the beginning of a project, where lead identification requires quick identification of potential targets, the authors recommend exploring compounds with output $\epsilon \geq -1$, and in lead optimization, one may need to exploit structures that have already been identified as active, a better threshold would be $\epsilon \geq +1$.

(47) *MDL Drug Data Report*, version 2002.2; MDL ISIS/HOST software, MDL Information Systems, Inc.: San Leandro, CA. http://www.mdli.com.

(48) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 59−67. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 747−750.

(49) Small molecules are particularly susceptible to heteroatom substitutions. This is a consequence of the way the fingerprint is generated via a hashing scheme that sets bits on or off depending on whether a particular molecular path exists in the target molecule. As a result, many of the fingerprints in MDDR's α2 adrenergic compounds contain few bits that are set. Therefore, similarity measurements based on direct comparisons of the bits that are set in common between the fingerprints of two molecules will be strongly influenced by small fingerprint differences.

(50) See Figures 2−5 in Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.

(51) In Figure 4, ref 50, the author used a database of 50 000 randomly selected compounds to pull compounds similar to a query molecule. The closest compound in the database was 0.6 Tanimoto. A compound was "artificially" added to the list in order to illustrate what a 0.8 Tanimoto compound may look like. Our experience with the MDDR and our corporate dataset is in agreement with this observation. Using a 0.3, 0.4, and 0.5 Tanimoto distance, few of our corporate compounds were found in the same cluster as the MDDR's. Essentially, clustering the merged MDDR and corporate datasets provides little advantage. In agreement with the observation in ref 50, few of the corporate compounds are within 0.5 Tanimoto from the MDDR. Furthermore, clusters are composed of either exclusively MDDR compounds or corporate compounds. This is also in line with our observation that

clusters that include active compounds in the MDDR usually do not include a significant number of remaining ("inactive") compounds.

(52) The compounds that were part of the training and validation sets are included in the Supporting Information. For each compound in the validation, we list the compound's nearest neighbor in the training set, as well the pair's Tanimoto similarity.

CI049732R