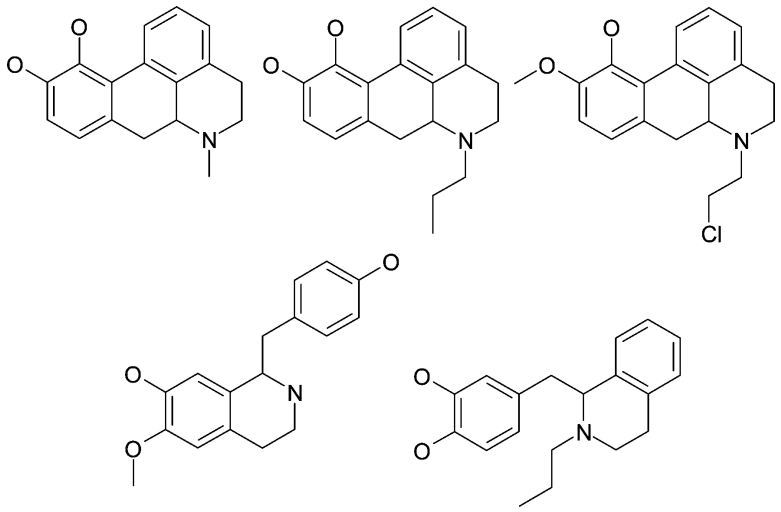
Journal of Medicinal Chemistry

Article

Application of Validated QSAR Models of D Dopaminergic Antagonists for Database Mining

Scott Oloff, Richard B. Mailman, and Alexander Tropsha

J. Med. Chem., **2005**, 48 (23), 7322-7332• DOI: 10.1021/jm049116m • Publication Date (Web): 13 October 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 15 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Journal of Medicinal Chemistry is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Application of Validated QSAR Models of D_1 Dopaminergic Antagonists for Database Mining

Scott Oloff,[†,‡] Richard B. Mailman,[†,‡,§] and Alexander Tropsha*[,‡]

Department of Pharmacology, School of Medicine, Departments of Psychiatry and Neurology, School of Medicine, and Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599

Received November 3, 2004

Rigorously validated quantitative structure−activity relationship (QSAR) models have been developed for 48 antagonists of the dopamine D_1 receptor and applied to mining chemical datasets to discover novel potential antagonists. Several QSAR methods have been employed, including comparative molecular field analysis (CoMFA), simulated annealing−partial least squares (SA-PLS), *k*-nearest neighbor (kNN), and support vector machines (SVM). With the exception of CoMFA, these approaches employed 2D topological descriptors generated with the MolConnZ software package (EduSoft, LLC. MolconnZ, version 4.05; http://www.eslc.vabiotech.com/ [4.05], 2003). The original dataset was split into training and test sets to allow for external validation of each training set model. The resulting models were characterized by cross-validated R^2 (q^2) for the training set and predictive R^2 values for the test set of (q^2/R^2) 0.51/0.47 for CoMFA, 0.7/0.76 for kNN, R^2 for the training and test sets of 0.74/0.71 for SVM, and training set fitness and test set R^2 values of 0.68/0.63 for SA-PLS. Validated QSAR models with $R^2 > 0.7$, (i.e., kNN and SVM) were used to mine three publicly available chemical databases: the National Cancer Institute (NCI) database of ca. 250 000 compounds, the Maybridge Database of ca. 56 000 compounds, and the ChemDiv Database of ca. 450 000 compounds. These searches resulted in only 54 *consensus* hits (i.e., predicted active by all models); five of them were previously characterized as dopamine D_1 ligands, but were not present in the original dataset. A small fraction of the purported D_1 ligands did not contain a catechol ring found in all known dopamine full agonist ligands, suggesting that they may be novel structural antagonist leads. This study illustrates that the combined application of predictive QSAR modeling and database mining may provide an important avenue for rational computer-aided drug discovery.

Introduction

The class of G protein-coupled receptors includes the dopamine receptors, made of two subclasses (D_1-like and D_2-like subtypes[2]) coded from five genes. The dopamine receptors play important roles such as modulation of motor function, cognition, memory, emotional activity, and various peripheral functions[3], and have been especially implicated in disorders such as Parkinson's disease and schizophrenia.[4] The consequences of activation or blockade of dopamine receptors are wide-ranging,[5−8] and perturbation of dopamine neurotransmission may result in profound neurological, psychiatric, or physiological signs and symptoms. For these reasons, there has been a great deal of research focused on the discovery of novel dopaminergic ligands as potential drug candidates. One aspect of this research has been the development of several pharmacophore and/or quantitative structure−activity relationship (QSAR) models of dopaminergic action.[9−18]

Due to the difficulties with crystallizing transmembrane proteins, X-ray structures for the dopamine receptors are not currently available. Traditionally, this leaves two computational approaches to discover novel ligands for the dopamine receptors: modeling of the receptor binding site, which can be used for structure-based design and ligand-based drug design approaches, e.g., active analogue approach[17] or QSAR.[19] The first approach can be aided by site-directed mutagenesis within the putative binding site to investigate its effect on binding of high affinity rigid ligands.[20−24] The second approach analyzes compounds with known specific binding affinity to a receptor and is used to build quantitative models relating affinity to molecular structure. Recent studies have emphasized the importance of model validation to ensure that it is capable of accurately predicting the binding affinity for a test set containing compounds not used in the model generation.[25] Once a series of validated models are collected, they then can be used for database mining to identify possible ligands for the receptor of interest.[9,26]

Herein, we report on the development of rigorously validated QSAR models for 48 D_1 dopaminergic antagonists. Despite many years of intensive research, QSAR modeling remains largely an empirical approach where the choice of the class of descriptors and the choice of an optimization technique resulting in successful models are not known a priori. For instance, our recent studies of fragrances using a combinatorial QSAR approach[27]

* To whom correspondence should be addressed. Mailing address: CB 7360, UNC−Chapel Hill, Chapel Hill, NC 27599. Tel: (919) 966 2955. Fax: (919) 966-0204. E-mail: tropsha@email.unc.edu.
† Department of Pharmacology, School of Medicine.
‡ Division of Medicinal Chemistry and Natural Products, School of Pharmacy.
§ Departments of Psychiatry and Neurology, School of Medicine.

10.1021/jm049116m CCC: $30.25
Published on Web 10/13/2005

indicated that only a small fraction of possible combinations of descriptor types and optimization techniques result in acceptable models. Thus, we chose to use several independent QSAR methods including CoMFA,[28] SA-PLS,[29] kNN,[30] and SVM[31] to determine which would result in the most predictive models. The purpose of this study was to generate a library of validated QSAR models with the highest external predictive power and use these models to screen available chemical databases to identify novel high affinity dopamine ligands. An important component of our studies was that, whenever possible, we have defined model applicability domains. These domains are defined quantitatively in the chemical descriptor space in the content of each specific QSAR model. They define the limits of structural diversity of external compounds for which prediction is considered possible based on the model-dependent estimate of the structural domain covered by the training set model.[32] Mining the publicly available NCI, ChemDiv, and Maybridge databases with QSAR models taking into account their applicability domains resulted in only 54 consensus hits with high predicted affinity to D_1 receptor. Five of these hits had experimentally confirmed binding to the dopamine D_1 receptor, but were not present in the original dataset.

This strategy for drug discovery that combines validated QSAR modeling and database mining has been under development in our group for several years[26,29] and was most recently applied successfully to the discovery of novel anticonvulsant agents.[29] We shall emphasize that this approach shifts the traditional focus of QSAR modeling from achieving statistically significant training set models (where the results are presented in the form of statistical parameters) to identify novel potentially active compounds on the basis of statistically significant and externally validated models (i.e., where the results are presented in the form of compounds). We believe that this shift brings QSAR modeling in tune with the ultimate needs of experimental medicinal chemists in novel compounds rather than models.

Methods

Datasets. Test and Training Set Selection. The pharmacological data for 48 D_1 antagonists used in this study were generated in one of our laboratories; the affinity of each compound against the D_1 receptor was experimentally measured in triplicate (reported elsewhere;[18,33−35] chemical structures of antagonists and the values of their binding affinity are given in the Supporting Information). MolConnZ[1] descriptors were calculated for each compound. The descriptors were then linearly normalized to fall within the range between zero and one based on the maximum and minimum values of each descriptor (i.e., range-scaled). Normalization was required to prevent unequal descriptor weighting during the QSAR model generation process as described later. The datasets were then subdivided into multiple training/test set pairs using the sphere exclusion method[36] developed in our laboratory. The number of compounds in the test set was varied to achieve the largest possible size of the test set, while ensuring that the training set models were still able to accurately predict the binding affinity of the test set compounds. The statistical significance of the training and test set models was characterized with leave-one-out cross-validated R^2 (q^2), or fitness criterion, or R^2 and a linear fit predictive R^2, depending on the modeling approach used (vide infra). The acceptability cutoffs were $q^2/R^2 > 0.6/0.6$[37] for CoMFA and kNN approaches; for SA-PLS we used the training set fitness and R^2 test set cutoffs of 0.6/0.6, and for SVM the R^2 training and R^2 test set cutoffs were 0.6/0.6. Models that did not meet these cutoff criteria were discarded. Additional details on the selection of training and test sets and model acceptance criteria are described elsewhere.[36,37]

Y-Randomization Test. This is a widely used technique to ensure the robustness of a QSAR model.[38] In this test, the dependent-variable vector, **Y**-vector, is randomly shuffled and a new QSAR model is developed using the original independent-variable matrix. This process is repeated several times. It is expected that the resulting QSAR models should generally have low training set statistics (q^2, fitness, R^2_{train}, etc.) and low test set R^2. It is likely that sometimes, though infrequently, high training set statistics may be obtained due to a chance correlation or structural redundancy of the training set.[39] If all QSAR models obtained in the **Y**-randomization test exhibit relatively high training set statistics and test set R^2 values, it implies that an acceptable QSAR model cannot be obtained for the given dataset by the current modeling method. This test was applied to all datasets considered in this study.

CoMFA. Structures were generated and CoMFA was performed within the QSAR module of the SYBYL version 6.8 molecular modeling package.[28] Default Sybyl settings were used, except as otherwise noted. Molecular mechanics calculations were performed with the standard Tripos force field with a convergence criterion requiring a minimum energy change between free energy minimization steps of 0.01 kcal/mol. The molecules were aligned in 3D space such that three or more structural features common to all of the compounds in the training set had approximately the same Cartesian coordinates. In this study, the amine group and the two aromatic carbons corresponding to those where hydroxyls would be present in the catechol ring of full D_1 agonists (i.e., beta-phenyl dopamine pharmacophore) were used as the three points of alignment. Charges were calculated using the Gasteiger−Huckel method as implemented in SYBYL. The steric and electrostatic field energies were calculated using sp^3 carbon probe atoms with a +1 charge. The CoMFA QSAR equations were calculated using the partial least-squares (PLS) algorithm. The optimal number of components (ONC) in the final PLS model was determined by the cross-validated R^2 (q^2) and standard error of prediction (SDEP) values, as obtained from the leave-one-out cross-validation technique. Statistical results reported by CoMFA are in q^2 (eq 1) and R^2 (eq 2) values, where q^2 represents the training set and R^2 represents the test set.

$$q^2 = 1 - \left(\frac{\sum_Y (Y_{\text{pred}} - Y_{\text{actual}})^2}{\sum_Y Y_{\text{actual}} - \bar{Y}} \right)^2 \quad (1)$$

where Y_{pred} = predicted affinity, Y_{actual} = actual affinity, and $\bar{Y}$ = mean actual affinity, and

$$R^2 = \left(\frac{\sum_Y (Y_{\text{pred}} - \bar{Y}_{\text{pred}})(Y_{\text{actual}} - \bar{Y})}{\sum_Y (Y_{\text{pred}} - \bar{Y}_{\text{pred}})^2 \sum_Y (Y_{\text{actual}} - \bar{Y})^2} \right)^2 \quad (2)$$

where $\bar{Y}_{\text{pred}}$ is the mean predicted affinity.

The number of components with the lowest SDEP and highest q^2 value was selected as the ONC.

Generation of 2D Molecular Descriptors. All chemical structures were generated using the SYBYL software package.[28] Molecular topological indices[40] were generated with the MolConnZ (MZ) software version 4.05.[1] Overall, MolConnZ produces over 400 different descriptors. Most of these descriptors characterize chemical structure, but several depend on the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, only 312 chemically relevant descriptors were initially calculated and a small subset of descriptors with zero values or zero variance was removed prior to model generation.

MZ descriptors were also range-scaled since the absolute scales for MZ descriptors can differ by orders of magnitude. Accordingly, our use of range scaling avoided giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional MZ descriptor space. All calculations were performed on an SGI Octane at the University of North Carolina's Molecular Modeling Laboratory.

SA-PLS Method. SA-PLS employs a combination of a simulating annealing driven sampling of the descriptor space with the use of a PLS algorithm to optimize the correlation between the descriptor values and the target property. This method is based on our earlier implementation of genetic algorithms for variable selection QSAR method (i.e., GA-PLS).[41] The optimal set of descriptors is determined by an iterative variable selection method. The iterative loop begins by randomly selecting a specified number of descriptors and building a model using PLS to determine the fitness of the model produced with that set of descriptors (eq 3).

$$\text{fitness} = 1 - \frac{n}{n-m-1}\frac{\sum_{Y}(Y_{\text{pred}} - Y_{\text{actual}})^2}{\sum_{Y}(Y_{\text{actual}} - \bar{Y})^2} \tag{3}$$

where n = number of compounds, m = number of principal components, Y_{pred} = predicted affinity, Y_{actual} = actual affinity, and $\bar{Y}$ = mean affinity of all the compounds. A Metropolis-like acceptance criterion (eq 4) is used to drive the descriptor selection process toward an optimized descriptor subset.

$$\vartheta \leq \mathrm{e}(F_{\text{new}} - F_{\text{old}})/T \tag{4}$$

where ϑ = a random number between 0 and 1, F_{new} = fitness of the new model, F_{old} = fitness of old model, and T = temperature.

After each iteration, the temperature is lowered by a specified percentage, which decreases the probability that the less predictive model is chosen. This allows the algorithm to find an internally predictive set of descriptors and increases the probability that the final model is an optimum solution where changing a small subset of the descriptors would most likely not improve the model. Additional details of the implementation of the SA procedure for variable selection QSAR are given elsewhere.[30]

kNN Method. The kNN QSAR method[30] employs the kNN pattern recognition principle[39] and a variable selection procedure. Briefly, a fixed size subset of descriptors is selected randomly in the beginning of the calculations. The model is built using this random descriptor selection with leave-one-out (LOO) cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the average activity of k most similar molecules (k = 1−5). The value k is optimized during the model building process to give the best prediction of the training set. The similarity is characterized by the Euclidean distance between compounds in multidimensional descriptor space. A method of simulated annealing with the Metropolis-like acceptance criterion (similar to the one described above for the SA-PLS method) is used to sample the entire descriptor space to converge on the subset of the same size which affords the highest value of LOO R^2 (q^2). The descriptor subsets of different sizes are optimized using this procedure to arrive at a variety of models with acceptable q^2 greater than certain threshold (we chose 0.6 as the default threshold). The training set models with acceptable q^2 are then validated on external test sets to select predictive models as discussed above. Further details of the kNN method implementation, including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere.[30]

The original kNN method[30] was enhanced in this study by using weighted molecular similarity. In the original method, the activity of each compound was predicted as the algebraic average activity of its k-nearest-neighbor compounds in the training set. In general, however, the Euclidean distances in the descriptor space between a compound and each of its k nearest neighbors are not the same. Thus, the neighbor with the smaller distance from a compound was given a higher weight in calculating the predicted activity as follows:

$$w_i = \frac{\exp(-d_i)}{\sum_{i=0}^{k}\exp(-d_i)} \tag{5}$$

$$\tilde{y} = \sum w_i y_i \tag{6}$$

Here d_i is the Euclidean distance between the compound and its k nearest neighbors, k is the number of nearest neighbors, w_i is the weight for every individual nearest neighbor, y_i is the actual activity value for nearest neighbor i, and $\tilde{y}$ is the predicted activity value. In summary, the kNN algorithm generates both an optimum k value and an optimal nvar subset of descriptors, which afford a QSAR model with the highest value of q^2. This modified algorithm was also applied recently to the kNN QSAR modeling of anticonvulsant agents.[42]

Applicability Domain of kNN QSAR Models. Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. Since the training set models are developed in the kNN QSAR approach by interpolating activities of the nearest neighbor compounds,[30] a special applicability domain (i.e., similarity threshold) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules.[25]

To measure similarity, each compound is represented by a point in the M-dimensional descriptor space (where M is the total number of descriptors in the descriptor pharmacophore) with the coordinates $X_{i1}, X_{i2}, \ldots, X_{iM}$, where X_is are the values of individual descriptors. The molecular similarity between any two molecules is characterized by the Euclidean distance between their representative points. The Euclidean distance d_{ij} between two points i and j (which correspond to compounds i and j) in M-dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{M}(X_{ik} - X_{jk})^2} \tag{7}$$

Compounds with the smallest distance between one another are considered to have the highest similarity. The distances (similarity) of compounds in our training set are compiled to produce an applicability domain threshold, D_{T}, calculated as follows:

$$D_{\text{T}} = \bar{y} + Z\sigma \tag{8}$$

Here, $\bar{y}$ is the average Euclidean distance of the k nearest neighbors of each compound within the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Based on successful results from previous studies, we set the default value of this parameter to 0.5, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation (assuming a Boltzmann distribution of distances between k nearest neighbor compounds in the training set). Thus, if the distance of the external compound from at least one of its nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable.

SVM Method. Support vector machines (SVM) was developed by Vapnik[31] as a general data modeling methodology where both the training set error and the model complexity are incorporated into a special loss function that is minimized during model development. The methodology allows one to regulate the importance of the training set error versus the model complexity to develop the optimum model that best predicts a test set. Later SVM was extended to afford the

development of SVM regression models for datasets with noninteger activities, such as QSAR.

We have implemented the SVM method for QSAR modeling as follows: Let m be the number of points representing the training set compounds with known biological activity in an n-dimensional descriptor space. The problem is to generate a hypersurface in the descriptor-activity $(n + 1)$ dimensional space that relates descriptor values to the biological activities. Thus, the biological activity of any compound can be predicted from its descriptors by placing the point corresponding to this compound on this hypersurface.

Given a training set of instance-label pairs (x_i, y_i), $i = 1, ..., m$, where $x_i \in R^n$ are the descriptors that describe each compound and y_i is the biological activity (e.g., IC_{50} value) of each compound, the sought correlation between structure and activity can be represented as $y_i = f(x_i)$. For simplicity, we will define $f(x_i)$ to be a linear function of the form

$$f(x_i) = \langle \omega_i, x_i \rangle + b \tag{9}$$

where ω is the coefficient vector of the linear function and b is the bias. One major goal of any regression algorithm is to minimize the errors between the predicted and the actual values as defined by ξ_i in the following equation:

$$|y_i - (\langle \omega_i, x_i \rangle + b)| = \xi_i \tag{10}$$

As a means of regulating generalization of the algorithm, SVM utilizes the following constraint to solve the optimization problem:

$$\min_{\omega,b,\xi} \frac{||\omega||}{2} + C\sum_{i=1}^{m} \xi_i \tag{11}$$

with the constraint

$$|y_i - (\omega\phi(x_i) + b)| = \xi_i$$

whereas the training vectors x_i are mapped into a higher dimensional space by a kernel function ϕ. Then SVMs finds a linear correlation between the actual activity and this higher dimensional space $\phi(x_i)$. For this study, we have implemented a linear kernel. $C > 0$ is the penalty parameter of the error term that controls the weight between the two terms in the SVM optimization problem.

In many cases, however, the binding activities may contain small errors or the kernel function may not be capable of perfectly representing the training compounds in a simplified manner. As a means of inhibiting the algorithm from producing an overly complicated training set correlation that would not accurately predict a test set, we included a slack variable, ϵ. This slack variable is a threshold of prediction error for any compound's activity before the algorithm is penalized for a poor prediction. Beyond the boundary ϵ the algorithm is penalized by the value of $\xi_i - \epsilon$. When combining the SVM optimization problem defined in (11) with this added slack variable, the following SVM loss function is obtained:

$$_{\min}\text{loss} = \frac{||\omega||}{2} + C\sum_{i=1}^{m} \begin{cases} 0, & \text{if } \xi_i \le \epsilon \\ \xi_i - \epsilon, & \text{if } \xi_i > \epsilon \end{cases} \tag{12}$$

The nature of SVMs requires one to specify the values of C and ϵ a priori since it is not known beforehand which values may work best for the dataset; thus, a parameter search must be performed. The goal is to identify good values of C and ϵ such that the model can accurately predict unknown data (i.e., testing data). In most circumstances, the highest training accuracy does not yield the best accuracy on a test set. Therefore, the optimum C and ϵ values are commonly selected based on the values that give the best test set results.

For this study we have chosen to use a "grid-search" on C and ϵ to identify the best parameters. There are several advanced methods which can save computational cost by estimating the best parameters. We chose a simple grid-search approach for the following two reasons. First, unlike alternative methods which use approximations or heuristics, grid search allows for an exhaustive parameter search. Second, the computational time to find good parameters by a grid-search is not much longer than the time required by advanced methods since there are only two optimization parameters. Furthermore, the grid-search can be easily parallelized because each parameter is independent. Many of the advanced methods for parameter estimation are iterative processes, e.g., walking along a path, which is difficult for parallelization.

Especially for large datasets, a complete grid-search may be overly time-consuming; therefore, we commonly use a coarse grid on a subset of available data first. A user may randomly choose a subset of the dataset, conduct a grid-search using those compounds, and then do a fine-tuned grid-search on the complete dataset over the parameter value ranges that exhibited the best results. For this study our coarse grid-search of C varied from 50 to 1000 with an increment of 63, and ϵ was varied from 0 to 1.5 with an increment of 0.1. Once the best parameters for C and ϵ were found, we then did a fine-tuned search surrounding those values with ranges of ± 200 and ± 0.3 for C and ϵ, respectively. In this fine-tuned search the increments were 5 for C and 0.05 for ϵ.

Applicability Domain of SVM QSAR Models. To prevent the inaccurate activity prediction of compounds dissimilar to the training set, we have created a simple pseudoboundary around our training set compounds to dictate where the training set resides within the chemistry space. Any test compound that lies in the chemistry space beyond this boundary may possess a different descriptor−activity relationship than the training set compound. For this reason we chose not to predict biological activities for compounds beyond this boundary.

Based on successful results we have generated using kNN's modeling with applicability domain, we chose to take a similar approach with the SVM QSAR. The SVM method does not employ a definitive variable selection technique as in kNN, where the weights of each descriptor are assigned a value of 0 or 1. Instead, the weight of each descriptor is a noninteger number that may have a positive (directly correlated with the biological property) or negative (inversely related to the biological property) value. Knowing this, we wanted to select a subset of descriptors that the model found to correlate well with compound activity. Since both high positive weights and low negative weights may be vital for activity prediction, we chose to use the absolute value of each of these weights as a measure of how important they were for predicting activity. Of course, the range of these weights may vary drastically between models, so we also normalized the weights between zero and one and then implemented a similar applicability domain criterion compared to the one used by kNN. To do that, the calculated Euclidean distances between a test compound and the training set were weighted based on the normalized, absolute weights of each descriptor assigned by the SVM model. This produces a similar pseudoboundary to kNN, except the boundary would be extended for descriptors whose weight was close to zero and the boundary would be narrow for descriptors whose normalized, absolute weight was close to one.

Database Mining. Although we have employed four different QSAR methods for model building, only two, i.e., kNN and SVM, have produced predictive models as discussed in the Results section below. Thus, only validated models built with these two algorithms were used for the data mining studies. Three publicly available chemical databases were screened including the NCI database containing ca. 250 000 compounds,[43] the ChemDiv database containing ca. 450 000 compounds,[44] and the Maybridge database with 54 000 compounds.[45] MZ descriptors were generated for each compound in the databases and linearly normalized based on the maximum and minimum values of each descriptor in the training set. Each validated kNN and SVM model was then used to predict activity for the database compounds that were

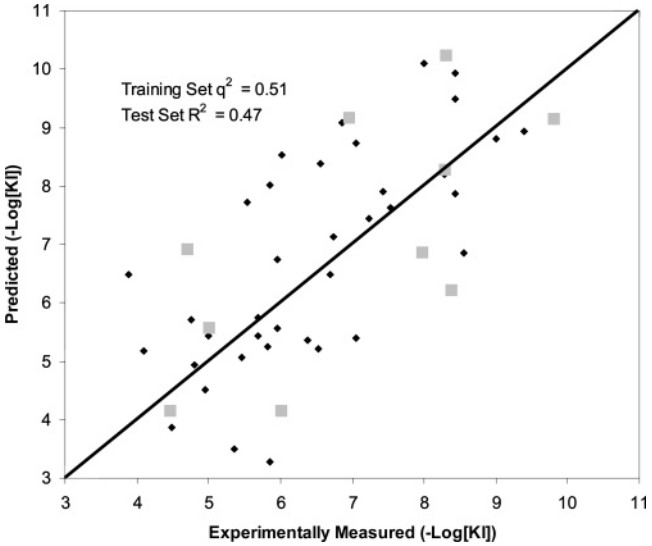


Figure 1. Comparison of actual vs predicted D1 antagonist binding affinity based on CoMFA models. The results are shown for both training (38 compounds; dark tilted squares) and test (10 compounds, gray squares).

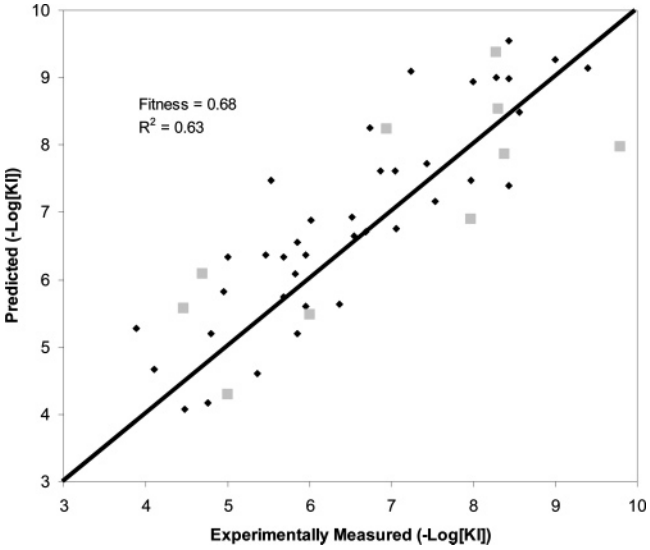


Figure 2. Comparison of actual vs predicted D1 antagonist binding affinity based on SA-PLS models. The results are shown for both training (38 compounds; dark tilted squares) and test (10 compounds, gray squares).

within the applicability domain for that model. The results for each individual prediction exercise were then combined and the mean predicted activity was calculated for each compound that was within the applicability domain of multiple models. The percentage of models that predicted each compound in the database and the standard deviation of those predictions were also recorded for each compound. It is our hypothesis that the higher the percentage of models with a stringent applicability domain that predict a compound's activity, the more likely the compound actually possesses the predicted activity. This may also apply to the standard deviation of the predictions made for a single compound.

The smaller the prediction variance across all models, the more confidence we have that the predicted biological activity for that compound is accurate. For these reasons, we selected a subset of compounds as hits that were predicted by at least 50% of the models and exhibited a small standard deviation across all models. We have also performed an additional estimate as to whether the hits resulting from database mining using variable selection models possessed the features essential (i.e., common) for all dopamine D_1 receptor ligands. This additional precaution was considered essential because variable selection procedure by default eliminates features (descriptors) that have the same values for all modeled compounds. To this end, the training set was used to calculate the standard deviation of the nearest neighbor distances using all descriptors. Any database hit that did not reside within 5 standard deviations of any training set compound in the entire descriptor space was discarded.

Results

QSAR Models Based on CoMFA. SCH23390 was used as a template to align the dopamine D_1 receptor antagonists. The CoMFA steric and electrostatic fields obtained using an sp^3 carbon with a +1 charge were calculated by multiplying the β-coefficient and standard deviation of columns in the QSAR table (stdev*coeff). From this, the CoMFA steric and electrostatic fields were employed to explain differences in the measured activity of the D_1 inhibitors.

Models were generated with q^2/R^2 values of 0.68/0.63 as shown in Figure 1. To prove that a method does not have the capability to fit random data to structural features, the ligand binding affinities were randomly rearranged and QSAR models were generated (cf. **Y**-randomization test in the Methods section). The best models for randomized data only produced a q^2 of 0.2 (data not shown), indicating that the models generated using true activity data were robust.

QSAR Models Developed with the SA-PLS Method. Predictive models were obtained with the training/test set fitness/R^2 values of 0.68/0.63, as shown in Figure 2. To prove that the model was robust, the ligand binding affinities were randomly rearranged and QSAR models were generated as discussed previously. The best models for randomized data only produced a fitness of 0.25 (data not shown), indicating that the models generated using true activity data are based on actual structure−activity correlations rather than a chance correlation.

QSAR Models Developed with the kNN Method. The kNN method produced highly predictive models with q^2/R^2 values of 0.7/0.76, as shown in Figure 3. To ensure that these models are not based on noise, the ligand activities were randomly shuffled within the training set. Models produced with these activity arrays lowered the mean q^2 value produced by kNN to approximately 0.2 (data not shown). This suggests that kNN does not have the ability to correlate descriptors from D_1 ligands to random activities and that the models produced represent a true structure−activity relationship.

QSAR Models Developed with the SVM Method. SVM produced statistically significant models with R^2_{train}/R^2_{test} values of 0.74/0.71, as shown in Figure 4. The optimum values of C and ϵ were found to be 160 and 0.3, respectively. The ϵ value, which corresponds to an acceptable error of 0.3 log unit, agrees with the expected experimental error that is commonly seen in ligand−receptor binding experiments. To ensure that the method does not have the capability to fit random data to structural features, the ligand binding affinities were randomly shuffled and SVM models were generated as discussed previously. The best models using randomized data only produced an R^2 for the test set of 0.3 (data not shown), providing further evidence that

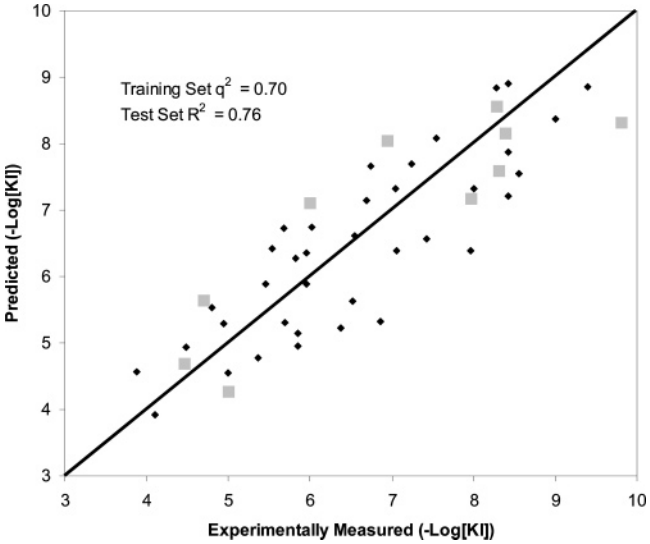


Figure 3. Comparison of actual vs predicted D_1 antagonist binding affinity based on kNN QSAR models. The results are shown for both training (38 compounds; dark tilted squares) and test (10 compounds, gray squares).

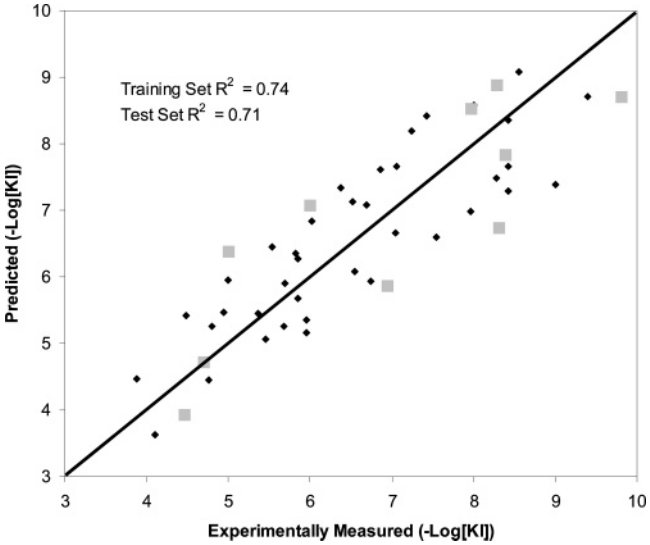


Figure 4. Comparison of actual vs predicted D_1 antagonist binding affinity based on SVM QSAR models. The results are shown for both training (38 compounds; dark tilted squares) and test (10 compounds, gray squares).

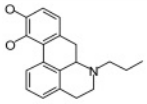
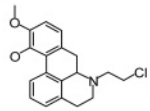
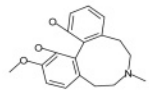
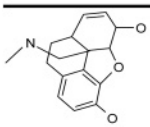
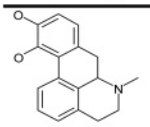
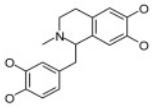
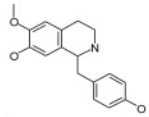
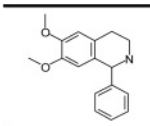
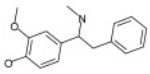
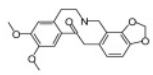
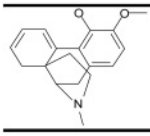
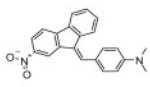
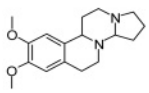
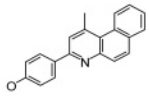
the models generated using real activity data are based on actual structure−activity relationships rather than a chance correlation.

Interpreting Predictive QSAR Models. Upon analysis of the kNN and SVM models built with MolconnZ descriptors, several commonly selected descriptors were found suggesting they played critical roles in relating chemical structure to D_1 binding affinity. Based on MolconnZ manual and personal communications with Dr. Lowell Hall, one of the principal developers of MolconnZ, these frequently selected descriptors fell into 5 classes: (1) The appearance of n4Pae23, n3Pad11, n2Pag22, n3Pad22, Tm, and Tg descriptors suggests a high importance of steric factors. These descriptors are based on the number of terminal vertex degrees; the high values of these descriptors suggest steric crowding. This is expected since several ligands in the dataset have hydrophobic accessory rings and amino-branched side chains that could create unfavorable steric interactions when bound to the D_1 receptor. (2) Atom type E-state descriptors such as SsCH3, SsI, and SssO appeared several times suggesting a high importance of electron accessibility for these atoms. In fact, it is well-known that the hydroxyls of the catechol ring on D_1 agonists are critical for forming hydrogen bonds to serine residues within the D_1 binding pocket.[46] (3) The appearance of the Hmax descriptor indicates the importance of polar hydrogen atoms in these molecules such as −OH, =NH, and −NH. This agrees with the well-known fact that the catechol hydroxyls and amine group of dopamine-like ligands play a large role in ligand binding. (4) The descriptor Xvch5 also appeared in most models suggesting the importance of a five-membered ring. There are 10 compounds in the training dataset that contain a five-membered ring. The presence of this five-membered ring, in the same relative position, may have caused steric interactions that led to the low binding affinity found for these compounds. (5) Last, several descriptors indicating the presence of iodine such as nI, nsI, and SsI appeared in several models. Only one compound in our training set contained an iodine atom located at the para position of the typical catechol ring suggesting that this substituent may play an important role in regulating receptor binding.

Database Mining with Predictive QSAR Models. Although robust models were generated in all cases, those obtained with both CoMFA and SA-PLS were not applied in database screening studies since their statistical parameters were below (CoMFA) or marginally above (SA-PLS) our acceptability criteria of q^2/R^2 greater than 0.6 for training/test set models, respectively. In addition, CoMFA is certainly unsuitable for the large scale virtual screening because of the difficulties associated with the conformational analysis and alignment that would have to be done for every compound in the screening set.

The ChemDiv, NCI, and Maybridge databases totaling over 750 000 compounds were screened by each of the validated (121) kNN and (106) SVM models with a defined applicability domain and predicted activities were averaged to yield a consensus value. This approach identified only 54 compounds with moderate to high predicted affinities out of the three chemical databases. There may in fact be additional D_1 ligands that our screening procedure missed; however, compounds in the database are not annotated with biological data. The chemical structures of the 54 compounds predicted as hits are shown in Table 1. Three compounds (shown in Table 1) did not contain a catechol ring found in most dopamine ligands, suggesting that they may be novel structural leads. The identification of a novel D_1 structural class may bring us closer to the ultimate goal of generating a D_1 selective ligand with high affinity. It should be noted that after manually searching with each of these compounds for known biological activity, five compounds were, indeed, found to have been previously characterized as dopamine D_1 receptor ligands[47−50] (see Table 2). The remaining 49 compounds could not be found in the scientific literature as ever tested against the D_1 receptor.

Table 1. Compounds Identified as Hits from Mining the NCI, Maybridge, and ChemDiv Databases[a]

Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted (-Log[KI])	Std. Dev. of Predictions
	18426-20-5	SVM	98.6	6.5	0.2
		kNN	81.8	6.9	1.1
	73378-11-7	SVM	85.2	7.5	0.4
		kNN	72.4	8.4	0.8
	139332-31-3	SVM	98.6	7.6	0.2
		kNN	61.2	7.3	1.1
	57-27-2	SVM	70.4	7.5	0.3
	314-19-2	kNN	56.5	7.2	1.1
	485-33-6	SVM	34.5	8.1	0.2
	2196-60-3	kNN	78.8	6.7	0.6
	4118-36-9	kNN	52.4	5.8	0.5
	6278-28-0	kNN	54.7	6.6	0.8
	6901-20-8	SVM	36.6	5.2	0.2
	15448-36-9	SVM	78.9	8.7	0.3
	18432-30-9	SVM	98.6	6.7	0.3
	20796-58-1	SVM	39.4	6.2	0.4
	22188-31-4	kNN	50.0	6.1	0.7

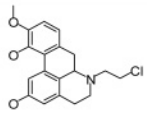
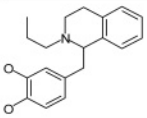
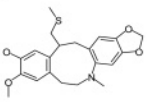
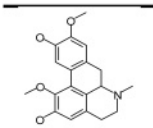
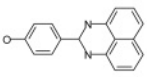
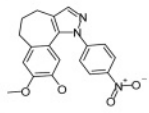
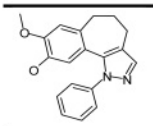
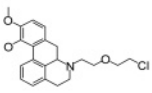
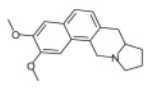
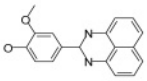
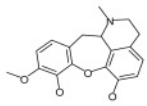
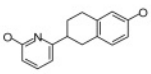
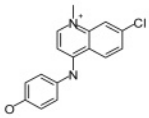
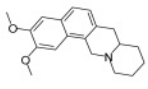
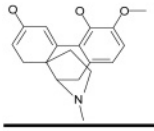
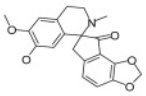
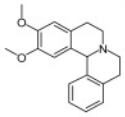
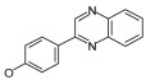
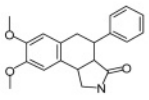
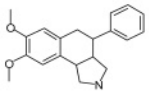
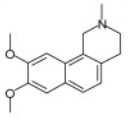
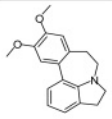
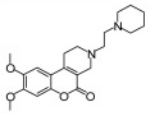
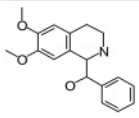
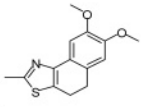
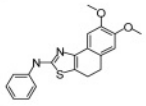
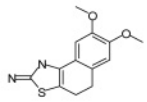
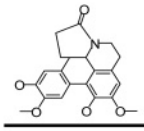
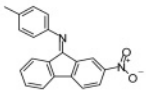
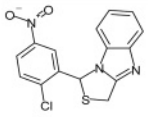
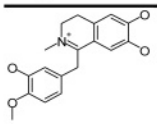
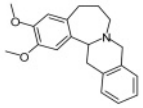
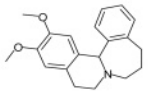
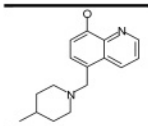
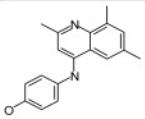
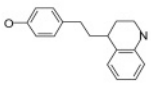
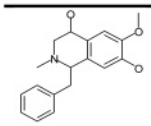
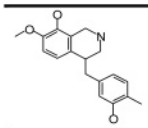
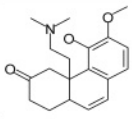
Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted (-Log[KI])	Std. Dev. of Predictions
	61752-28-1	SVM	81.7	7.7	0.4
		kNN	39.4	8.7	0.9
	87590-49-6	SVM	84.5	6.5	0.1
		kNN	86.5	6.7	0.6
	58939-39-2	SVM	85.2	5.3	0.3
	60755-81-9	SVM	31.0	8.3	0.2
	64573-23-5	kNN	60.0	5.7	0.5
	71007-75-5	SVM	94.4	7.5	0.4
	71007-73-3	SVM	59.9	7.3	0.3
	74427-12-6	SVM	85.2	6.7	0.2
	82589-60-4	SVM	76.1	5.1	0.4
	85557-36-4	kNN	60.6	5.7	0.7
	87035-67-4	SVM	98.6	7.5	0.3
	92850-90-3	kNN	58.2	6.3	0.6
	101272-47-3	kNN	95.9	7.1	0.6
	110288-05-6	SVM	78.9	5.4	0.6

Table 1 (Continued)

Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted (-Log[KI])	Std. Dev. of Predictions
	24887-59-0	SVM	42.3	9.0	0.3
	28230-70-8	SVM	47.2	9.1	0.5
	30562-57-3	SVM	85.2	5.0	0.4
	33707-91-4	kNN	52.9	6.1	0.3
	35202-58-5	SVM	51.4	6.8	0.3
	35202-51-8	SVM	84.5	5.9	0.4
	37707-95-2	SVM	71.1	6.2	0.6
	40360-74-5	SVM	84.5	6.1	0.6
	41806-52-4	SVM	41.5	6.9	0.3
	57420-79-8	kNN	55.3	6.4	0.6
	57595-89-8	SVM	38.0	6.3	0.4
	57595-87-6	SVM	62.7	5.7	0.3
	57595-86-5	SVM	38.0	6.7	0.2

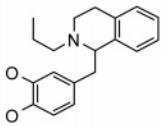
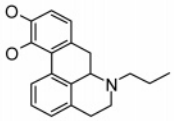
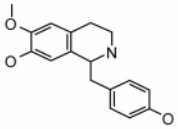
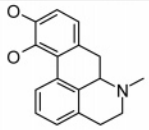
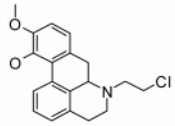
Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted (-Log[KI])	Std. Dev. of Predictions
	110717-92-5	SVM	38.0	7.8	0.3
	111796-34-0	SVM	98.6	6.4	0.3
	138226-16-1	SVM	71.8	3.3	0.4
	138596-84-6	SVM	82.4	8.4	0.3
	166537-49-9	SVM	85.2	5.2	0.5
	251306-28-2	SVM	85.2	4.3	0.4
	313395-24-3	kNN	74.1	6.3	0.5
	474087-83-7	kNN	51.8	7.5	0.8
	500565-28-6	kNN	97.1	6.0	0.3
	500566-75-6	kNN	62.9	7.4	0.9
	500574-70-9	kNN	51.8	7.4	0.9
	521282-59-7	SVM	98.6	8.8	0.4

[a] None of these compounds were present in our training set.

Discussion

For our dataset of 48 D_1 antagonists, thousands of QSAR models were built and as expected not all methods used were equally predictive. The top 10% of CoMFA models ranked on the basis of the test set accuracy exhibited R^2 ranging from 0.38 to 0.54, for SA-

Table 2. Selected Computational Hits Resulting from Mining the NCI, Maybridge, and ChemDiv Databases That Were Identified Previously as D_1 Antagonists with Measured D1 Affinity[a]

Chemical Structure	CAS #	Prediction Algorithm	% Models that Predicted each compound	Predicted (-Log[KI])	Std. Dev. of Predictions	Published (-Log[KI])
	87590-49-6	SVM	84.5	6.5	0.1	5.6[47]
		kNN	86.5	6.7	0.6	
	18426-20-5	SVM	98.6	6.5	0.2	6.5[48]
		kNN	81.8	6.9	1.1	
	2196-60-3	kNN	78.8	6.7	0.6	6.9[49]
	314-19-2	kNN	56.5	7.2	1.1	8[47]
	73378-11-7	SVM	85.2	7.5	0.4	<6.1[50]
		kNN	72.4	8.4	0.8	

[a] None of these compounds were present in the training set.

PLS models it was 0.55 to 0.69, kNN yielded ranges from 0.67 to 0.87, and SVM exhibited values from 0.68 to 0.73. We then chose to employ the validated SVM and kNN, exhibiting test set R^2 values greater than 0.7. This high number of models used a broad range of descriptors suggesting that they are diverse. By ranking predictions based on the percentage of models that have a particular compound within their applicability domains, one can obtain a sense of confidence that the predicted compound is in fact a ligand of the target of interest. For this study, only compounds that passed our global similarity screen and were predicted by over 50% of the models were considered candidate ligands for the D_1 receptor. Additional measures of confidence can be taken from the standard deviation of prediction between the models for any one algorithm, if multiple QSAR algorithms predict the same compound, and how close are the mean predictions between multiple algorithms that are validated against the initial dataset.

The use of database mining in this research project also helped point out current pitfalls in the process and gave us the chance to discover ways to avoid those pitfalls. The most important factor when performing a database search is the ability of a predictive model to search an external database for drug lead identification. If a model is unable to predict the activities of the training or test set compounds with high accuracy, then it most likely will not accurately predict activities of compounds from an external database. Another obstacle that was realized is that a small number of variable selection QSAR models may not be used efficiently for database mining. For example, kNN may select only 15 descriptors to produce an acceptable QSAR model. By default, a QSAR model reflects a correlation between variation in descriptor values and that of the target property. A small subset of essential descriptors, however, may exist that are relatively constant for the training set (and therefore may be essential determinants of the compound pharmacological class; cf. conventional pharmacophores), so they may not be included in the model. Therefore, if one searches a database with a small number of variable-selection QSAR models, a rough similarity screen of the database may be necessary in addition to model-based activity prediction. This should be done to ensure that compounds predicted active are not characterized by dissimilar values for constant essential descriptors that naturally were excluded in model development. In fact, after screening the chemical libraries, we first identified over 7100 compounds that were predicted by at least one SVM or kNN model with stringent applicability domain cutoffs. However, after a very coarse similarity screen using the entire descriptor space, most of these compounds were eliminated because they were not found within the applicability domain of at least 30% of the models. This reduced the list of potential hits to the final collection of 54 candidate dopamine D_1 receptor hits.

Conclusions

While CoMFA and SA-PLS produced reasonable models (q^2 or fitness and R^2 values greater than 0.5), the models with much higher predictive power were generated with kNN and SVM. Mining of the ChemDiv, NCI, and Maybridge chemical databases for D_1 antagonists, based upon kNN and SVM models, resulted in several hits with and without catechol rings. A few of

the predicted compounds are known D_1 ligands not included in the training set, validating the use of our approach for lead identification. Compounds lacking catechol rings that are predicted to be D_1 ligands may in fact lead to a novel structural class that exhibits an atypical pharmacological profile. Although further experimental work is needed to verify this hypothesis, the identification of known ligands suggests that models produced in this study are capable of detecting novel D_1 compounds from large chemical databases.

As we discussed briefly at the onset of this paper, our general approach to QSAR modeling goes beyond the traditional boundaries of this method. Although QSAR modeling is generally regarded as a ligand optimization approach that may lead to rational design of novel compounds, the examples of rationally designed compounds are rare in any traditional QSAR modeling paper. Most of the publications present models that are capable of reproducing training set compound activity with high accuracy (in some cases, test set compound predictions are included but those already have their biological activity determined). Thus, a typical outcome of a traditional QSAR modeling study is a set of statistical characteristics such as q^2, R^2, F value, etc., mostly for the training set, which provide little help to chemists interested in the design of novel molecules (CoMFA presents a notable exception by formally providing structural design hypothesis based on "fields").

The approach applied herein to QSAR modeling of D_1 antagonists does not stop when one could obtain a statistically significant training set model. Our approach places the emphasis of the *entire* QSAR modeling study on making reliable predictions of *chemical structures* expected to have the desired biological activity, *rather than* on respectable statistical characteristics of (training set) models. These predicted structures either are already available in existing chemical databases or are synthetically feasible (i.e., included in virtual combinatorial chemical libraries, which can also be mined with QSAR models). We believe that this extended view of the entire QSAR modeling approach exemplified both by our recent studies of anticonvulsants[42] and those of D_1 antagonists presented in this paper brings the focus of the modeling closer to the needs of medicinal chemists who both supply computational chemists with experimental structure−activity data and expect novel structures rather than equations and statistical parameters in return. We suggest that our approach that combines predictive QSAR modeling and database mining provides an important general avenue toward drug discovery that can be explored for many pharmacological datasets.

Acknowledgment. We would like to thank Alexander Golbraikh, Min Shen, Yunde Xiao, and Yuanyuan Qiao for their assistance and constructive criticism throughout this research project. We are especially grateful to Prof. Lowell Hall for his valuable help with the interpretation of MolConnZ descriptors that were found important in our models. The project was supported by the NIH grants MH60328, MH40537, and GM066940.

Supporting Information Available: Chemical structures and competitive binding affinities used to develop the structure−activity models. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) EduSoft, LLC. *MolconnZ*, version 4.05; http://www.eslc.vabiotech.com/ [4.05], 2003.
(2) Kebabian, J. W.; Calne, D. B. Multiple receptors for dopamine. *Nature* **1979**, *277*, 93−96.
(3) Strange, P. G. *Brain Biochemistry and Brain Disorders*; Oxford University Press: New York, 1993.
(4) Seeman, P.; Bzowej, N. H.; Guan, H. C.; Bergeron, C.; Reynolds, G. P.; Bird, E. D.; Riederer, P.; Jellinger, K.; Tourtellotte, W. W. Human brain D1 and D2 dopamine receptors in schizophrenia, Alzheimer's, Parkinson's, and Huntington's diseases. *Neuropsychopharmacology* **1987**, *1*, 5−15.
(5) Creese, I.; Iversen, S. D. Blockage of amphetamine induced motor stimulation and stereotypy in the adult rat following neonatal treatment with 6-hydroxydopamine. *Brain Res.* **1973**, *55*, 369−382.
(6) Phillips, A. G.; Fibiger, H. C. Dopaminergic and noradrenergic substrates of positive reinforcement: differential effects of d- and l-amphetamine. *Science* **1973**, *179*, 575−577.
(7) Pijnenburg, A. J.; Honig, W. M.; Van der Heyden, J. A.; Van Rossum, J. M. Effects of chemical stimulation of the mesolimbic dopamine system upon locomotor activity. *Eur. J. Pharmacol.* **1976**, *35*, 45−58.
(8) Ungerstedt, U.; Arbuthnott, G. W. Quantitative recording of rotational behavior in rats after 6-hydroxy-dopamine lesions of the nigrostriatal dopamine system. *Brain Res.* **1970**, *24*, 485−493.
(9) Hoffman, B. T.; Kopajtic, T.; Katz, J. L.; Newman, A. H. 2D QSAR modeling and preliminary database searching for dopamine transporter inhibitors using genetic algorithm variable selection of Molconn Z descriptors. *J. Med. Chem.* **2000**, *43*, 4151−4159.
(10) Ghosh, D.; Snyder, S. E.; Watts, V. J.; Mailman, R. B.; Nichols, D. E. 9-Dihydroxy-2, 3, 7, 11b-tetrahydro-1H-naph[1,2,3-de]-isoquinoline: a potent full dopamine D1 agonist containing a rigid-beta-phenyldopamine pharmacophore. *J. Med. Chem.* **1996**, *39*, 549−555.
(11) Tomita, T.; Inami, Y.; Terada, Y. Structure-activity relationships of dopamine- and norepinephrine-uptake inhibitors. *Chem. Pharm. Bull. (Tokyo)* **1990**, *38*, 1563−1569.
(12) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure−activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.
(13) Lien, E. J.; Nilsson, J. L. QSAR of N-alkylated 2-aminotetralins as central dopamine receptor stimulating agents. *Acta Pharm. Suec.* **1983**, *20*, 271−276.
(14) Norinder, U.; Hogberg, T. A quantitative structure-activity relationship for some dopamine D2 antagonists of benzamide type. *Acta Pharm. Nord.* **1992**, *4*, 73−78.
(15) Van de, W. H.; el Tayar, N.; Testa, B.; Wikstrom, H.; Largent, B. Quantitative structure-activity relationships and eudismic analyses of the presynaptic dopaminergic activity and dopamine D2 and sigma receptor affinities of 3-(3-hydroxyphenyl)piperidines and octahydrobenzo[f]quinolines. *J. Med. Chem.* **1987**, *30*, 2175−2181.
(16) Wilcox, R. E.; Tseng, T.; Brusniak, M. Y.; Ginsburg, B.; Pearlman, R. S.; Teeter, M.; DuRand, C.; Starr, S.; Neve, K. A. CoMFA-based prediction of agonist affinities at recombinant D1 vs D2 dopamine receptors. *J. Med. Chem.* **1998**, *41*, 4385−4399.
(17) Mottola, D. M.; Laiter, S.; Watts, V. J.; Tropsha, A.; Wyrick, S. D.; Nichols, D. E.; Mailman, R. B. Conformational analysis of D1 dopamine receptor agonists: pharmacophore assessment and receptor mapping. *J. Med. Chem.* **1996**, *39*, 285−296.
(18) Charifson, P. S.; Bowen, J. P.; Wyrick, S. D.; Hoffman, A. J.; Cory, M.; McPhail, A. T.; Mailman, R. B. Conformational analysis and molecular modeling of 1-phenyl-, 4-phenyl-, and 1-benzyl-1,2,3,4-tetrahydroisoquinolines as D1 dopamine receptor ligands. *J. Med. Chem.* **1989**, *32*, 2050−2058.
(19) Tropsha, A. Recent Trends in Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 6th ed.; Abraham, D., Ed.; John Wiley & Sons: New York, 2003; pp 49−77.
(20) Wilson, J.; Lin, H.; Fu, D.; Javitch, J. A.; Strange, P. G. Mechanisms of inverse agonism of antipsychotic drugs at the D(2) dopamine receptor: use of a mutant D(2) dopamine receptor that adopts the activated conformation. *J. Neurochem.* **2001**, *77*, 493−504.
(21) Schetz, J. A.; Benjamin, P. S.; Sibley, D. R. Nonconserved residues in the second transmembrane-spanning domain of the D(4) dopamine receptor are molecular determinants of D(4)-selective pharmacology. *Mol. Pharmacol.* **2000**, *57*, 144−152.

(22) Sartania, N.; Strange, P. G. Role of conserved serine residues in the interaction of agonists with D3 dopamine receptors. *J. Neurochem.* **1999**, *72*, 2621−2624.
(23) Alberts, G. L.; Pregenzer, J. F.; Im, W. B. Contributions of cysteine 114 of the human D3 dopamine receptor to ligand binding and sensitivity to external oxidizing agents. *Br. J. Pharmacol.* **1998**, *125*, 705−710.
(24) Lundstrom, K.; Turpin, M. P.; Large, C.; Robertson, G.; Thomas, P.; Lewell, X. Q. Mapping of dopamine D3 receptor binding site by pharmacological characterization of mutants expressed in CHO cells with the Semliki Forest virus system. *J. Recept. Signal Transduction Res.* **1998**, *18*, 133−150.
(25) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(26) Tropsha, A.; Cho, S. J.; Zheng, W. "New Tricks For an Old Dog": Development and application of novel QSAR methods for rational design of combinatorial chemical libraries and database mining. In *Rational Drug Design: Novel Methodology and Practical Applications*; ACS Symposium Series 719; Parrill, A. L., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 2001.
(27) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y. D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−595.
(28) Tripos Inc. *Sybyl User's Manual*, version 7.8; Tripos, Inc.: St. Louis, MO, 2002.
(29) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(30) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(31) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
(32) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(33) Charifson, P. S.; Wyrick, S. D.; Hoffman, A. J.; Simmons, R. M.; Bowen, J. P.; McDougald, D. L.; Mailman, R. B. Synthesis and pharmacological characterization of 1-phenyl-, 4-phenyl-, and 1-benzyl-1,2,3,4-tetrahydroisoquinolines as dopamine receptor ligands. *J. Med. Chem.* **1988**, *31*, 1941−1946.
(34) Minor, D. L.; Wyrick, S. D.; Charifson, P. S.; Watts, V. J.; Nichols, D. E.; Mailman, R. B. Synthesis and molecular modeling of 1-phenyl-1,2,3,4-tetrahydroisoquinolines and related 5,6,8,9-tetrahydro-13bH-dibenzo[a,h]quinolizines as D1 dopamine antagonists. *J. Med. Chem.* **1994**, *37*, 4317−4328.
(35) Schulz, D. W.; Wyrick, S. D.; Mailman, R. B. [3H]SCH23390 has the characteristics of a dopamine receptor ligand in the rat central nervous system. *Eur. J. Pharmacol.* **1984**, *106*, 211−212.
(36) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241−253.
(37) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Diversity* **2002**, *5*, 231−243.
(38) Wold, S. a. E. L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1995; pp 309−318.
(39) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics;* John Wiley & Sons: New York, 1986.
(40) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(41) Cho, S. J.; Zheng, W.; Tropsha, A. Rational combinatorial library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity probe and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 259−268.
(42) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356−2364.
(43) National Cancer Institute. NCI Cancer Database. http://dtp.nci.nih.gov/docs/3d_database/structural_information/smiles_strings.html. 2004.
(44) Chemical Diversity. ChemDiv Chemical Database. www.chemdiv.com. 2004.
(45) Maybridge Chemical Company. Maybridge. http://www.daylight-.com/products/databases/Maybridge.html. 2004.
(46) Lewis, M.; Jassen, A.; Oloff, S.; Hoffman, B.; Henage, L.; Nicholas, R.; Dawson, E.; Tropsha, A.; Lybrand, T.; Nichols, D.; Mailman, R. Docking of D1 dopamine agonists reveals specific residues involved in agonist recognition and receptor activation. *Mol. Pharmacol.* **2004**.
(47) Bradbury, A. J.; Costall, B.; Naylor, R. J.; Neumeyer, J. L. Motor inhibition induced by aporphine derivatives in the mouse. *J. Pharm. Pharmacol.* **1983**, *35*, 494−499.
(48) Gao, Y. G.; Ram, V. J.; Campbell, A.; Kula, N. S.; Baldessarini, R. J.; Neumeyer, J. L. Synthesis and structural requirements of N-substituted norapomorphines for affinity and activity at dopamine D-1, D-2, and agonist receptor sites in rat brain. *J. Med. Chem.* **1990**, *33*, 39−44.
(49) Bermejo, A.; Protais, P.; Blazquez, M.; Rao, S.; Zafra-Polo, C.; Cortes, D. Dopaminergic isoquinoline alkaloids from roots of Xylopia papuana. *Nat. Prod. Lett.* **1995**, *6*, 57−62.
(50) Baldessarini, R. J.; Kula, N. S.; Arana, G. W.; Neumeyer, J. L.; Law, S. J. Chloroethylnorapomorphine, a proposed long-acting dopamine antagonist: interactions with dopamine receptors of mammalian forebrain in vitro. *Eur. J. Pharmacol.* **1980**, *67*, 105−110.

JM049116M