Published in final edited form as:
*J Chem Inf Model*. 2011 March 28; 51(3): 521–531. doi:10.1021/ci100399j.

# Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS) and Its Application on Modeling Ligand Functionality for 5HT-subtype GPCR Families

**Chao Ma**[1,2,3], **Lirong Wang**[2,1,3], and **Xiang-Qun Xie**[2,1,3,*]

[1] Department of Computational Biology, Joint Pitt/CMU Computational Biology Program, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15260, USA

[2] Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, Pittsburgh, PA 15260, USA

[3] Pittsburgh Center for Chemical Methodologies & Library Development (PCMLD) and Drug Discovery Institute, University of Pittsburgh, Pittsburgh, PA 15260, USA

## Abstract

Advanced high-throughput screening (HTS) technologies generate great amounts of bioactivity data, and this data needs to be analyzed and interpreted with attention to understand how these small molecules affect biological systems. As such, there is an increasing demand to develop and adapt cheminformatics algorithms and tools in order to predict molecular and pharmacological properties based on these large datasets. In this manuscript, we report a novel machine-learning-based ligand classification algorithm, named Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS), for data-mining and modeling of large chemical datasets to predict pharmacological properties in an efficient and accurate manner. The performance of LiCABEDS was evaluated through predicting GPCR ligand functionality (agonist or antagonist) using four different molecular fingerprints, including Maccs, FP2, Unity and Molprint 2D fingerprints. Our studies showed that LiCABEDS outperformed two other popular techniques - classification tree and Naive Bayes classifier - on all four types of molecular fingerprints. Parameters in LiCABEDS, including the number of boosting iterations, initialization condition, and a “reject option” boundary, were thoroughly explored and discussed to demonstrate the capability of handling imbalanced datasets, as well as its robustness and flexibility. In addition, the detailed mathematical concepts and theory are also given to address the principle behind statistical prediction models. The LiCABEDS algorithm has been implemented into a user-friendly software package that is accessible online at http://www.cbligand.org/LiCABEDS/.

## 1. INTRODUCTION

As a complement to modern high-throughput screening, one of the primary goals of virtual screening and cheminformatics techniques is to explore the enormous chemical and biological properties in a time-efficient manner as well as to help reduce the cost of

*Author to whom correspondence should be addressed: Sean Xie, xix15@pitt.edu; Tel.: +1-412-383-5276; Fax: +1-412-383-7436.

Supporting Information Available: The Supporting Information is divided in four parts. The structures of the compounds and the generation of training and testing datasets are given in part 1. An outline of computational protocol is provided in Part 2. Part 3 lists the performance of computational models and molecular descriptors on each testing dataset. Part 4 describes the effect of training iterations on training and testing errors, together with more details regarding cross-validation. This information is available free of charge via the Internet at http://pubs.acs.org/

experimental screening[1]. In particular, great emphasis is placed on the "druggability" or "drug-likeness" of compounds using cheminformatics tools in the early stages of drug development, with the hope of increasing the probability of "lead" compounds, or their derivatives, to pass through the later phases of drug clinical trials[2].

Despite numerous molecular properties and various mathematical models, the prediction of ligand binding activity and biological properties can be addressed by two types of approaches: a classification model for categorical response and a regression model for continuous response. For example, some pharmaceutical properties, such as mutagenicity, can be modeled by ligand classification[3]. To build up a quantitative structure-property relationship (QSPR) model, pattern recognition methodology can be applied to map molecular descriptors to continuous value or categorical value via regression or classification [4]. These molecular descriptors are usually binary or continuous vectors describing various aspects of molecular attributes or structural patterns. Many ligand properties pertaining to drug discovery have been successfully modeled with hundreds of molecular descriptors or fingerprints through statistical or machine learning techniques[5]. As one of the representative regression models, Comparative Molecular Field Analysis (CoMFA)[6] applies partial least square regression to make predictions from the principal components that are linear combinations of electrostatic and steric energy fields at 3D grids. CoMFA was successfully applied in the prediction of membrane flux[7], modeling structure-pharmacokinetic relationships[8] and antagonist binding affinities at cannabinoid receptor subtype[9]. A CoMFA model was also developed to distinguish 5-$HT_{1A}$ agonists and antagonists[10], which is also one of the focuses in this manuscript. Another classification technique, Naive Bayes classifier, has also been used to model quantitative structure-selectivity relationships[11].

Despite the advance of cheminformatics methodology, it remains a challenge to develop a robust, reliable, and interpretable ligand classifier to tackle different scenarios in computer-aided drug design. Although any regression method like CoMFA can be adapted as a ligand classifier, such an approach often suffers from overfitting due to the model complexity of the regression method. In addition, the ability to find a 3D bioactive conformer remains as one of the limits [12]. Many existing modeling methods may require researchers to perform variable selection. However, variable selection is still a complicated procedure that ultimately has a large effect on the final predictive model. Free parameters are manually specified in most computational models, for example, the number of components in the CoMFA method. Besides, cross-validation is often carried out to find optimal values of those parameters, but this practice could be computationally inefficient, and its performance also heavily relies on the choice of cross-validation datasets.

Thus, reliable and robust ligand classifiers are needed to aid scientists and researchers in discovering compounds with desired properties in both the lead discovery as well as the drug development process. In this regard, we report our recent work in developing boosting-based classifier for prediction of pharmaceutical properties. The adaptive boosting algorithm, or Adaboost, introduced by Freund and Schapire[13], is a general method used to produce a "strong" classifier by combining a series of "weak learners". Sharing certain resemblance with the support vector machine (SVM) algorithm, Adaboost is also a maximum-margin classifier and tends not to overfit the training data[14]. Advantageously, the number of boosting rounds is the only essential parameter in Adaboost training, which simplifies the computational process of machine learning algorithms. In spite of the advantages, this algorithm has rarely been applied and discussed in drug discovery. In this study, we presented a novel ligand classifier, LiCABEDS, by adaptively boosting sets of decision stumps based on 2D molecular fingerprints. In our established algorithm, important features are automatically selected and weighted accordingly to build "weak learners" in

model training. The performance and the characteristics of our novel algorithm are demonstrated and tested through the application on modeling ligand functionality for serotonin receptors or 5-hydroxytryptamine (5HT) receptors, belonging to an important family of G protein-coupled receptors (GPCRs). In addition, across-target studies indicate the potential application of LiCABEDS on orphan receptors. In this manuscript, we also describe the detailed mathematical concepts of the LiCABEDS algorithm. It is anticipated that LiCABEDS, as a general-purpose ligand classifier, can be applied to model more biochemical and pharmacological properties. The model development is free of conformation search and is readily automated with the robustness of 2D molecular fingerprints. Its performance and application are described below. Finally, the algorithm is implemented in a freely available and user-friendly software package, allowing the easy importing of datasets and model development. The fully functioning software package is available online to the scientific community.

## 2. MATERIALS, METHODS AND CALCULATIONS

### 2.1. Computational Methods

The detailed mathematics concepts of Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS) and its application on modeling ligand functionality are described below. As case studies, LiCABEDS was used to model the ligand functionality for the 5HT-subtype GPCR families by predicting a given ligand to be either an agonist or an antagonist. For a parallel study, the performance of LiCABEDS was compared to two other popular data-mining methods: classification tree[15] and Naive Bayes classifier[16]. The underlying theory of these two methods is also introduced in this section.

**2.1.1. Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS)—**Adaptive boosting, initially introduced by Yoav Freund and Robert Schapire[13], is a general machine learning technique to create a strong classifier by combining a series of "weak learners" for improving the accuracy of prediction. In LiCABEDS, "decision stumps" are designed to be the weak learners. As illustrated in Figure 1a, the "decision stump" denotes a heuristic classification hypothesis that a compound will be classified as an agonist (+1) if the $i^{th}$ bit of fingerprint ($x_i$ is equal to a target value (t); or as an antagonist (−1), otherwise.

Instead of using the graphic representation, a "decision stump" can be formulated by a function: $y(x, i, t) = 2I(x_i = t) - 1$, where $I$ is an indicator function, $I(Z) = 1$ if the statement $Z$ is true; $I(Z) = 0$, otherwise. $x$ is the molecular fingerprint vector, $i$ is the index of the fingerprint, and $t$ is the target value. For example, if $x_i$, the $i^{th}$ bit of fingerprint, is equal to the target value $t$, then $I(x_i = t) = 1$ and $y(x, i, t) = 1$ (agonist). If $x_i$, the $i^{th}$ bit of fingerprint is different from $t$, then $I(x_i = t) = 0$ and $y(x, i, t) = -1$ (antagonist).

Different from many other machine-learning algorithms, LiCABEDS, as an ensemble method, is designed to achieve stronger classification power by boosting many "weak" classification hypotheses. As illustrated in Figure 1b, a series of "decision stumps" with corresponding weights $a_m$ vote for the final prediction, which can be formulated as the weighted summation of the outcome of every "decision stump":

$$Y_M = sign\left(\sum_m^M a_m y_m(x, i_m, t_m)\right) \quad (1)$$

sign(z) = 1 if $z > 0$, or −1 otherwise. The unknown variables, $a_m$, $i_m$ and $t_m$, for each weak classifier *m* can be "learned" from training datasets using the following algorithm:

1. Initialize the sample weights for each training compound *n*, $w_n = 1/N$, $n = 1, \ldots, N$, *N* is the total number of training compounds.
2. For each round of calculation $m = 1, \ldots, M$

   Find $i_m$, $t_m$ for weak learner $y_m$ by minimizing the weighted error function

$$(i_m, t_m) = \arg\min_{i_m, t_m} \sum_{n=1}^{N} w_n I(y_m(X_n, i_m, t_m) \neq l_n) \quad (2)$$

   where argmin is the function to return the arguments which minimize the object function, $X_n$ is the descriptor vector for compound *n;* $l_n = \pm 1$ is the label of compound *n*. $i_m$, $t_m$ uniquely define a "decision stump", and their optimal values can be found by enumerating all possible combinations of $i_m$, $t_m$.

   Evaluate the quantities:

$$\varepsilon = \frac{\sum_{n=1}^{N} w_n I(y_m(X_n, i_m, t_m) \neq l_n)}{\sum_{n=1}^{N} w_n}; a_m = \ln \frac{1-\varepsilon}{\varepsilon} \quad (3)$$

   $a_m$ becomes the weight for the "decision stump" *m*. Then update the weights of training compounds for next round of calculation:

$$w_n \leftarrow w_n \exp(a_m I(y_m(X_n, i_m, t_m) \neq l_n)) \quad (4)$$

The number of training steps, *M*, is the only parameter that must be specified manually in the algorithm. Cross-validation is one of the options to specify the optimal value of *M*, $M_{optimal}$. Training error is steadily minimized as M increases. While the training algorithm aims to minimize the exponential loss function, boosting algorithm may have potential to overfit the training data as pointed by others [17]. Despite such potential, Freund and Schapire have shown the underlying mechanism that adaptive boosting does not often suffer from overfitting [14]. Discussion is given later on the difference between a large value of *M* (by default) and $M_{optimal}$ in order to address the overfitting issue.

Training compound datasets may potentially be overwhelmed by one category of training samples. In this case, the majority class is usually favored in the prediction. To minimize the effect of disproportionate training samples in each category, balanced class weight can be set as an alternative initialization condition to equal initial weight. In other words, the total weights for each class are equal at the initialization step: $\sum_{n=1}^{N} w_n I(l_n = 1) = \sum_{n=1}^{N} w_n I(l_n = -1)$. For example, all the labeled agonists in the training set may have initial weights $1/N_{+1}$, where

$N_{+1}$ is the total number of agonists in the training data. Similarly, all of the antagonists may have an initial weight $1/N_{-1}$.

Heuristically, the absolute value of $A=\sum_{m}^{M} a_m y_m(x, i_m, t_m)$ indicates the degree of confidence in the prediction, because a relatively large population of "decision stumps" vote for the corresponding class. On the other hand, a low absolute value of A indicates uncertainty in the prediction. Better prediction accuracy is anticipated by avoiding uncertain cases, which we also refer to as "reject option". In our study, a prediction is only made for a test compound if $|A|>c$, where $c$ is rejection threshold. Otherwise, an "unknown" label is assigned to the test compound.

**2.1.2. Classification Tree—**Classification tree is a straightforward and effective data-mining technique. It has been widely applied to different areas of computer-aided drug design, such as virtual screening[18], drug-likeness prediction[19] and ligand blood-brain-barrier passage[20].

A classification tree consists of a set of split criterions and leaf nodes. The split criterions control the region that a ligand belongs to, while the leaf nodes represent classification hypotheses that are derived from training datasets in the same regions. The structure of a decision tree can be induced from training datasets in a greedy manner. By recursively partitioning the entire training dataset into regions, impurity *impurity*(*t*) is minimized regarding each possible partitioning *t*:

$$\min_t impurity\ (t)=\sum_s \lambda_s(t),$$

where *s* is the new region created from split *t*, and $\lambda_s(t)=1-\sum_{j=0}^{1}\widehat{p_s}(j)^2$. In this study, splitting rule is chosen from $x_i = 0$ *or* $x_i = 1$, where $x_i$ is the $i^{th}$ bit of descriptor vector. $\hat{p_s}(j)$ is the maximum likelihood estimator of a ligand being j, j = *0* or *1* (*0* represents antagonists, and *1* represents agonists), in region *S*. Training data can be perfectly fitted by growing the tree until 100% purity is achieved at each node. To avoid overfitting, k-fold cross-validation is commonly employed to control the "height" of a decision tree. After "pruning" the whole tree according to the cross-validation score that is defined as the percentage of correct predictions on cross-validation sets in this study, the optimal tree structure will be used to make predictions for novel ligands.

**2.1.3. Naive Bayes Classifier—**The Naive Bayes classifier method is a simple classification method based on applying Bayes' theorem with independence assumptions[16]. The method relies on the assumption that the presence or absence of a particular feature or class is unrelated to the presence or absence of any other feature. This independence assumption, with regard to molecular fingerprints simplifies the estimation of the likelihood function, which makes the method applicable to many computer-aided drug design tasks, such as virtual screening[21] and selectivity prediction[11]. In this study, Naive Bayes classifier was used to model the probability of one ligand being an agonist or an antagonist, given its molecular fingerprint: Pr(*Cl*|*Fp*) where *Fp* is the molecular fingerprint vector, and *Cl* = 1 for agonist or 0 for antagonist. By applying Bayesian theory, $\Pr(Cl|Fp) \propto \Pr(Fp|Cl)\Pr(Cl)$, the predicted class of a given ligand is antagonist, $\hat{Cl} = 0$, if $\Pr(Cl = 0|Fp) \geq 0.5$; and $\hat{Cl} = 1$, otherwise. Pr(*Fp*|*Cl* ) can be approximated by applying the independence assumption to molecular fingerprints:

$$\Pr(Fp|Cl)=\prod_i \Pr(Fp_i|Cl) \text{ where } Fp_i \text{ is the } i^{th} \text{ bit of fingerprint.}$$

where *Fp* is the $i^{th}$ bit of fingerprint.

Due to the difference between Molprint 2D and other types of fingerprints, the equation used in calculating the likelihood was also different. For example, we had Molprint 2D string "2;0-1-0; 2;0-2-2;" and Maccs fingerprint "0101" for a testing compound (only for illustration). The likelihood of the Molprint 2D fingerprint was calculated as:

$$\Pr(Fp=2;0-1-0;2;0-2-2;|Cl)=\Pr(Fp_1=2;0-1-0;|Cl)\times\Pr(Fp_2=2;0-2-2;|Cl)$$

The likelihood of Maccs key was calculated as:

$$\Pr(Fp=0101|Cl)=\Pr(Fp_1=0|Cl)\times\Pr(Fp_2=1|Cl)\times\Pr(Fp_3=0|Cl)\times\Pr(Fp_4=1|Cl)$$

The presence or absence of predefined Maccs features are considered in the likelihood calculation, while only present Molprint 2D features are modeled in the calculation.

## 2.2. Calculation

### 2.2.1. Dataset Preparation, Molecular Fingerprint and Computation Protocol—

To evaluate the performance of LiCABEDS, all the labeled human 5-$HT_{1A}$, 5-$HT_{1B}$, 5-$HT_{1D}$, and 5-$HT_{4R}$ agonists and antagonists were retrieved from the GLIDA database [22]. The ligand quantity and their properties are summarized in Table 1 (properties were calculated using the Sybyl8.0 www.tripos.com). With the published compound datasets, the prediction accuracy of different data-mining methods along with different molecular descriptors was assessed on the labeled agonists and antagonists of the human 5-$HT_{1A}$ subtype G-Protein Coupled Receptor (GPCR) by ten rounds of calculation. For each round of calculation, three classification methods were compared, including LiCABEDS, classification tree, and Naive Bayes classifier. Each was trained on the same randomly selected training compounds. The set of training compounds was composed of 75% labeled agonists and antagonists (827 5-$HT_{1A}$ agonists and 446 5-$HT_{1A}$ antagonists). The remaining 25% ligands (275 5-$HT_{1A}$ agonists and 149 5-$HT_{1A}$ antagonists) were used as a testing dataset in order to evaluate the prediction accuracy of different methods. The prediction accuracy was estimated by comparing the predictions to the real ligand labels (agonists or antagonists). With Molprint 2D[21, 23] as descriptor, the across-target ligand functionality prediction was also made by LiCABEDS. In this case, a LiCABEDS model was trained on all the labeled human 5-$HT_{1A}$ ligands (totally 1697 ligands), and then predictions were made on the ligands for human 5-$HT_{1B}$, 5-$HT_{1D}$, and 5-$HT_{4R}$ receptors. The details regarding the datasets can be found in Supporting Information Part 1.

In this study, four types of molecular fingerprints were generated for each compound, including Maccs key[24], Unity (www.tripos.com), FP2[25], and Molprint 2D [21, 23] fingerprint. The Maccs key fingerprint was calculated by Chemistry Development Kit (CDK)[26] and the Unity fingerprint was calculated by Sybyl 8.0 (www.tripos.com). Openbabel[25] was used to generate the FP2 fingerprint, and Molprint 2D package was used to generate the Molprint 2D fingerprint. Variable selection was not carried out before model trainings, so all the dimensions of these molecular fingerprints were exposed to the learning algorithms. To ensure a fair amount of overlapping in Molprint 2D patterns, a three-layer atom environment

was used for predicting ligand functionality for the human 5-$HT_{1A}$ receptor, while a two-layer atom environment was preferred for across-target predictions. Each feature defined by Molprint 2D was mapped to a unique bit in the descriptor vector by an in-house program (see Supporting Information Part 2).

The implementation of the classification tree presented in this study came from a Tree package in "R" [15, 27]. To avoid overfitting, a tree node was not split unless more than ten training compounds were observed in the parent node and more than five were present in both child nodes. Lastly, each classification tree was "pruned" according to ten-fold cross-validation scores. The implementation of Naive Bayes classifier on the Molprint 2D fingerprint was from the Molprint 2D software package. An in-house Naive Bayes classifier was also developed for other fingerprints. The implementation of LiCABEDS is discussed in the following section (2.3. LiCABEDS Software Package).

LiCABEDS models were initially developed using a large value of *M* (*M*=10000), for all fingerprint types to ensure a convergence of training error. Furthermore, the influence of *M* was studied, and optimal values of M were determined by running cross-validation with 10% of training compounds as a cross-validation set. LiCABEDS models were developed using balanced weights when compared to Classification Tree and Naive Bayes classifier on human 5-$HT_{1A}$ ligand datasets. Next, equal weights were compared to balanced weights as initialization conditions, which was conducted on the same training and testing datasets. Finally, the prediction accuracy was assessed with the "reject option" boundary ranging from 0 to 3.

### 2.3. LiCABEDS Software Package

A user-friendly interface was developed for LiCABEDS in order to simplify the steps involved in project management, model training and making predictions. The software integrates automated importing of training and testing datasets. The training module features automatic cross-validation, flexible initialization and interruptible model development. The "Reject option" is also implemented for making predictions. The graphical user interface allows for flexible model editing, prediction browsing, and result exporting. In addition, a work session can be saved to local hard disk, so that the previous workspace can be restored by the program. The program has been tested on Intel i7 860 2.8GHz CPU to evaluate the computational time. To iterate 10000 steps on 1697 compounds, model training takes 44 minutes for Molprint 2D fingerprint, 5 minutes for FP2 or Unity fingerprints, and 1 minute for Maccs fingerprint. The calculation time for model training is related to the amount of training samples, the number of boosting iterations and the dimension of descriptors. Once the model is established, the predictions can be made instantly by the program. More details regarding the software can be found in Supporting Information Part 2, and the user manual is available on our website.

## 3. RESULTS AND DISCUSSION

According to the distribution of physical properties listed in Table 1, simple classification hypotheses do not distinguish agonists from antagonists very well, if even at all. On the other hand, molecular fingerprints encode a large amount of chemical information regarding structural patterns of small molecules. The strength of the LiCABEDS method lies in its ability to robustly process this fingerprint information and make better predictions on the small molecules. In this section, we discuss the performance of different fingerprints and computational models. The effect of different parameters in LiCABEDS is also discussed in detail.

### 3.1 Classification Accuracy of LiCABEDS, Classification Tree, and Naive Bayes Classifier

Even if the predictions are made in the same descriptor space, the derived decision boundaries of different machine learning algorithms rarely agree, because of the discrepancy of underlying model assumptions, as well as object optimization functions. The results of systematic comparisons among classification tree, Naive Bayes classifier and LiCABEDS are plotted in Figure 2. A summary of the results can also be found in Table 2, which reports the distribution of prediction accuracy out of ten rounds of calculation on human 5-$HT_{1A}$ ligands (more details can be found in Table S1 in Supporting Information Part 3).

As shown in Figure 2 and Table 2, LiCABEDS uniformly outperforms both classification tree and Naive Bayes classifier regardless of the choice of molecular fingerprints. First, LiCABEDS exhibits the highest average prediction accuracy on ten different testing compound datasets. When Unity and FP2 fingerprints are used as the descriptor, the highest number of mistakes made by LiCABEDS on testing sets is still lower than the lowest of the Tree or Naive Bayes classifier methods. With the Molprint 2D fingerprint as descriptor, the lowest accuracy from LiCABEDS is 0.894, which is almost the same as the highest accuracy from Naive Bayes classifier, 0.896 (shown and highlighted in Table S1). Not only does LiCABEDS show the highest average prediction accuracy, but also it possesses the lowest standard deviation on four kinds of fingerprints. This indicates model stability as well as model reliability. This is further seen in the standard deviation of prediction accuracy from LiCABEDS. As listed in Table 2, the standard deviation is 0.008 on the Molprint 2D fingerprint, while the standard deviation of both the Classification Tree and Naive Bayes classifier methods is 0.013. A similar pattern is also observed using the other three types of molecular fingerprints. The standard deviation from LiCABEDS ranges from 0.010 to 0.016 using the FP2, Unity and Maccs fingerprints. On the other hand, the standard deviation of both the Tree and Naive Bayes methods range from 0.018 to 0.027 using these three fingerprints. Therefore, LiCABEDS is less affected by the distribution of training compounds compared to the Tree and Naive Bayes methods.

Molprint 2D was the most predictive descriptor among the four types of fingerprints. In this study, a total of 6839 features were defined in the whole human 5-$HT_{1A}$ dataset. The length of the Unity and FP2 fingerprints were 992 and 1024, respectively. The Maccs key had the shortest bit length of 168. Although Molprint 2D encoded many structural patterns in comparison to the Unity or FP2 fingerprints, it did not significantly improve the performance of the classification tree. As the “height” of tree was limited after “tree pruning” to avoid overfitting, a limited number of features could be considered in the classification hypothesis. The LiCABEDS method, on the other hand, consisted of 10000 weighted “decision stumps” and many factors contributed to the final prediction. This might explain the reason why LiCABEDS yielded more accurate and reliable predictions than the classification tree method.

The Naive Bayes method outmatched the Tree method using Molprint 2D as a descriptor, but the Tree method outmatched the Naive Bayes method with other fingerprints. As previously mentioned in the method section, the Naive Bayes models were slightly different with different fingerprints. Molprint 2D, as an atom environment descriptor, only considered the features present, while FP2, Maccs and Unity predefined a set of substructures and modeled both the presence and absence of structural patterns. However, when the Naive Bayes classifier from Molprint 2D was applied to the other three fingerprints, the test calculation showed that the result was even worse. The Naive Bayes classifier treated each dimension in the fingerprint equally. Thus, the performance could be affected by noise from irrelevant features. The independence assumption in the Naive Bayes model was not necessarily true for molecular fingerprints, which was one of the factors impairing the estimation of the likelihood function. In addition, the training algorithm of LiCABEDS

selected the most predictive "decision stump" and assigned its weight accordingly in order to build the classifier systematically. In that sense, not all the dimensions of the fingerprint vectors contributed to the prediction equally, and predictive features and corresponding "decision stumps" were emphasized with relatively large weights, $a_m$. Without much assumption regarding the fingerprints, LiCABEDS built robust models and produced more accurate predictions than the Naive Bayes classifier.

### 3.2 Initialization Condition

As previously mentioned in the Methods section, the equal initial weight in the training algorithm considers each training compound equally, while the balanced initial weight considers two compound categories (agonist and antagonist) equally. The two different initializations in LiCABEDS were compared on the same ten sets of training and testing compounds with $M$ = 10000. Figure 6 shows the distribution of the overall accuracy of predictions from combinations of different initialization conditions and molecular fingerprints. Table 3 lists the average accuracy and standard deviation for each ligand category, as well as for the whole testing dataset. According to Figure 3 and Table 3, these two initialization conditions result in differences with respect to the overall performance, even if equal initial weight is slightly better. As displayed in Table 3, the balanced initial weight correctly predicts antagonists at a percentage of 87.9%, 86.4%, 83.0% and 79.2% with Molprint 2D, FP2, Unity and Maccs descriptors, while the equal initial weight predicts antagonists at the accuracy of 85.6%, 84.8%, 81.6% and 70.2% on the same descriptors. The opposite pattern is observed for agonist prediction, in which the equal initial weight uniformly outperforms the balanced initial weight. To explain this, LiCABEDS training algorithm with equal initial weight aims to minimize the error function by making fewer mistakes on the training datasets, while balanced weight emphasizes both ligand categories and each training sample. For example, at the initial step of training algorithm with balanced weights, the cost to make a mistake is $1/N_{-1}$ for one antagonist and $1/N_{+1}$ for one agonist. As there are 827 agonists and 446 antagonists in the training datasets, LiCABEDS may tend to avoid making mistakes on antagonists because one mistake on an antagonist costs more than the one on an agonist ($1/N_{-1} > 1/N_{+1}$). On the other hand, the equal initial weight favors the majority category, because the mistake on any training compound costs $1/N$. Although the weights for each training sample are updated in the follow-up training iteration, the initialization condition still significantly affects the model development. As a result, the balanced initial weight makes the predictions that are more accurate on antagonists. In reality, training sets are sometimes overwhelmed by one category of samples, but correct predictions are still desired for the minority group. The balanced initial weight seeks a tradeoff between the accuracy for each category and the overall performance, which makes the algorithm generally applicable to many data mining situations.

### 3.3 Training Parameter

Besides the initialization condition, another parameter crucial to the LiCABEDS training algorithm is $M$, the number of boosting iterations. To minimize overfitting, the optimal value of $M$, $M_{\text{optimal}}$ ($M_{\text{optimal}} < 10000$) can be determined through cross-validation. In this process, a fraction of the training compounds (10% of the whole training sets) was left out as a cross-validation set. $M_{\text{optimal}}$ was then compared to the default condition, a large value of $M$, $M$ = 10000, on the same training and testing datasets. Models were developed using a balanced initial weight and the four types of fingerprints. The percentages of correct predictions on the ten testing datasets are shown in Figure 4. More details regarding cross-validation and the effect of M on training and testing errors can be found in Table S2 and Figure S1 of Supporting Information Part 4. According to the distribution shown in Figure 4, models developed by $M_{\text{optimal}}$ iterations are moderately better than $M$ = 10000 on FP2, Unity and Maccs fingerprints, but not as good as $M$ = 10000 on Molprint 2D fingerprint.

Because $M$ = 10000 is much larger than the length of FP2, Unity and Maccs fingerprints, some dimensions in the fingerprint are overrepresented in the classifier. This may result in overfitting. If this is the case, running cross-validation could control the overfitting and improve the performance. However, the length of the Molprint 2D fingerprint used in this study is 6839, which is at the magnitude of $M$ = 10000. Thus, cross-validation is not essential for the Molprint 2D fingerprint.

Hypothesis testing was carried out to quantify the difference between $M_{\text{optimal}}$ and $M$= 10000. The distribution of correct predictions on the testing datasets was examined. The null hypothesis was "the models trained with and without cross-validation have the same performance", while the alternative hypothesis was "cross-validation improves the model performance". Student's t-test showed that the one-sided p-values for the four types of fingerprints (Molprint 2D, FP2, Unity and Maccs) were 0.954, 0.357, 0.290 and 0.354, respectively. This result does not significantly favor the alternative hypothesis, indicating cross-validation does not significantly influence on the prediction generated. Our data mining studies and the results presented also indirectly support the conclusion that LiCABEDS is not so susceptible to overfitting in the studied datasets, which is also supported by boosting theory[14]. Although cross-validation is not strictly required by LiCABEDS, parameter tuning may still be beneficial under certain circumstances, such as the application of LiCABEDS on FP2, Unity and Maccs fingerprints.

### 3.4 Reject Option

To assess the confidence-rated predictions, LiCABEDS uses the raw value of $A=\sum_{m}^{M} a_m y_m(x, i_m, t_m)$ to address the degree of belief for each prediction. By applying the concept of "reject option", accurate prediction is anticipated, provided a high absolute value of $A$, whereas an "unknown" label is output to prevent uncertain prediction for a low absolute value of $A$. To validate this hypothesis, predictions were made on the ten testing compound datasets with different "reject" boundaries and molecular descriptors. The difference in the average performance of different "reject option" boundaries is reported in Figure 5.

When Molprint 2D was used as the descriptor, an average accuracy of 90.1% (Table 2) was reported without using the "reject option". As shown in Figure 5A, the average proportion of predictions made on the testing datasets readily decreases when the "reject" boundary increases from 0.5 to 2. Because more "unknown" labels are output when a higher boundary value is specified, a relatively smaller fraction of predictions is made. In the meantime, the average prediction accuracy increases from the original value of 90.1% (reported in Table 2) to 91.1%, 92.1%, and 93.8% with corresponding boundaries being 0.5, 1 and 2, respectively. A similar trend can be observed with the other three types of fingerprints as well. For example, using the FP2 fingerprint, the predictions are made on 92.3% of the testing compounds when the boundary is set to 2. An accuracy of 91.1% is obtained from the "reject option", which is a noticeable improvement from the original accuracy of 87.9% (reported in Table 2). Therefore, LiCABEDS is not only able to attain better performance by making selective predictions, but is also able to estimate the classification risk for testing compounds through the absolute value of $A$, or the confidence-rated prediction. In practice, it is sometimes economical to sacrifice some testing compounds to achieve accurate predictions. The boundary value can be determined by examining the distribution of $A$ and leaving a fair prediction ratio.

### 3.5 Across-target Ligand Functionality Prediction

In addition to the ligand functionality classification, we have explored the potential of across-target ligand bioactivity prediction using LiCABEDS program. With the assumption that agonists and antagonists might share some common pharmacological features for similar receptor subtypes, the LiCABEDS model, which was developed from human 5-$HT_{1A}$ ligands on Molprint 2D fingerprint, was used to predict the ligand functionality for other human 5-HT subtype receptors, including 5-$HT_{1B}$, 5-$HT_{1D}$, and 5-$HT_{4R}$ receptors. 5-$HT_{1A}$, 5-$HT_{1B}$ and 5-$HT_{1D}$ GPCRs can be classified as serotonin receptor subtype 1 (or 5-$HT_1$) while 5-$HT_{4R}$ belongs to the family of serotonin subtype four. 5-$HT_{1B}$ receptor has the shortest evolution distance to 5-$HT_{1A}$. On the other hand, 5-$HT_{4R}$ receptor has the largest evolution distance to 5-$HT_{1A}$ of all. As sufficient number of known agonists and antagonists has been reported for these receptors, the correlation between model predictivity and target similarity can be studied in order to understand the scope of application of established models.

The performance of LiCABEDS models for each category of the 5-HT ligands, as well as entire datasets, can be seen in Table 4. The sequence similarity scores compared to human 5-$HT_{1A}$ are calculated by blastp[28] and Sybyl Biopolymer[29], respectively. The calculations based on the 5-$HT_{1A}$ model show that 85.9% of predictions are correct on 5-$HT_{1B}$ ligands, with 87.5% accuracy for agonists and 81.6% accuracy for antagonists, respectively. The data is congruent with the relative high sequence similarity between 5-$HT_{1B}$ and 5-$HT_{1A}$ (blastp score: 304; Sybyl score: 397.80). Our studies show that the model trained from 5-$HT_{1A}$ ligands is still predictive for 5-$HT_{1B}$ ligands. However, the overall accuracy for 5-$HT_{1D}$ ligands drops to 74.5%, which may be attributed to the lower sequence similarity between 5-$HT_{1D}$ and 5-$HT_{1A}$ (blastp score: 279; Sybyl score: 393.60). 5-$HT_{4R}$, which possesses the lowest sequence similarity to 5-$HT_{1A}$ (blastp score: 142; Sybyl score: 340.90), was also evaluated, and its prediction is not necessarily better than a random guess. The results are consistent with the known data that the drugs Ergotamine (agonist) and Methiothepin (antagonist) are active to 5-$HT_{1A}$, 5-$HT_{1B}$ and 5-$HT_{1D}$ receptors but not to 5-$HT_{4R}$. The results may also suggest that LiCABEDS prediction models may have potential of applying to other targets with limited known ligands, as long as the models are developed for a closely related receptor family. This concept could extend the application of LiCABEDS to the drug discovery process targeting at orphan receptors that has no known ligand reported.

### 3.6 Model Interpretation

The interpretability of the LiCABEDS model may help us understand the underlying classification mechanism and significant features regarding ligand properties. The model developed on the first set of 5-$HT_{1A}$ training compounds, which consist of randomly selected 827 agonists and 446 antagonists, is used to demonstrate this process. As presented in the method section, each “decision stump” contributes to the final prediction according to its weight, $a_m$, as described in equation (1). In the LiCABEDS model, four out of 6839 highly weighted Molprint 2D features, which are the few highly weighted ones to distinguish agonists and antagonists, are listed in Table 5. Feature 1 and 2 are favored by agonists, while feature 3 and 4 are preferred in antagonists. In order to illustrate the structural patterns of these features, Figure 6 shows a graphical atom environment according to the four features. For example, feature 1 (0;0-1-0;0-1-4;1-3-0;2-3-0;) translates to a substructure of a central sp3 carbon atom (C.3) neighbored by one sp3 carbon atom and one sp3 nitrogen atom (N.3), and surrounded by three sp3 carbon atoms (C.3) located two or three bonds away. The Molprint2D features in Table 5 are generated by Molprint2D software package, and the detailed explanation can be found in original publication[21, 23].

Figure 7 lists seven compounds selected from the testing compound dataset in order to exemplify the four features. The first three compounds (L008638, L006280, L006274) are labeled as agonists, and the other four compounds (L022154, L018611, L001620, L018612) are labeled as antagonists. It is worth pointing out that thousands of features are involved agonist/antagonist prediction, but only four highly weighted Molprint 2D features are picked up to illustrate model interpretation. Molprint2D fingerprint (feature) is a highly sparse descriptor, and the number of features that a compound possesses is equal to the number of its heavy atoms. The agonists from the testing set, which possess both feature 1 and 2 (listed in Table 5), are involved in Model Interpretation. The same analogy is applied to antagonists. Even if the agonists or antagonists do not share the same structural scaffold, certain substructures may still match in three-dimensional space. Figure 8 displays that feature 1, 2 from the three agonists are well aligned, with the central carbon atom from feature 1 labeled in grey and the central oxygen atom from feature 2 labeled in red. Similarly, Figure 9 displays the alignment of feature 3, 4 for the four antagonists. The result suggests that those features might be related to ligand functionality and ligand-protein interaction. The interpretability of LiCABEDS models is rooted in the explanation of each "decision stump", especially the highly weighted ones. Therefore, LiCABEDS models can be easily understood and interpreted, which could potentially guide chemical modification to achieve better pharmacological or physicochemical profile.

### 3.7 Model Robustness

Model robustness is the potential to handle diverse training data and provide consistent predictions. Section 3.1 has shown that LiCABEDS models render the most consistent and accurate predictions on any of the molecular fingerprints. To analyze the composition of the classifiers, important dimensions of Molprint 2D fingerprints are extracted from the LiCABEDS models, which are developed on ten different 5-$HT_{1A}$ training datasets. All the Molprint 2D features are observed totally more than 50 times in all the models and possess weights, $a_m$, larger than 0.08. To visualize the major components of the classifiers, Figure 10 shows the distribution of occurrence of six important Molprint 2D features that are favored in agonists, for which LiCABEDS training algorithm may select a feature several times to minimize generalization error. The occurrence of features mainly ranges from 4 to 7, except dimension 2828. Even if the ten models are developed on the randomly selected training datasets, only three outliers (labeled as circles in Figure 10, two in dimension 1694, one in dimension 2828) are identified in the boxplot. Thus, the occurrence of the six major components has moderate variance in each of the ten models. The stability of the influential "weak-learners" leads to consistent prediction accuracy, although only a few features are visualized.

## 4. CONCLUSION

We have reported a novel ligand classification algorithm, Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS), and thoroughly investigated it through the case studies of ligand functionality prediction for the GPCR 5-HT subtypes. The performance of LiCABEDS is compared to the Classification Tree model and Naive Bayes classifier using four types of molecular fingerprints: Molprint 2D, FP2, Unity and Maccs. Our results show that LiCABEDS uniformly produces the most accurate and consistent predictions, especially with Molprint 2D fingerprints as the descriptor. Additionally, unique characteristics of LiCABEDS make it applicable to model various ligand properties. The flexible initialization conditions of LiCABEDS allow the development of predictive models and emphasize minority categories on unbalanced training datasets. Parameterization is usually a complicated procedure in many machine-learning algorithms, however, model development in LiCABEDS is simplified because the number of boosting iterations, *M*, is

the only parameter required for model training. The result from cross-validation suggests that a large value of *M* still yields satisfactory performance, which makes the model training process simplified in practice. Another valuable characteristic of LiCABEDS is the "reject option", which returns the degree of confidence for each prediction. Higher prediction accuracy can be achieved by rejecting some "low-confident" testing samples. The capability of LiCABEDS is further demonstrated through the application on a cross-target prediction. The interpretation of LiCABEDS models may reveal the correlation between structural pattern and molecular properties of interest. The robustness of LiCABEDS models is further demonstrated by examining the principal components of "decision stumps". Lastly, LiCABEDS has been implemented into an easy-to-use and freely available (http://www.CBLigand.org/LiCABEDS/) software platform that provides a graphical user interface for automating model development and predictions. As a general classifier, LiCABEDS may also have great potential for modeling and predicting other ligand properties, such as ADME prediction and other applications on in-silico drug design research. These are ongoing projects and will be reported in future studies.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This project is supported by grants from NIH R01 DA025612 and NIGMS P50-GM067082.

## References

1. (a) Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug Discovery Today. 2002; 7(17):903. [PubMed: 12546933] (b) Wilton DJ, Harrison RF, Willett P, Delaney J, Lawson K, Mullier G. Virtual Screening Using Binary Kernel Discrimination:Analysis of Pesticide Data. Journal of Chemical Information and Modeling. 2006; 46(2):471. [PubMed: 16562974] (c) X-QX. Exploiting PubChem for Virtual Screening. Expert Opinion for Drug Discovery. 2010; 5:1–16. (in press).
2. Clark DE, Pickett SD. Computational methods for the prediction of "drug-likeness". Drug Discovery Today. 2000; 5(2):49. [PubMed: 10652455]
3. Papa E, Pilutti P, Gramatica P. Prediction of PAH mutagenicity in human cells by QSAR classification. SAR and QSAR in Environmental Research. 2008; 19(1):115–127. [PubMed: 18311639]
4. Dudek AZ, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen. 2006; 9(3):213–28. [PubMed: 16533155]
5. (a) Grover II, Singh II, Bakshi II. Quantitative structure-property relationships in pharmaceutical research - Part 2. Pharm Sci Technolo Today. 2000; 3(2):50–57. [PubMed: 10664573] (b) Grover M, Singh B, Bakshi M, Singh S. Quantitative structure-property relationships in pharmaceutical research - Part 1. Pharmaceutical Science & Technology Today. 2000; 3(1):28. [PubMed: 10637598] (c) Wassermann AM, Geppert H, Bajorath J. Searching for Target-Selective Compounds Using Different Combinations of Multiclass Support Vector Machine Ranking Methods, Kernel Functions, and Fingerprint Descriptors. Journal of Chemical Information and Modeling. 2009; 49(3):582. [PubMed: 19249858]
6. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988; 110(18):5959.
7. Liu R, Matheson LE. Comparative molecular field analysis combined with physicochemical parameters for prediction of polydimethylsiloxane membrane flux in isopropanol. Pharm Res. 1994; 11(2):257–66. [PubMed: 8165185]

8. Van der Graaf PHN, Van Schaick J, Danhof EAM. Multivariate quantitative structure-pharmacokinetic relationships (QSPKR) analysis of adenosine A1 receptor agonists in rat. J Pharm Sci. 1999; 88(3):306–12. [PubMed: 10052988]
9. Chen JZ, Han XW, Liu Q, Makriyannis A, Wang J, Xie XQ. 3D-QSAR studies of arylpyrazole antagonists of cannabinoid receptor subtypes CB1 and CB2. A combined NMR and CoMFA approach. J Med Chem. 2006; 49(2):625–36. [PubMed: 16420048]
10. Agarwal A, Taylor EW. 3-D QSAR for intrinsic activity of 5-HT1A receptor ligands by the method of comparative molecular field analysis. Journal of Computational Chemistry. 1993; 14(2):237–245.
11. Stumpfe D, Geppert H, Bajorath J. Methods for Computer-Aided Chemical Biology. Part 3: Analysis of Structure Selectivity Relationships through Single- or Dual-Step Selectivity Searching and Bayesian Classification. Chemical Biology & Drug Design. 2008; 71(6):518–528. [PubMed: 18482335]
12. Martin YC. 3D QSAR: Current State, Scope, and Limitations. 3D QSAR in Drug Design. 2002:3.
13. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. journal of computer and system sciences. 1997; 55:119–139.
14. Freund Y, Schapire RE. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence. 1999; 14(5):771.
15. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and Regression Trees. Wadsworth; Monterey: 1984.
16. Mitchell, TM. Machine Learning. New York: McGraw-Hill; 1997.
17. Bishop, CM. Pattern Recognition and Machine Learning. Springer; 2006.
18. Plewczynski D, Spieser SA, Koch U. Assessing different classification methods for virtual screening. J Chem Inf Model. 2006; 46(3):1098–106. [PubMed: 16711730]
19. Walters WP, Murcko MA. Prediction of ‘drug-likeness’. Advanced Drug Delivery Reviews. 2002; 54(3):255. [PubMed: 11922947]
20. Deconinck E, Zhang MH, Coomans D, Vander Heyden Y. Classification tree models for the prediction of blood-brain barrier passage of drugs. J Chem Inf Model. 2006; 46(3):1410–9. [PubMed: 16711761]
21. Bender A, Mussa HY, Glen RC, Reiling S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. Journal of Chemical Information and Computer Sciences. 2004; 44(5):1708. [PubMed: 15446830]
22. (a) Okuno Y, Yang J, Taneishi K, Yabuuchi H, Tsujimoto G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. Nucleic Acids Res. 2006; 34(Database issue):D673–7. [PubMed: 16381956] (b) Okuno Y, Tamon A, Yabuuchi H, Niijima S, Minowa Y, Tonomura K, Kunimoto R, Feng C. GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update. Nucleic Acids Res. 2008; 36(Database issue):D907–12. [PubMed: 17986454]
23. Bender A, Mussa HY, Glen RC, Reiling S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. Journal of Chemical Information and Computer Sciences. 2003; 44(1):170. [PubMed: 14741025]
24. MACCS Structural keys. Symyx Software; San Ramon, CA: 2005.
25. Hutchison, G. Open Babel: File Translation for Computational Chemistry and Nanoscience. NNIN/CNF Fall Workshop; 2005.
26. (a) Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. Journal of Chemical Information and Computer Sciences. 2003; 43(2):493. [PubMed: 12653513] (b) Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr Pharm Des. 2006; 12(17):2111–20. [PubMed: 16796559]
27. Ripley, BD. Pattern Recognition and Neural Networks. Vol. Chapter 7. Cambridge University Press; Cambridge: 1996.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

29. Tripos, Sybyl Biopolymer. Sybyl version 7.1. Tripos, Inc; 2006. http://tripos.com/index.php?family=modules,SimplePage,sybyl_biopolymer

**Figure 1.**
(a) Illustration of a “decision stump for ligand function prediction, based on molecular fingerprint. In this model, agonists are labeled as +1 and antagonists are labeled as −1. (b) Illustration of the composition of LiCABEDS.

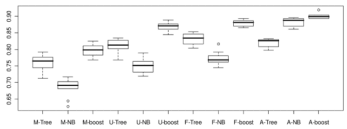


**Figure 2.**

The box-plot showing the distribution of prediction accuracy from ten rounds of calculation. M: Maccs key; U: Unity fingerprint; F: FP2 fingerprint; A: Molprint2D fingerprint. Tree stands for classification tree, NB stands for Naive Bayes classifier and boost is short for LiCABEDS.

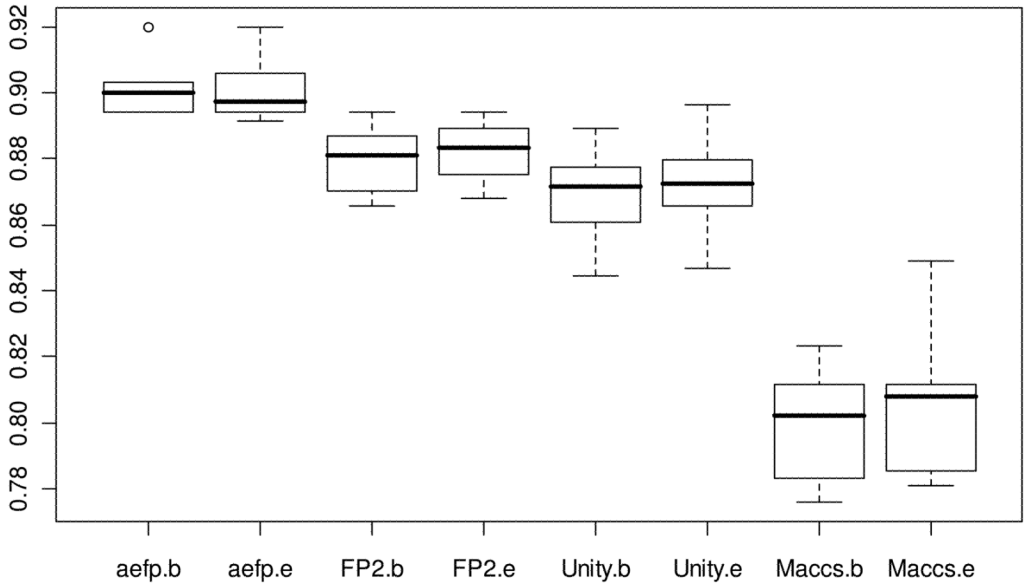


**Figure 3.**
Boxplot showing prediction accuracy with different initialization conditions. **aefp** stands for Molprint2D, ***b*** stands for balanced initial weight and ***e*** stands for equal initial weight.

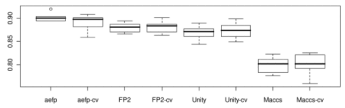


**Figure 4.**

The boxplot showing the effect of *M*, number of boosting rounds. The x-axis denotes the choice of fingerprint and *M*. aefp stands for Molprint2D fingerprint. cv means that number of boosting rounds is set to $M_{\text{optimal}}$. The label without cv means $M$=10000 by default.

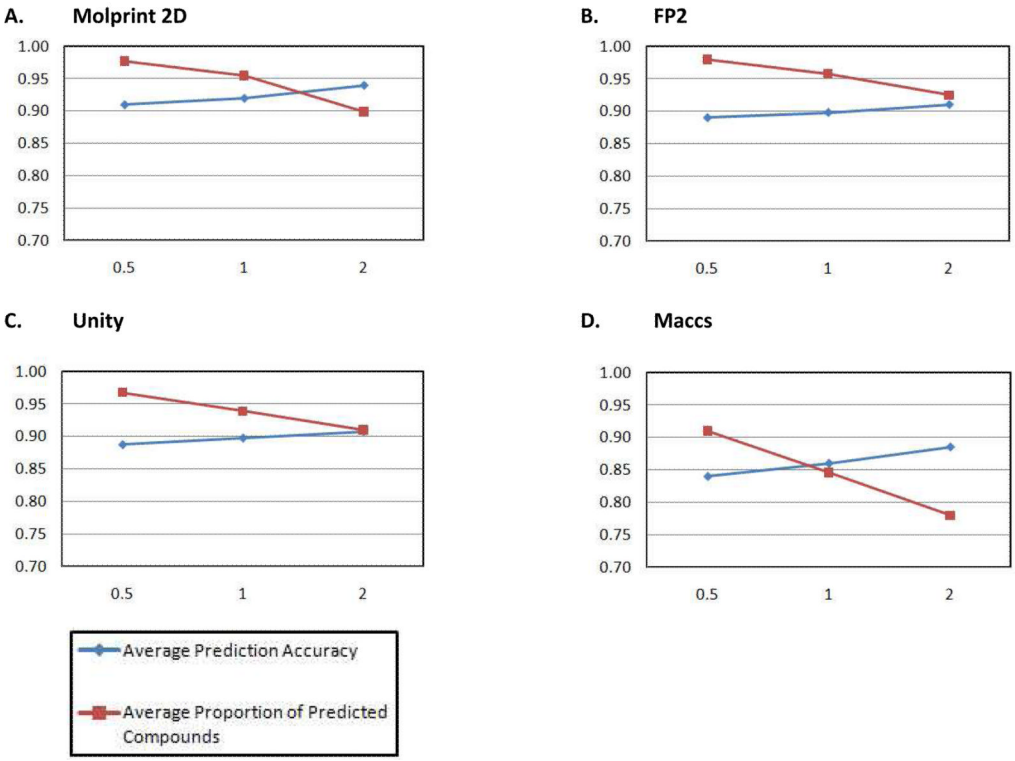


**Figure 5.**
The plot showing average prediction accuracy and percentage of prediction made on testing compounds. Each figure corresponds to a type of fingerprint. X-axis denotes the value of decision boundary.

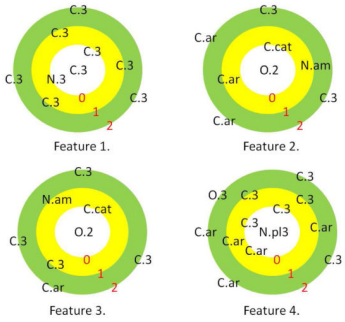


**Figure 6.**

Four sample Molprint 2D features in graphic representation. Each feature depicts a central atom and its atom environment up to a specific topological distance. The atom environment in Molprint2D is defined as the quantity of heavy atoms surrounding the central atom. Heavy atoms are distinguished by Sybyl atom types.

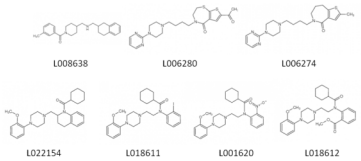


**Figure 7.**
Compounds used to exemplify four features in Figure 9. L008638, L006280 and L006274 are labeled as agonists, and the other four are antagonists.

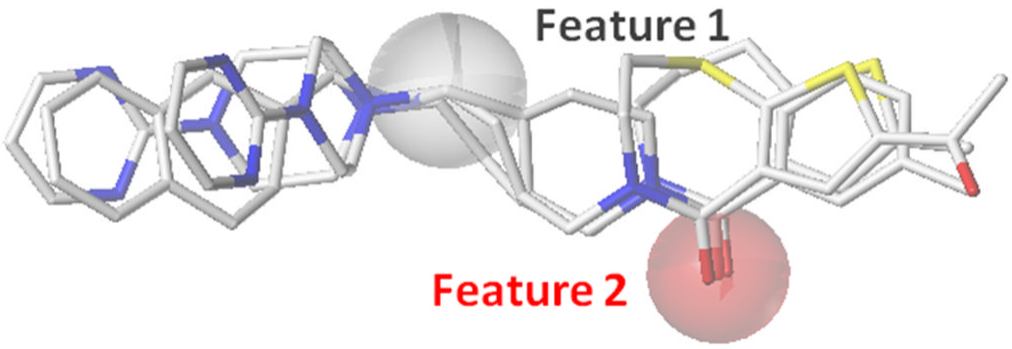


**Figure 8.**
3-D alignment of three agonists.

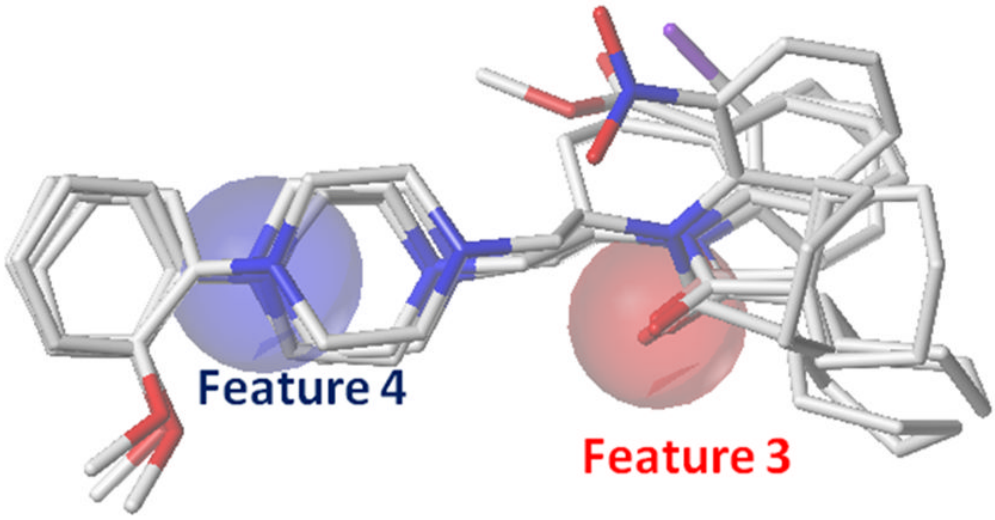


**Figure 9.**
3-D alignment of four antagonists.

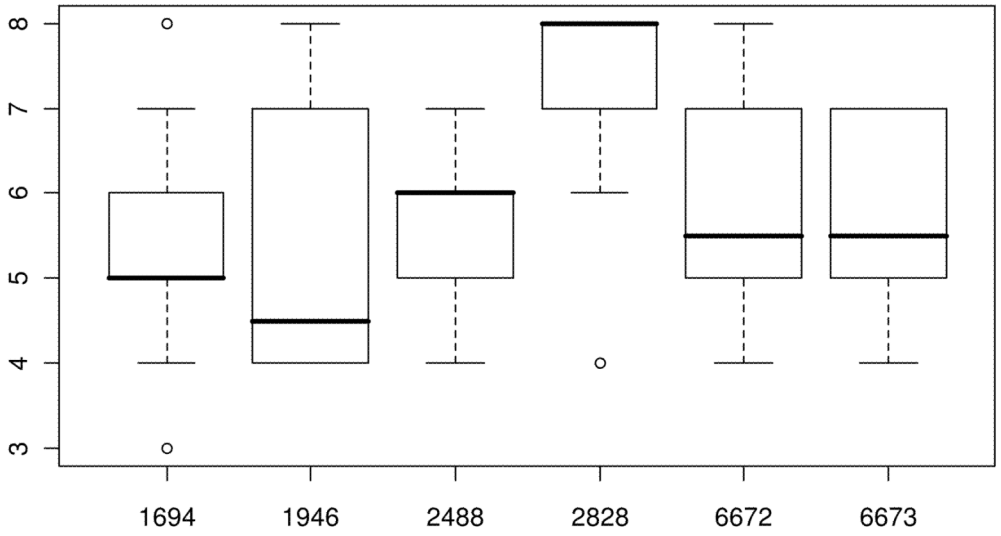


**Figure 10.**
The boxplot showing the occurrence of the six major LiCABEDS components in ten 5-$HT_{1A}$ models

**Table 1**

The quantity and molecular properties of agonists and antagonists. Molecular properties are given by sample average and standard deviation.

| Receptor | Ligand Category | Quantity | Molecular Weight | No. of Rotatable Bond | No. of H-Bond Acceptor | No. of H-Bond Donor |
|---|---|---|---|---|---|---|
| HUMAN | Agonist | 1102 | 346.7 ± 72.6 | 5.4 ± 2.4 | 2.8 ± 1.5 | 1.3 ± 0.9 |
| 5-$HT_{1A}$ | Antagonist | 595 | 385.8 ± 70.7 | 5.0 ± 2.5 | 3.2 ± 1.3 | 1.7 ± 0.9 |
| HUMAN | Agonist | 104 | 348.4 ± 104.1 | 4.3 ± 2.5 | 3.6 ± 1.8 | 1.2 ± 0.8 |
| 5-$HT_{1B}$ | Antagonist | 38 | 359.6 ± 88.4 | 4.1 ± 1.8 | 2.6 ± 1.4 | 1.2 ± 0.7 |
| HUMAN | Agonist | 685 | 339.1 ± 82.3 | 5.8 ± 2.6 | 3.3 ± 1.6 | 1.2 ± 0.7 |
| 5-$HT_{1D}$ | Antagonist | 335 | 420.2 ± 71.0 | 4.1 ± 2.2 | 3.9 ± 1.4 | 1.7 ± 0.9 |
| HUMAN | Agonist | 287 | 369.4 ± 65 | 3.9 ± 2.5 | 2.8 ± 1.1 | 1.1 ± 0.5 |
| 5-$HT_{4R}$ | Antagonist | 262 | 367.3 ± 61.5 | 5.7 ± 2.1 | 3.6 ± 1.4 | 1.1 ± 0.4 |

## Table 2

The sample mean and standard deviation of prediction accuracy from different computational models and molecular fingerprints.

| Model Fingerprint | Tree | Naive Bayes | LiCABEDS |
|---|---|---|---|
| Maccs | 0.759 ± 0.026 | 0.685 ± 0.027 | 0.799 ± 0.016 |
| Unity | 0.810 ± 0.023 | 0.753 ± 0.023 | 0.869 ± 0.013 |
| FP2 | 0.831 ± 0.018 | 0.771 ± 0.021 | 0.879 ± 0.010 |
| Molprint2D | 0.820 ± 0.013 | 0.883 ± 0.013 | 0.901 ± 0.008 |

## Table 3

The sample mean and standard deviation of prediction accuracy for each category of ligands, using equal initial weight and balanced initial weight.

| | | Agonist | Antagonist | Overall |
|---|---|---|---|---|
| Molprint 2D | Balanced weight | 0.913 ± 0.018 | 0.879 ± 0.041 | 0.901 ± 0.008 |
| | Equal weight | 0.924 ± 0.018 | 0.856 ± 0.038 | 0.900 ± 0.009 |
| FP2 | Balanced weight | 0.888 ± 0.021 | 0.864 ± 0.026 | 0.879 ± 0.010 |
| | Equal weight | 0.901 ± 0.021 | 0.848 ± 0.029 | 0.882 ± 0.009 |
| Unity | Balanced weight | 0.891 ± 0.018 | 0.830 ± 0.023 | 0.869 ± 0.013 |
| | Equal weight | 0.901 ± 0.020 | 0.816 ± 0.022 | 0.871 ± 0.014 |
| Maccs | Balanced weight | 0.803 ± 0.025 | 0.792 ± 0.033 | 0.799 ± 0.016 |
| | Equal weight | 0.860 ± 0.018 | 0.702 ± 0.049 | 0.805 ± 0.020 |

**Table 4**

The prediction accuracy for across-target ligand functionality prediction for 5-HT receptor subtypes based on 5-$HT_{1a}$ LiCABEDS model

| Receptor | Accuracy for agonists | Accuracy for antagonists | Overall accuracy | Blastp similarity score[a] | Sybyl similarity score[b] |
|---|---|---|---|---|---|
| Human 5-$HT_{1B}$ | 0.875 | 0.816 | 0.859 | 304 | 397.80 |
| Human 5-$HT_{1D}$ | 0.812 | 0.609 | 0.745 | 279 | 393.60 |
| Human 5-$HT_{4R}$ | 0.826 | 0.267 | 0.559 | 142 | 340.90 |

[a] Blastp similarity score is calculated by blastp using default scoring parameters

[b] Sybyl similarity score is calculated by Sybyl Biopolymer[29] using “pmutation” scoring matrix.

## Table 5

List of important features, indices in the bit vector and weights in the sample LiCABEDS model. The interpretation of the listed features are presented graphically in Figure 6.

| | Bit Index | Molprint 2D feature[a] | Weight |
|---|---|---|---|
| Feature 1 | 3474 | 0;0-1-0;0-1-4;1-3-0;2-3-0; | 0.869 |
| Feature 2 | 808 | 8;0-1-1;1-1-2;1-1-27;2-2-0;2-2-2; | 0.328 |
| Feature 3 | 362 | 8;0-1-1;1-1-0;1-1-27;2-3-0;2-1-2; | 0.343 |
| Feature 4 | 4322 | 18;0-2-0;0-1-2;1-2-0;1-2-2;2-2-2;2-2-4;2-1-7; | 0.639 |

[a]Features listed in this table are extracted from Molprint 2D software package.