

THE QSAROME OF THE RECEPTOROME: QUANTITATIVE STRUCTURE-ACTIVITY  
RELATIONSHIP MODELING OF MULTIPLE LIGAND SETS ACTING AT MULTIPLE  
RECEPTORS

Guiyu Zhao

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in  
partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Division  
of Chemical Biology and Medicinal Chemistry at Eshelman School of Pharmacy

Chapel Hill  
2011

Approved by:

Dr. Alexander Tropsha

Dr. Bryan Roth

Dr. Steve Marron

Dr. Qisheng Zhang

Dr. Shawn Gomez

## ABSTRACT

GUIYU ZHAO: The QSARome of the Receptorome: Quantitative Structure-Activity Relationship Modeling of Multiple Ligand Sets Acting at Multiple Receptors  
(Under the direction of Alexander Tropsha)

Recent advances in High Throughput Screening (HTS) led to the rapid growth of chemical libraries of small molecules, which calls for improved computational tools and predictive models for Virtual Screening (VS). Thus this dissertation focuses on both the development and application of predictive Quantitative Structure-Activity Relationship (QSAR) models and aims to discover novel therapeutic agents for certain diseases.

First, this dissertation adopts the combinatorial QSAR framework created by our lab, including the first application of the Distance Weighted Discrimination (DWD) method that resulted in a set of robust QSAR models for the 5-HT<sub>7</sub> receptor. VS using these models, followed by the experimental test of identified compounds, led to the finding of five known drugs as potent 5-HT<sub>7</sub> binders. Eventually, droperidol ( $K_i = 3.5 \text{ nM}$ ) and perospirone ( $K_i = 8.6 \text{ nM}$ ) proved to be strong 5-HT<sub>7</sub> antagonists. Second, we intended to enhance VS hit rate. To that end, we developed a cost/benefit ratio as an evaluation performance metric for QSAR models. This metric was applied in the Decision Tree machine learning method in two ways: (1) as a benchmarking criterion to compare the prediction performances of different classifiers and (2) as a target function to build QSAR classification trees. This metric may be more suitable for imbalanced HTS data that include few active but many inactive compounds.

Finally, a novel QSAR strategy was developed in response to the polygenic nature of most psychotic disorders, related mainly to G-Protein-Coupled Receptors (GPCRs), one class of molecular targets of greatest interest to the pharmaceutical industry. We curated binding data for thousands of GPCR ligands, and developed predictive QSAR models to assess the GPCR binding profiles of untested compounds that could be used to identify potential drug candidates. This comprehensive study yielded a compendium of validated QSAR predictors (the GPCR QSARome), providing effective *in silico* tools to search for novel antipsychotic drugs.

The advances in results and procedures achieved in these studies will be integrated into the current computational strategies for rational drug design and discovery boosted by our lab, so that predictive QSAR modeling will become a reliable support tool for drug discovery programs.

## ACKNOWLEDGEMENT

I give my truehearted thanks to my mentor, Dr. Alexander Tropsha, for his scientific guidance and education throughout my graduate study and research. His insight, courage, and determination all inspired me to dedicate myself to hard work and the courage to inquire. I thank him for nurturing me from a naïve computational scientist. I received great inheritance from Alex's lab and hope to pass it on in my future life.

I am very grateful to Dr. Denis Fourches, Dr. Eugene Muratov, Dr. Clark Jeffries, and Dr. Xiang (Simon) Wang, for their invaluable help and scientific inspirations. Without their professional skills and close mentorship of my projects, my research would never have been completed.

I also want to thank my other committee members Dr. Bryan Roth, Dr. Steve Marron, Dr. Qisheng Zhang, and Dr. Shawn Gomez. They provided me with wonderful ideas for my projects. Special thanks are given to Dr. Bryan Roth for providing the fruitful PDSP database and insightful interpretation of our computational results.

I give thanks to other members and former members of the molecular modeling lab for their friendship and support, especially Dr. Alexander Golbraikh, Dr. Alexander Sedykh, Yen Low, Dr. Jui-Hua Hsieh, Dr. Hao Tang, Dr. Liying Zhang, and Dr. Hao Zhu, for their daily ardent help with my research projects.

# TABLE OF CONTENTS

LIST OF TABLES.....	IX
LIST OF FIGURES.....	X
LIST OF ABBREVIATIONS.....	XII
CHAPTER 1. INTRODUCTION.....	1
1.1. Overview.....	1
1.2. Quantitative Structure-Activity Relationship (QSAR).....	3
1.3. Validation Criteria for Virtual Screening.....	5
1.4. Thesis Outline.....	8
CHAPTER 2. APPLICATION OF CURRENT CHEMINFORMATIC TECHNIQUES TO HUMAN 5-HT <sub>7</sub> DATASETS TO BUILD VALIDATED AND PREDICTIVE QSAR MODELS FOR DRUG REPURPOSING.....	11
2.1. Introduction.....	11
2.2. Materials and Methods.....	13
2.2.1. Data.....	13
2.2.2. Generation of MolConnZ Descriptors.....	14
2.2.3. Applicability Domain (AD).....	14
2.2.4. Structural Outliers Exclusion.....	15
2.2.5. Modeling and External Evaluation Sets.....	16
2.2.6. DWD Classification Method.....	16
2.2.7. Sphere Exclusion Algorithm.....	17
2.2.8. kNN Regression Modeling Method.....	18

2.2.9. Y-Randomization Test .....	20
2.2.10. Consensus Prediction and Virtual Screening (VS) using kNN Regression Models .....	21
2.2.11. Experimental Validation of Screening Hits .....	22
2.3. Results and Discussion .....	22
2.3.1. DWD Classification Modeling .....	22
2.3.2. kNN QSAR Regression Modeling.....	23
2.3.3. Models Validation using External Evaluation Set .....	24
2.3.4. QSAR-based Virtual Screening.....	25
2.3.5. Experimental Validation .....	25
2.3.6. Data quality and QSAR modeling.....	27
2.4. Conclusions .....	27
<b>CHAPTER 3. DEVELOPMENT OF ALGORITHM ECONOMIC RATIO (ER) AS BOTH A COST FUNCTION AND A VALIDATION MERIT FOR CLASSIFICATION QSAR MODELS.....</b>	<b>42</b>
3.1. Introduction .....	42
3.2. Materials.....	43
3.2.1. Datasets .....	44
3.2.2. Descriptors.....	45
3.3. Theoretical basis.....	47
3.3.1. Evaluation of QSAR prediction performance.....	47
3.3.2. Evaluating QSAR classifiers using cost/benefit ratio .....	48
3.4. Expanding ER as a Target Function.....	49
3.4.1. Decision tree construction.....	50
3.4.2. Tree construction algorithm .....	50
3.4.3. The advantage of application of ER as a target function.....	53
3.5. Results and Discussion .....	54

3.5.1. Model construction .....	54
3.5.2. 5-fold external cross-validation .....	54
3.5.3. Chemical similarity analysis .....	56
3.5.4. Discussion .....	56
3.6. Conclusions .....	57
CHAPTER 4. PILOT STUDY FOR THE QSAROME PROJECT: 5-HT <sub>1A</sub> .....	66
4.1. Introduction .....	66
4.2. Materials and Methods .....	68
4.2.1. Dataset .....	68
4.2.2. Chemical Data Curation .....	69
4.2.3. Biological Data Curation .....	72
4.2.4. Generation of Descriptors .....	73
4.2.5. kNN Modeling Algorithm .....	74
4.2.6. Support Vector Machines (SVM) Modeling Method .....	74
4.2.7. Five-fold External Cross-Validation .....	76
4.2.8. Virtual Screening (VS) .....	77
4.3. Results and Discussion .....	77
4.3.1. Curated Datasets .....	77
4.3.2. Activity Analysis .....	78
4.3.3. QSAR Modeling .....	79
4.3.4. Virtual Screening (VS) .....	81
4.4. Conclusions .....	81
CHAPTER 5. QSAROME OF THE RECEPTOROME: QSAR MODELING OF MULTIPLE LIGAND SETS ACTING AT MULTIPLE RECEPTORS .....	97
5.1. Introduction .....	97
5.2. Datasets and Methods .....	99

5.2.1. Datasets .....	99
5.2.2. Descriptor Generation .....	100
5.2.3. Building Models with Support Vector Machines (SVM) .....	101
5.2.4. Applicability Domain (AD).....	101
5.2.5. Data Division for Model Building and Validation.....	102
5.2.6. Y-Randomization test.....	103
5.2.7. Gap Filling and Virtual Screening (VS) .....	104
5.3. Results and Discussion .....	104
5.3.1. Curated Data Matrix .....	104
5.3.2. Validation of SVM Models .....	105
5.3.3. Prediction Performance for the External Matrix .....	106
5.3.4. Filling the Gaps in the Matrix.....	106
5.3.5. Experimental Testing .....	107
5.4. Conclusions .....	108
CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS .....	118
6.1. Application of current cheminformatic techniques to human 5-HT <sub>7</sub> datasets to build validated and predictive QSAR classification models for drug repurposing. ....	118
6.2. Development of algorithm Economic Ratio (ER) as both a cost function and a validation merit for classification QSAR. ....	119
6.3. QSARome of the Receptorome: QSAR Modeling of Multiple Ligand Sets Acting at Multiple Receptors .....	120
Appendices .....	123
References .....	135



## LIST OF TABLES

Table 2.1. Frequently Used R Values and the Corresponding Critical Values of Zc for One-Tail Test.....	38
Table 2.2. Confusion matrices for DWD and SVM predicted on 18 external compounds ....	39
Table 2.3. Consensus prediction of external test sets by kNN Models passed acceptance criteria based on PDSP and WOMBAT datasets. ....	39
Table 2.4. Results of the experimental assays and drug information of seven hit compounds.....	40
Table 2.5. Examples of activity variation for the 5-HT <sub>7</sub> binders extracted from public PDSP K <sub>i</sub> database. ....	41
Table 3.1. Examples of prior probability tables of two hypothetical QSAR classifiers called Q1 and Q2.....	63
Table 3.2. Conventional performance measurements of the two QSAR models derived from the prior probabilities.....	63
Table 3.3. Datasets used for decision tree model construction and evaluation. ....	64
Table 3.4. Performance statistics of 5-fold external cross-validation.....	65
Table 4.1. 5-HT <sub>1A</sub> dataset.....	89
Table 4.2. Four different datasets based on the resources and tested organisms.....	89
Table 4.3. Numbers of compounds for all continuous and classification modeling sets.....	90
Table 4.4. Prediction performance ( $R^2$ ) of developed WinSVM continuous models. ....	91
Table 4.5. Prediction performance ( $R^2$ ) of developed kNN continuous models. ....	92
Table 4.6. Prediction performance (CCR) of developed WinSVM classification models. ....	93
Table 4.7. Prediction performance (CCR) of developed kNN classification models. ....	94
Table 4.8. Prediction performance of consensus WinSVM classification models. ....	95
Table 4.9. Prediction accuracies (CCR) of cross prediction between Datasets I-IV. ....	96
Table 5.1. Examples of chemical structure processing during basic data curation. ....	116
Table 5.2. Prazosin – an example of excluded compound with multiple records and high activity deviation in both ChEMBL and PDSP.....	117

## LIST OF FIGURES

Figure 1.1. R&D model yielding costs to successfully discover and develop a single new molecular entity (NME).....	10
Figure 1.2. The confusion matrix used to evaluate a classifier.....	10
Figure 2.1. The workflow of QSAR model building, validation and virtual screening applied to the 5-HT <sub>7</sub> dataset and WDI database.....	30
Figure 2.2. Distribution of pK <sub>i</sub> values of each dataset extracted from PDSP and WOMBAT.....	31
Figure 2.3. The q <sup>2</sup> and R <sup>2</sup> distribution for 5-HT <sub>7</sub> kNN QSAR models built with actual activity data versus models generated with randomized data (Y-randomization).....	32
Figure 2.4. Comparison of actual and predicted pK <sub>i</sub> for the external evaluation set using the best kNN models.....	33
Figure 2.5. Functional assay of hit compounds against the 5-HT <sub>7</sub> receptor.....	34
Figure 2.6. Toy Example, illustrating potential for “data piling” problem in HDLSS settings, for discrimination using SVM.....	35
Figure 2.7. DWD and SVM classification models based on the 5- HT <sub>7</sub> dataset.....	36
Figure 2.8. The 5-HT <sub>7</sub> dataset promiscuous binding matrix on 11 receptors including 5-HT <sub>7</sub> .....	37
Figure 3.1. Decision tree for PGP dataset grown to minimize ER <sub>c</sub> .....	58
Figure 3.2. Decision tree for PGP dataset optimized to maximize WER <sub>c</sub> .....	59
Figure 3.3. Main steps of ER-based Decision Tree Algorithm.....	60
Figure 3.4. Decision tree for PGP dataset optimized to maximize weighted CCR <sub>c</sub> .....	61
Figure 3.5. Distribution of Tanimoto coefficients (T <sub>c</sub> ) for all pairs of compounds in PGP modeling set and virtual screening hits.....	62
Figure 4.1. General dataset curation workflow and number of compounds kept after each step.....	83
Figure 4.2. The nitro group which can be represented by five different patterns.....	84
Figure 4.3. Distribution of activity variation for the duplicate cases.....	85
Figure 4.4. Overlaps between the datasets.....	86

Figure 4.5. Correlation of $pK_i$ values for the compounds from Datasets I-IV.....	87
Figure 4.6. Types of descriptors, modeling methods, and variable properties of QSAR models based upon Datasets I-IV.....	88
Figure 5.1. An example of wrong translation from literature to ChEMBL.....	110
Figure 5.2. Data distribution and $pK_i$ cutoff values for each dataset.....	111
Figure 5.3. Prediction performance of SVM models. CCR is the cumulative CCR values of 5 external folds.....	111
Figure 5.4. Prediction performance for the external matrix composed of 13 drugs.....	112
Figure 5.5. Number of binders (either experimental or predicted) to each GPCR target.....	113
Figure 5.6. Distribution of compounds targeting various numbers of GPCRs.....	114
Figure 5.7. The heat map of final curated matrix (part).....	115

## LIST OF ABBREVIATIONS

AD	Applicability Domain
ADME	Absorption, distribution, metabolism, and excretion
CADD	Computer-aided Drug Design
CCR	Correct Classification Rate
CV	Cross Validation
GPCR	G Protein-Coupled Receptor
HTS	High Throughput Screening
HTS	High Throughput Screening
kNN	k-Nearest Neighbors
LBVS	Ligand-based Virtual Screening
LOO	Leave One Out
MLR	Multiple Linear Regression
PLS	Partial Linear Squares
QSAR	Quantitative Structure Activity Relationship
RF	Random Forest
RMSD	Root Mean Square Deviation
ROC	Receiver Operating Characteristic
SBVS	Structure-based Virtual Screening
SE	Sensitivity
SP	Specificity
SVM	Support Vector Machines
TC	Tanimoto Coefficient
VS	Virtual Screening

# CHAPTER 1

## INTRODUCTION

### 1.1. Overview

The drug discovery and development pipeline is a notoriously time-consuming and costly process. To successfully launch one New Chemical Entity (NCE) from the discovery stage to market takes about 15 years and hundreds of millions (Figure 1.1)[1]. Applying Computer-Aided Drug Design (CADD) strategies could provide both time- and cost-savings for drug research and development programs (i.e., integration of computational tools into the standardized pipeline should further raise the efficiency of drug design).

As an integral part of CADD, Quantitative Structure-Activity Relationships (QSAR) is experiencing one of the most important periods in its history, highlighted by the availability of vast chemical databases with abundant bioactivity data, such as ChEMBL[2], PDSP[3], and dozens of others[4]. The explosive growth of such data provides a good opportunity for large-scale QSAR modeling across diverse pharmaceutically relevant targets. Resulting QSAR models could become valuable tools for identifying novel molecular probes and potential leads for drug discovery.

To develop statistically robust QSAR models, our lab built a rigorous workflow for development and validation of QSAR models. Major stages of this workflow include the division of the original datasets into training, test, and external validation sets; Y-

randomization validation; model selection based on given statistical performance; and Virtual Screening (VS)[5–8]. The underlying components of this workflow, such as data curation, development of ensemble models, hit rate of VS, or Applicability Domain (AD)[9][9], are all active research areas and further improvement in overall model predictivity can be expected through their advancements.

This dissertation focuses on the target class of G Protein-coupled Receptors (GPCRs), a group of molecular targets of great interest to pharmaceutical industry[10]. As of 2003, the number of GPCRs in human genome from five main families (glutamate, rhodopsin, adhesion, frizzled/taste2, and secretin) had been estimated at over 800[11]. However, the true number is much higher now due to the known existence of alternatively spliced variants and editing isoforms of GPCRs. In addition, GPCRs with unknown functions (i.e., lack of known natural transmitters), called “orphan” GPCRs, account for a large portion of newly identified GPCRs[12].

The impact of GPCRs on drug discovery is phenomenal. Previous studies suggest that at least one-third[13], and perhaps up to half[14] of currently marketed drugs target GPCR family members, which represent only around 3% of known molecular targets[15]. Actively ongoing studies of GPCRs such as deorphanization of orphan GPCRs provide huge opportunities for new drug discovery. For instance, the majority of drug targets related to central nervous system (CNS) disorders (e.g., depression, schizophrenia and bipolar disorder) belong to this receptor family. However, most antipsychotic drugs have complex GPCR polypharmacology, leading either to therapeutic effects or undesired adverse events. Thus, it will be beneficial to understand functioning bioprofiles of GPCR ligands to enhance their potential therapeutic effects and avoid possible adverse reactions. Our goal of focusing on

this receptor class is to search for antipsychotic drugs, both selective to a specific GPCR and those that non-selectively target a combination of critical GPCRs.

## 1.2. Quantitative Structure-Activity Relationship (QSAR)

Previous studies (e.g., SAR analysis) have shown that structural features of small molecules have significant effect on their physicochemical and biological properties. Compared with conventional SAR analysis, the QSAR analysis intends to *quantitatively* explain the relationship between chemical structures and the corresponding activity. The QSAR analysis is based on the assumption that compounds with similar structures are expected to exhibit similar properties (the Similarity Property Principle[16]). This assumption serves as a foundation behind experimental SAR studies by medicinal chemists, as well as the basis for computational QSAR studies since the 1960s when Dr. Corwin Hansch established the very first QSAR analysis to predict chemical solubility[17]. However, the definition of similarity is not straightforward because the estimated degree of similarity depends on a number of underlying factors such as molecular descriptors, variable selection methods, and the similarity metrics.

To briefly explain the fundamental concepts, any QSAR method can be generally expressed in the following form[18]:

$$P_i = \hat{k}(D_1, D_2, \dots, D_n) \dots \dots \dots (1.1)$$

Where  $P_i$  is the biological activity of molecule I (dependent variable),  $D_1, D_2, \dots, D_n$  are independent variables, which are either calculated molecular descriptors or experimentally measured properties of molecule i, and  $k(D_i)$  is a function that relate the descriptors to the biological activity  $P_i$ .  $k(D_i)$  could be either linear (whose output is directly proportional to its input variables) or nonlinear (whose output is not directly proportional to

its input variables) function, depending on the expected relationship between the descriptor values  $D$  (input variables) and target property  $P$  (output). In essence, all machine learning techniques aim to find such mathematical representation of  $k(D_i)$  that would best reproduce the trend in biological activities.

The recent explosive growth of experimental data due to the technological advances in High Throughput Screening (HTS)[19–22] calls for the use of fast QSAR methods to establish QSAR models of large and complex data sets. During the past few decades of development, the field of QSAR has grown rapidly in terms of novel molecular descriptors, nonlinear regression methods, QSAR for toxicity and ADME (Absorption, Distribution, Metabolism, and Excretion), and 3D QSAR[23–28]. The differences among various QSAR approaches mainly depend on the descriptors used to characterize the molecules and the machine learning methods used to establish relationships between input descriptor values and biological activities. To list a few popular methods, nonlinear approaches of multivariate analysis include the Decision Trees[29], Random Forest (RF)[30], Artificial Neural Networks (ANN)[31],  $k$  Nearest Neighbors ( $k$ NN)[32], and Support Vector Machines (SVM)[33]. However, the most serious issue faced by these methods is the High-Dimension Low-Sample Size (HDLSS) problem, which means that the number of descriptors (usually from hundreds to thousands) is much greater than the number of samples in the studied dataset (less than a hundred compounds is common). To overcome this problem, we have applied recent developed Distance Weighted Discrimination (DWD) method[34] that was developed as a more robust alternative to SVM and is capable of handling HDLSS problem common for small modeling datasets.



### 1.3. Validation Criteria for Virtual Screening

Aside from interpretation of found relationships, important practical application of validated QSAR models is to screen large untested databases to assist the discovery of novel bioactive chemical entities [6,7]. At this point, two important aspects should be clarified: the classification of QSAR approaches based on the nature of the modeled response variable (**target property**), and the importance of rigorous model validation[9].

Generally speaking, QSAR approaches can be grouped in to three classes according to the target properties (referred to as dependent variables or response variables in statistical data modeling sense): classification, category, and continuous QSAR[35]. To explain in more detail, classes of target properties are different from categories in terms of whether or not they can be ordered in some scientifically meaningful way. The former, also regarded as categorical unrelated, cannot be rank ordered, i.e., classes do not relate to each other in any continuum. For example, compounds belonging to different pharmacological classes (interacting with different receptors) or classified as drugs vs. non-drugs cannot be rank ordered. On the other hand, the categorical related, can be rank ordered as the classes of target properties that cover certain ranges of values, e.g., very active, active, moderately active, and inactive. For the purpose of subsequent analysis, such classes are often encoded numerically (for example, one for active or zero for inactive). Continuous QSAR is based on the real values covering certain range, e.g.,  $pK_i$  (log-transformed binding constant values),  $IC_{50}$ ,  $ED_{50}$ , *etc.* Understanding this classification is very important when considering the nature of target property, its data quality, the choice of molecular descriptors and associated modeling techniques. Often, continuous activity data can be categorized and modeled as such to avoid fitting models to the experimental noise.

The choice of validation procedures and criteria is often dictated by the type of target properties which defines the classes of QSAR practices. For validation of QSAR models, Y-randomization test (randomization of the response variable)[36] is often used to check for the possibility of chance correlation[37]. Y-randomization procedure is discussed in detail in Chapter 2. In addition, the most critical way to ensure the predictive power of a QSAR model is estimating its performance on a validation (test) set which was not used in model development[38]. The model must demonstrate a significant correlation between predicted and observed target activities of compounds in such an external dataset. The practical way to achieve this is to divide experimental data into the training and test sets[39]. The criteria to select models from the training set, however, can be subject to a series of filtering rules. Many authors apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedure to the entire modeling dataset, which is now considered as insufficient for rigorous model validation[40]. For continuous QSAR models, the outcome of this procedure is a cross-validated correlation coefficient  $q^2$  for the training set of compounds, and  $R^2$  for the test set, which are calculated respectively by the formulas[38]:

$$q^2 = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y})^2} \dots\dots\dots(1.2)$$

$$R^2 = \frac{\sum (y_i - \bar{y})^2 (\tilde{y}_i - \bar{\tilde{y}})^2}{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}})^2} \dots\dots\dots(1.3)$$

where  $y_i$ ,  $\tilde{y}_i$ , and  $\bar{y}$  (or  $\bar{\tilde{y}}$ ) are the actual, predicted, and the average actual (or predicted) activities, respectively. We emphasize highly on the ability of the models to predict the activity of compounds in an external validation set, instead of only considering high  $q^2$  as an

indicator or even the ultimate proof of the predictive power of a QSAR model, which is often misleading and cannot guarantee the extrapolation power of respective models.

Correct classification rate (*CCR*) is often used to evaluate the predictivity of a binary classification model (i.e., for a two categories of activity that are usually called “active” and “inactive”). *CCR* is the average of sensitivity (*SE*) and specificity (*SP*), which are calculated by below formulas[41]:

$$\text{Sensitivity}(SE) = \frac{TP}{TP + FN} \dots\dots\dots(1.4)$$

$$\text{Specificity}(SP) = \frac{TN}{TN + FP} \dots\dots\dots(1.5)$$

$$CCR = \frac{SE + SP}{2} \dots\dots\dots(1.6)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \dots\dots\dots(1.7)$$

where *TP*, *TN*, *FP*, *FN* are true positives (accurately predicted actives), true negatives (accurately predicted inactives), false positives (inactives predicted as actives), and false negatives (actives predicted as inactives), respectively. Together they compose a confusion matrix (Figure 1.2) which is the common resort to evaluate a classifier. *CCR* is preferred as a performance measure of a classifier since it is not biased to the major class in the case of imbalanced data in which the minority class is often more important (e.g., active compounds are often fewer than inactive ones). However, predictive *accuracy* (Equation 1.7) simply favors performance of the majority class.

It should be noted that data balancing is an important issue to consider before running a classification modeling. Most machine learning algorithms assume that their training sets are well balanced, and demonstrate poor performance when they deal with imbalanced data

sets[42]. However, in many cases we cannot control the influence of this imbalance issue by simply re-sampling the data sets (e.g., removal of certain cases in the major class). As a result, we need to resort to other algorithms resistant to the issue.

#### 1.4. Thesis Outline

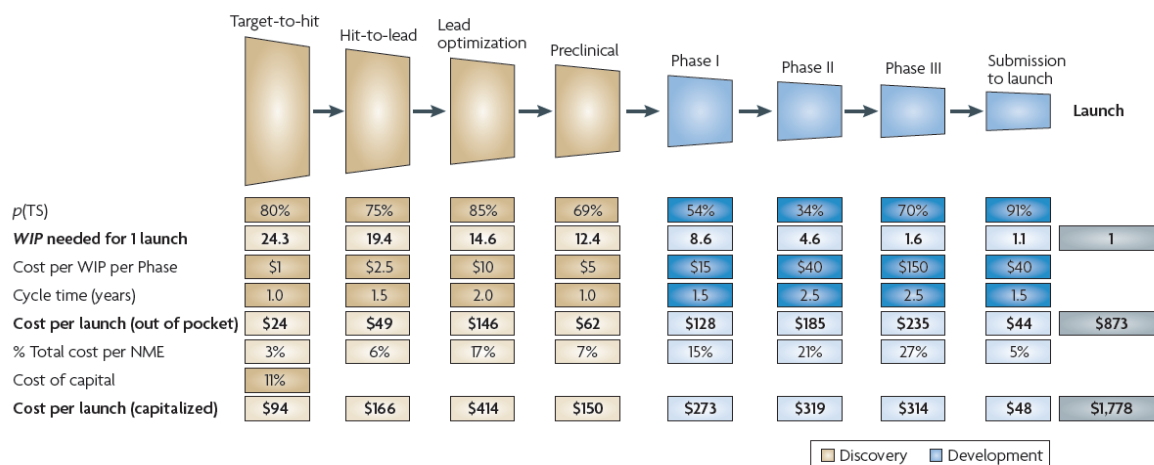
This dissertation concentrates on the application of QSAR approaches to discover novel antipsychotic drugs. Extensive efforts have been made in terms of data collection and curation, QSAR modeling, and the quest for suitable evaluation criteria.

Chapter 2 illustrates the successful practice of identifying FDA-approved drugs with newfound 5-HT<sub>7</sub> binding affinity by applying rigorous QSAR modeling workflow. The 5-HT<sub>7</sub> receptor, a member of the GPCR family, is postulated to be a potential drug target for psychotic disorders, especially for schizophrenia. A combi-QSAR approach established by our lab was used to develop predictive continuous models using  $k$  Nearest Neighbor ( $k$ NN) and classification models using Distance Weighted Discrimination (DWD). Models were rigorously validated by Y-randomization and demonstrated high accuracy in predicting external datasets. VS of the publically available compound database World Drug Index (WDI) followed by experimental testing successfully identified five known drugs with first identified 5-HT<sub>7</sub> binding affinity. Two of these drugs have been confirmed as 5-HT<sub>7</sub> receptor antagonists, which could be repositioned to treat schizophrenia.

In Chapter 3, we propose a new evaluation metric, the Economic Ratio (ER), not only as a performance parameter for the developed models, but also as a target function during model training. After applying this metric with the Decision Tree (DT) machine learning method to various datasets, we found that some trees generated using ER differ in structure and performance from those generated using traditional metrics for branch selection. The

cost/benefit economic ratio, ER, can thus be used in two different but complementary ways: (1) as a benchmarking criterion to compare the prediction performances of various classifiers and (2) as a target function to build QSAR classification trees.

In Chapter 4 and Chapter 5, we extend our modeling strategy to develop multiple robust predictors for a set of GPCR targets. The resulting models can be used to predict the binding bioprofiles of untested chemicals. In summary, we curated and integrated binding data for thousands of GPCR ligands extracted from both ChEMBL and PDSP databases. First, we used 5-HT<sub>1A</sub> as a scheme to decide what properties should be applied for data collection and modeling processes, which resulted in rigorous standards for both chemical and biological data curation. We then developed robust classification QSAR models based on the ligands of the large set of GPCRs; that is, 34 GPCRs in total including 5-HT<sub>1A</sub>. The validated models were applied to assess the GPCR binding profiles of 13 drugs not present in the modeling sets, and we found the accuracy was as high as 70.5%. This extensive study yielded a compendium of validated QSAR potency predictors, the GPCR QSARome, providing an effective *in silico* means to search for novel antipsychotic drugs and to unveil their complex polypharmacological nature.



**Figure 0.1. R&D model yielding costs to successfully discover and develop a single new molecular entity (NME).**

Money unit: million. Work in process, WIP. Probability of successful transition from one stage to the next,  $p(TS)$ .

(Modified from *Steven M. Paul, et al. Nature Reviews Drug Discovery, 2010, 9: 203-214.*)

	Predicted Negative	Predicted Positive
Actual negative	TN	FP
Actual Positive	FN	TP

**Figure 0.2. The confusion matrix used to evaluate a classifier.**

**Note:** The columns are the predicted classes and the rows are the actual classes. TN is the number of negative cases correctly predicted (True Negatives), FP is the number of negative cases incorrectly predicted as positive (False Positives), FN is the number of positive cases incorrectly predicted as negative (FN), and TP is the number of positive cases correctly predicted as positive (True Positives).

## CHAPTER 2

# APPLICATION OF CURRENT CHEMINFORMATIC TECHNIQUES TO HUMAN 5-HT<sub>7</sub> DATASETS TO BUILD VALIDATED AND PREDICTIVE QSAR MODELS FOR DRUG REPURPOSING

### 2.1. Introduction

5-hydroxytryptamine (5-HT) receptors are involved in a large number of physiological and behavioral functions[43–46]. Many antipsychotic drugs act through multiple molecular targets including 5-HT receptors. Although it received little attention when first cloned in 1993, the 5-HT<sub>7</sub> receptor has become the most studied member of the 5-HT receptor family now[47]. Several distribution studies indicated that the 5-HT<sub>7</sub> receptors are located mainly in thalamus, hippocampus, and hypothalamus with relatively lower concentrations in the amygdala and cerebral cortex[48,49]. Additionally, 5-HT<sub>7</sub> receptor subtypes are found in smooth muscle cells and other peripheral tissues[50]. Scientific research on the 5-HT<sub>7</sub> receptor has mainly focused on its therapeutic effects for psychiatric disorders, especially for major depression[51] and schizophrenia[52]. Previous studies show that 5-HT<sub>7</sub> antagonists modulate the level of 5-HT and thus increase neurogenesis, indicating 5-HT<sub>7</sub> receptor is a promising molecular target for antidepressants[53]. A series of studies identify 5-HT<sub>7</sub> receptors as critical in hippocampus-dependent functions including learning and memory[54–56]. In addition, the presence of 5-HT<sub>7</sub> receptor subtypes in smooth muscle cells suggests

that 5-HT<sub>7</sub> ligands could be effective therapeutic agents for migraine[57]. Other possible roles for the 5-HT<sub>7</sub> receptor include hypertension[58] and irritable bowel syndrome[50,59].

Despite the fact that the 5-HT<sub>7</sub> receptor is an attractive target for psychiatric disorders and many other diseases, no 5-HT<sub>7</sub> ligands (antagonists or agonists) are currently available for clinical use. Recent projects to identify potentially therapeutic 5-HT<sub>7</sub> ligands have relied on high throughput screening or chemical optimization of existing 5-HT<sub>7</sub> ligands[60–69]. Their results were limited by low hit rate or only modification of existing chemical scaffolds, and none of the ligands produced by their studies have proved successful in clinical practice.

To address these problems, we employed QSAR modeling to search for potent 5-HT<sub>7</sub> ligands. Our laboratory has used a combinatorial QSAR strategy involving the machine learning method *k* Nearest Neighbor (*k*NN) for several years (cf., Figure 2.1). Previous studies demonstrate that virtual screening with combinatorial QSAR models yields diverse chemicals targeting the protein or pathway of interest[8,70–74]. In this study, we incorporated the Distance Weighted Discrimination (DWD)[34] method into our combi-QSAR approach. Two datasets used as the basis for our modeling were provided by the NIMH Psychoactive Drug Screening Program (PDSP) and extracted from the World of Molecular Bioactivity (WOMBAT) database, respectively (see **Materials and Methods**).

Little study has been done in this area with a search of literature uncovering only two publications focused on ligand-based modeling of 5-HT<sub>7</sub> ligands. Based on a set of 22 5-HT<sub>7</sub> inverse agonists, Vermeulen *et al.*[75] built both a pharmacophore model and a CoMFA model ( $R^2=0.97$ ,  $SE=0.18$ ). Using the pIC<sub>50</sub> values of 81 quinazolinone derivatives, Jalali-Herav *et al.*[76] used a modified modeling approach which combined ant colony optimization (ACO) and adaptive neuro-fuzzy interference system (ANFIS) to generate



QSAR models ( $R^2=0.775$ ). Neither of these previous studies performed external validation of the their models or applied their models to virtual screening.

Through use of rigorous external validation, we demonstrated that our QSAR models are predictive. Application of *k*NN models along with DWD classification models to virtual screening identified droperidol and perspironone as potent 5-HT<sub>7</sub> binders. The subsequent binding assays confirmed that their binding affinity were at nano-molar level. Neither drug had previously been studied for its relationship to the 5-HT<sub>7</sub> receptor and therefore, are both novel binders. Most of the confirmed hits produced by the study are marketed drugs, indicating the methods may be useful in drug repurposing.

## 2.2. Materials and Methods

### 2.2.1. Data

The original dataset provided by PDSP contained 137 compounds including 67 5-HT<sub>7</sub> binders ( $K_i$  values greater than 10,000nM) and 70 non-binders ( $K_i$  values less than 10,000nM), whereas the dataset extracted from the WOMBAT database had 80 binders (See Supporting Information). Both datasets were curated according to the “Trust, but Verify” protocol established by our lab[77]. The binding affinity of 5-HT<sub>7</sub> binders ( $K_i$ ) provided by PDSP were measured as described previously[78].  $K_i$  values, which spanned over four orders of magnitude for both datasets (cf., Figure 2.2), were converted to the  $pK_i$  scale ( $-\log K_i$ ) in which higher values indicated exponentially greater binding affinity.

Virtual screening (VS) using developed models based on the datasets was applied to the curated World Drug Index (WDI) database containing ca. 52,000 compounds[79]. The VS hits were assayed by the same lab providing the PDSP dataset.

### 2.2.2. Generation of MolConnZ Descriptors

The MolConnZ4.09 (MZ4.09) software[80] was used for the computation of a wide range of topological indices (descriptors) of molecular structures[81–88]. MZ4.09 produced more than 800 different descriptors. Descriptors with zero variance were eliminated from consideration. Also, if a pair of descriptors had correlation coefficient larger than 0.95, one would be eliminated. MZ4.09 descriptors were range-scaled because the absolute values of individual types could differ by orders of magnitude[32]. Range scaling prevents undesirable overweighting of descriptors with high ranges of values when calculating compound similarities in QSAR modeling procedures. Descriptor values for the entire dataset were scaled to be within the interval of [0,1].

### 2.2.3. Applicability Domain (AD)

A QSAR model can predict the target property of any compound for which chemical descriptors can be calculated. However, if a compound is highly dissimilar from all compounds of the modeling set, reliable prediction of its activity is unlikely. A concept of AD was developed to avoid such an unjustified extrapolation in activity prediction.

In regression modeling, AD was defined as the Euclidean distance threshold  $D_T$  between a compound under prediction and the compound's closest nearest neighbor in the training set. Euclidean distance  $d_{ij}$  between any two molecules  $i$  and  $j$  in the  $M$ -dimensional descriptor space ( $M$  is the number of selected descriptors) and  $D_T$  was calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (2.1)$$

$$D_T = \bar{y} + Z\sigma \quad (2.2)$$

Here,  $\bar{y}$  is the average Euclidean distance between each compound and its  $k$ -nearest neighbors in the training set (where  $k$  is the parameter optimized in the course of QSAR modeling, and the distances are calculated using descriptors selected by the optimized model only),  $\sigma$  is the standard deviation of these Euclidean distances, and  $Z$  is an arbitrary cutoff parameter to control the significance level. We set the default cutoff value of  $Z$  ( $Z_{\text{cutoff}}$ ) at 0.5, which formally placed the allowed distance threshold at the mean plus one-half of the standard deviation.

We also defined the AD in the entire descriptor space to exclude outliers for VS databases. The same formula (2.2) was used with  $k=1$  and  $Z_{\text{cutoff}}=0.5$ , but Euclidean distances were calculated using all calculated descriptors.

Thus, if the distance of the external compound from its nearest neighbor in the training set within either the entire descriptor space or the selected descriptor space exceeds these thresholds, a prediction is not made. This approach has been applied in our recent work[5,8,70,89].

#### 2.2.4. Structural Outliers Exclusion

A similarity search based on Euclidean distance was performed prior to modeling procedures to exclude structural outliers, i.e., compounds that are highly dissimilar to the majority in the dataset. In our studies, molecular dissimilarity is denoted by the Euclidean distance (Equation 1). For the PDSP dataset,  $Z_{\text{cutoff}}$  values were set as 1.5 to keep 62 binders for regression modeling and 3.0 to include another 38 non-binders to compose a classification modeling set. For the WOMBAT dataset,  $Z_{\text{cutoff}}$  value was 0.8 to exclude 14 outliers from the 80 5-HT<sub>7</sub> binders.

### 2.2.5. Modeling and External Evaluation Sets

For all regression and classification datasets, 18% of the compounds were randomly selected and designated as the external evaluation set while the remaining compounds were deemed the modeling set. The compounds with the highest and lowest activity values were kept in the modeling set for regression modeling. The modeling sets were used for construction and selection of models while the external evaluation sets, which were not involved in the process of model development, were used to validate the predictive power of accepted models.

### 2.2.6. DWD Classification Method

Distance Weighted Discrimination (DWD) was developed by Marron et al.[34] to classify High-Dimension Low-Sample Size (HDLSS) datasets in a wide range of applied contexts. The method was invented to circumvent the drawback of the popular Support Vector Machines (SVM) method[33] in HDLSS settings, as illustrated in Figure 2.3. The toy data used there consist of two classes, each with 20 data points, that follow a 50 dimensional spherical Normal distribution. The only difference is that the means have been shifted. When the data have been projected onto the direction of shift, the result is shown in Figure 2.3a. The projections of the data are shown on the horizontal axis, and a random height is used for good visual separation. Smooth histograms also show the subpopulation nature of the data. Note that good separation of the classes is available from projecting the data onto this optimal direction. However, this direction is unknown, and is challenging to find from the full 50 dimensional data set. A potential candidate, called the Maximal Data Piling (MDP) direction, studied in Ahn and Marron (not published yet), is considered in Figure 2.3b. Note that in this direction, projections of the data for each class pile up at a single point (hence the

name). To make this happen, the MDP direction must be driven by small scale noise artifacts of the data. This indicates that this classification direction will have poor generalizability properties, which is confirmed by the large angle ( $58^\circ$ ) to the optimal direction. SVM provides a major improvement over MDP, as shown in Figure 2.3c. These projections show even better separation of the data (in fact this separation is maximized by the SVM direction). The much better generalizability of SVM is shown by the much smaller angle ( $36^\circ$ ) to the optimal direction. However, note that there are also some data piling issues for SVM, with a number of points piled on the margins (the boundaries of the empty region between classes). This again suggests undue influence from small scale noise artifacts which may be hampering generalizability. DWD overcomes this problem by modifying the underlying optimization problem, to one which allows all data points to have a stronger influence on the direction vector. In the resulting projection plot shown in Figure 2.3d, the projections are similarly separated, but there is no data piling. The beneficial effect of this is better generalizability, reflected by the even smaller angle of  $26^\circ$ .

Like SVM, the computation of DWD is based on computationally intensive optimization, but while SVM uses well-known quadratic programming algorithms[90], DWD uses recently developed interior-point methods for so-called Second-Order Cone Programming (SOCP) problems[91]. Detailed comparison of DWD with SVM is given by Marron et al.[34]. In this study, Matlab-based algorithms of SVM and DWD were used (available at [http://www.unc.edu/~marron/marron\\_software.html](http://www.unc.edu/~marron/marron_software.html)).

### 2.2.7. Sphere Exclusion Algorithm

For regression modeling, the modeling set was divided into multiple training and test sets. Ideally, the test sets should satisfy the following criteria: 1) The distribution of activities

in training and test sets should be similar. 2) Training set should be distributed within the entire area of the dataset distribution. 3) All points of the test set should be within the AD defined by the training set at least in the entire descriptor space. 4) Each point of the training set should be close to at least one point of the test set. Requirement 4) can be satisfied by dividing a dataset into a small number of bins and selecting one compound from each bin as well as most active and most inactive compound into the training set. The Sphere Exclusion algorithm developed by our lab and used in several publications[39,92] was applied to address criteria 2) – 4).

#### 2.2.8. *k*NN Regression Modeling Method

The *k* Nearest Neighbor (*k*NN) QSAR method used in this study employs the *k*NN pattern recognition principle[93] and sphere exclusion mentioned above. In short, a subset of variables (descriptors) is selected randomly as a Hypothetical Descriptor Pharmacophore (HDP)[7]. The HDP is validated by leave-one-out cross-validation (LOO-CV), where each compound is eliminated from the training set and its 5-HT<sub>7</sub> binding affinity is predicted as the weighted average of the binding affinity of the *k* most similar molecules (*k* varies from 1 to 5):

$$w_i = \frac{e^{-d}}{\sum_{i=1}^k e^{-d}} \quad (2.3)$$

$$\tilde{y} = \sum_{i=1}^k w_i y_i \quad (2.4)$$

In equations (2.3) and (2.4),  $\tilde{y}$  is the predicted activity value of the compound and  $y_i$  is the experimentally measured activity value of its *i*-th nearest neighbor. The dissimilarity  $d_i$  between the query molecule and its *i*-th nearest neighbor was represented by the Euclidean

distance (equation 2.1) between the corresponding points in the multidimensional descriptor space. Essentially, the neighbor with the smaller distance from a compound is given a higher weight  $w_i$  in calculating the predicted activity.

Simulated annealing was used to select the optimal list of variables, i.e. HDP. Details of the  $k$ NN method implementation including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere[32].

The following **acceptance criteria** were used for selecting regression QSAR models[38]: (i) leave-one-out (LOO) cross-validated  $q^2$  (which is also used as the target function, i.e., it is optimized by the QSAR modeling procedure); (ii) square of the correlation coefficient  $R$  ( $R^2$ ) between the predicted and observed activities; (iii) coefficients of determination (predicted versus observed activities  $R_0^2$ , and observed versus predicted activities  $R_0'^2$ ); (iv) slopes  $k$  and  $k'$  of regression lines (predicted versus observed activities, and observed versus predicted activities) through the origin. These **criteria** were calculated according to the following formulas:

$$q^2 = 1 - \frac{\sum_{i=1}^N (y_i - \tilde{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2.5)$$

$$R = \frac{\sum_{i=1}^n (y_i - \bar{y})(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\tilde{y}_i - \bar{\tilde{y}})^2}} \quad (2.6)$$

$$R_0^2 = 1 - \frac{\sum_{i=1}^n (\tilde{y}_i - \tilde{y}_i^{r_0})^2}{\sum_{i=1}^n (\tilde{y}_i - \bar{\tilde{y}})^2} \quad (\text{predicted vs. observed}) \quad (2.7a)$$

$$R_0'^2 = 1 - \frac{\sum_{i=1}^n (y_i - y_i^{r_0})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (\text{observed vs. predicted}) \quad (2.7b)$$

$$k = \frac{\sum_{i=1}^n y_i \tilde{y}_i}{\sum_{i=1}^n y_i^2} \quad (\text{predicted vs. observed}) \quad (2.7a)$$

$$k' = \frac{\sum_{i=1}^n y_i \tilde{y}_i}{\sum_{i=1}^n \tilde{y}_i^2} \quad (\text{observed vs. predicted}) \quad (2.7b)$$

where  $y_i$  and  $\tilde{y}_i$  are observed and predicted activities,  $R_0^2$  and  $R_0'^2$  are the coefficients of determination for regressions through the origin for predicted vs. observed, and observed vs. predicted activities, respectively,  $k$  and  $k'$  are the corresponding slopes, and  $\tilde{y}^{r_0} = ky$  and  $y^{r_0} = k'\tilde{y}$  are the regressions through the origin for predicted vs. observed and observed vs. predicted activities.

In this study, **acceptance criteria** for  $k$ NN regression models were set to (i)  $q^2 \geq 0.7$ ; (ii)  $R^2 \geq 0.7$ ; (iii)  $(R^2 - R_0^2)/R^2 < 0.1$  and  $0.90 \leq k \leq 1.10$ , or  $(R^2 - R_0'^2)/R^2 < 0.1$  and  $0.90 \leq k' \leq 1.10$ ; (iv)  $|R_0^2 - R_0'^2| < 0.2$ .

### 2.2.9. Y-Randomization Test

To establish model robustness, the Y-randomization (randomization of the response variable) test was carried out. This test consists of repeating all the calculations with scrambled activities of the training sets. The goal of the procedure is to evaluate the possibility that good statistical results are due to over-fitting or chance correlation. The statistical significance of QSAR models for training sets was evaluated with the standard



hypothesis testing method[94,95]. In this approach, two alternative hypotheses are formulated: (1) for  $H_0$ ,  $h=\mu$ ; (2) for  $H_1$ ,  $h>\mu$ , where  $\mu$  is the average value of  $q^2$  for random models and  $h$  is that for the actual models. The null hypothesis,  $H_0$ , states that the QSAR models for the actual dataset are not significantly better than random models whereas the alternative hypothesis,  $H_1$ , assumes the opposite (i.e., that the actual models are significantly better than the random models). Hypothesis rejection is based on a standard one-tail test, which involves the following procedure.

(1) Determine the average value of  $q^2$  ( $\mu$ ) and its standard deviation ( $\sigma$ ) for random models;

(2) Calculate the Z score that corresponds to the average value of  $q^2$  ( $h$ ) for the actual models:

$$Z = (h - \mu) / \sigma \quad (2.8)$$

(3) Compare this Z score with the tabular critical values of  $Z_c$  at different levels of significance ( $\alpha$ )[96] to determine the level at which  $H_0$  should be rejected. If the Z score is higher than tabular values of  $Z_c$  (cf., Table 2.1), one concludes that at the level of significance that corresponds to that  $Z_c$ ,  $H_0$  should be rejected while  $H_1$  should be accepted.

#### 2.2.10. Consensus Prediction and Virtual Screening (VS) using $k$ NN Regression Models

It is critical to validate QSAR models by assessing their prediction accuracy for an external evaluation set which was not used in model building and selection. Our previous experience suggests that results obtained by consensus prediction (i.e. by averaging predictions from multiple QSAR models) are often more accurate than predictions made by individual models. Compounds in the external evaluation set were predicted by all models that passed **acceptance criteria**. Each compound was predicted by a model only if the

compound was determined to be in the AD for that model. The average predicted activity, the variance of the prediction values, and the fraction of models that predict the activity ( $\geq 90\%$  in this study) were calculated for each compound.

Predictive QSAR models were used to virtually screen the WDI database. MZ4.09 descriptors were calculated for each compound in the database and normalized based on the minimal and maximal values of each descriptor for the regression datasets and classification dataset, respectively. As illustrated in the workflow (cf. Figure 2.1), first, the DWD classification model was applied to filter out compounds that were predicted as non-binders; then consensus prediction using rigorously validated *k*NN-QSAR models was applied. Each compound was required to be within the global AD defined by the entire descriptor space. Furthermore, a compound was considered out of the AD if it was found out of the ADs of more than 10% of all used models. There was also a threshold on the standard deviation of estimations across all used models: if it was higher than 0.80, prediction was considered unreliable. Finally, available consensus hits were submitted for experimental validation.

#### 2.2.11. Experimental Validation of Screening Hits

Experimental validation of the screened hits was performed as per the Assay Protocol Book available on the NIMH Psychoactive Drug Screening Program (PDSP) website (<http://pdsp.med.unc.edu/>), including both binding assay and functional assay methods[3].

### 2.3. Results and Discussion

#### 2.3.1. DWD Classification Modeling

DWD and SVM classification models were built and validated with the dataset composed of 100 compounds (62 binders and 38 non-binders) after similarity search ( $Z_{\text{cutoff}} = 3.0$ ). It was then split into a training set (51 binders and 31 non-binders) and a test set (11

binders and 7 non-binders). The classification models discriminated binders and non-binders based on their projections on the DWD or SVM direction vectors. Compounds with projections larger than the DWD or SVM thresholds were predicted as binders, while those with smaller projections were predicted as non-binders. Correct Classification Rate (CCR) was calculated to evaluate the performance of the classification models:

$$CCR = \frac{1}{2} \left( \frac{TP}{N_1} + \frac{TN}{N_0} \right) \quad (2.9)$$

where  $N_1$  and  $N_0$  are the number of binders and nonbinders in the dataset, TP and TN are the number of known binders predicted as binders (true positives) and the number of non-binders predicted as non-binders (true negatives). CCRs for the training and test sets were denoted as  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$ , respectively.

The threshold of the SVM model (0.2) was determined by the average of the boundaries for binders and non-binders, whereas that of the DWD model (0.8) was determined by the intersection point of the probability distribution curves of the two classes.

Results of the DWD and SVM classification models built on the same dataset are shown in Figure 2.4.  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  of the DWD model were 0.88 and 0.93, respectively.  $CCR_{\text{train}}$  of the SVM was 1.00; however, the more meaningful  $CCR_{\text{test}}$  was only 0.72 (cf., Table 2.2). Furthermore, the range of the entire data points for SVM was narrower than that for DWD. As a result, we conclude that DWD is a better method to classify the status of two classes, especially when dealing with a relatively small dataset.

### 2.3.2. *k*NN QSAR Regression Modeling

Similarity analysis was performed as described above to exclude structural outliers in the original datasets. Consequently, 62 binders from PDSP and 66 binders from WOMBAT were remained to build respective regression models ( $Z_{\text{cutoff}}=1.5$  and 0.8, respectively).

The PDSP dataset was first randomly divided into a modeling set (51 binders) and an external evaluation set (11 binders). After applying the Sphere Exclusion algorithm to the modeling set, 29 splits of training and test sets were ultimately accepted. 216 models were accepted for further consensus prediction. These models satisfied all **acceptance criteria**.

Y-randomization test was carried out with randomized activity values for the training sets. The largest  $q^2$  obtained from random dataset was 0.76 with poor  $R^2=0.002$ . The Z score that corresponds to the average  $q^2$  value in the standard one-tail hypothesis test was 2.88, indicating that the level of significance  $\alpha < 0.01$  (cf., Figure 2.5 and Table 2.1). The result of Y-randomization test confirmed that the result obtained for the actual dataset was statistically better than those obtained for random datasets at the given level of significance, meaning that the actual  $k$ NN models were robust.

### 2.3.3. Models Validation using External Evaluation Set

The 216  $k$ NN/MZ4.09 models based on the PDSP dataset that satisfied all **acceptance criteria** were used to make a consensus prediction of the binding affinity of the external evaluation set. Figure 2.6 shows the correlation between experimentally measured and predicted binding affinity for this external evaluation set. Statistical results suggested the models have reasonable predictive abilities, with  $R^2$  of 0.61 and  $R_o^2$  of 0.59 for 11 compounds. While most compounds were predicted within a reasonable range of log units, the compound ketanserin was a significant activity outlier (prediction error as large as 1.86 log units).

Modeling based on the WOMBAT dataset followed the same procedures, and the results from both PDSP and WOMBAT are summarized in Table 2.3.

#### 2.3.4. QSAR-based Virtual Screening

The accepted DWD and *k*NN models were selected for virtual screening based on their strong performance in model validation. The DWD classification model was applied first and filtered out about 39% compounds as non-binders from the WDI database. Then the 216 *k*NN/MZ4.09 models based on the PDSP dataset with defined applicability domains were applied to screen the remained compounds. In consensus prediction, predicted  $pK_i$  values from individual model were averaged. Finally, 43 structurally diverse hits with predicted  $pK_i$  values greater than 7.98 were prioritized as the most potent consensus hits.

For each consensus hit, we searched published literature and databases using both PubMed and the ChemoText knowledge base[97], but found that none of the compounds had ever been reported as 5-HT<sub>7</sub> receptor binders. Some of the compounds, however, were reported as potent binders of related neural receptors, or possessing antipsychotic activities with unknown mechanisms of actions.

#### 2.3.5. Experimental Validation

Based on the commercial availability of the concensus hits, seven hits were purchased (see supporting information) and submitted for binding affinity assay and then functional assay if identified as a binder.

As shown in Table 2.4, droperidol and perospirone had the best  $K_i$  values as 3.5 *nM* and 8.6 *nM*, respectively. Altanserin and clomipramine had moderate  $K_i$  values as 143 *nM* and 46 *nM*, respectively. Pravadoline possessed micro-molar scaled binding affinity as 3.18  $\mu M$ . The remaining two compounds showed no binding affinity to 5-HT<sub>7</sub> receptor. The results showed a high hit rate for our virtual screening strategy (5 out of 7). Although the sample size was

too small to estimate the overall hit rate, more potential receptor binders are expected among those hits that are not currently available.

Predictions of all above compounds by the *k*NN models based on the WOMBAT dataset are also summarized in Table 2.4. The results are very consistent with the predictions by the PDSP models.

A functional assay of six hit compounds were then applied to identify their binding functions (agonist or antagonist). According to the primary functional assay results, droperidol and perospirone completely inhibited 5-HT<sub>7</sub> receptor activity at the concentration of 10 μM, indicating both were 5-HT<sub>7</sub> receptor antagonists. Other tested compounds including altanserin and clomipramine showed neither agonist activity nor antagonist activity (cf., Figure 2.7).

Amisulpride was recently identified as a potent 5-HT<sub>7</sub> receptor antagonist[94], a finding confirmed by [3H]LSD labelled competition binding assay ( $K_i = 11.5 \pm 0.7$  nM). However, amisulpride had a lower affinity for [3H]5-CT labelled 5-HT<sub>7a</sub> receptors ( $K_i = 135.5 \pm 15.8$  nM), which was consistent with our VS prediction ( $K_i = 174.7$  nM). Recently, Keiser MJ, et al. used chemical similarity to predict new molecular targets for known drugs - in nature, a drug-repurposing practice[95]. Their study confirmed that the drug N,N-dimethyltryptamine (DMT) was associated with serotonergic receptors including the 5-HT<sub>7</sub> receptor. The  $K_i$  value of DMT to 5-HT<sub>7</sub> receptor was reported as 210 nM in that publication, while the predicted  $K_i$  value by our models was 920 nM, which was in the acceptable range of prediction error. Lately, four compounds tested by Roth's lab were found in our prediction list. These are fluspirilene (predicted  $pK_i=7.52$ ; experimental  $pK_i=7.39$ ), raloxifene (predicted  $pK_i=7.87$ ; experimental  $pK_i=5.91$ ), DO-897 (predicted  $pK_i=6.74$ ; experimental  $pK_i=7.85$ ),

and fendiline (predicted  $pK_i=6.15$ ; experimental  $pK_i=5.51$ ). The predictions of fluspirilene and fendiline by our models were nearly consistent with the tested results. These compounds discussed above were not included in our assay list, because their VS prediction scores were not on the top of our VS hit list. The confirmation of these hits as 5-HT<sub>7</sub> ligands give further evidence that our models are reliable for virtual screening.

#### 2.3.6. Data quality and QSAR modeling

It is also noticed in this study that data quality affected modeling results significantly. We also extracted 81 5HT<sub>7</sub> binders with unique structures from the public online PDSP  $K_i$  database. Preparation of this dataset immediately brought our attention to the quality of the data. Most of the binders were studied in pharmacological context, and the same compound had variations in activity among different publications. Take a few of them for example (cf., Table 2.5), ergotamine had  $K_i$  values of 17.37, 138.03, and 1291 nM from three different labs;  $K_i$  values of ketanserin ranged from 206 to over 7943.28 nM; NAN-190 even had conflict  $K_i$  values (less and greater than 1000 nM, respectively). As mentioned before, ketanserin is also included in the external evaluation set of the PDSP dataset used successfully in this study, but it was an obviously activity outlier according to the external validation (cf., Figure 2.6A). Thus the accuracy of  $K_i$  value for ketanserin used in this study (162.5 nM) is questionable. Variations of more than three folds usually cause unreliability. Several treatments of this data according to species, radioligands, and even exclusion of compounds with the greatest activity variations all failed to generate successful QSAR models.

## 2.4. Conclusions

Our group has established a robust hit identification strategy that combines rigorously validated QSAR models and virtual screening[98,99]. It has been proved that the workflow is

capable of identifying potent compounds with novel chemical scaffolds. Specific cases are anticonvulsant agents[41], D1 dopaminergic antagonists[89] and HDAC inhibitors[72]. This paper shows that we can achieve even more promising results through a modified modeling strategy, and the strategy is quite sensitive to the data quality. To highlight the features in our current protocol, first, *k*NN models are built using variable selection to select the subset of descriptors that are predictive of biological activity. Second, DWD is a new statistical strategy to circumvent the overfitting problem commonly suffered by several popular classification methods like SVM. In this case, it classified 5HT<sub>7</sub> binders and non-binders better than SVM (Figure 2.4) as indicated by  $CCR_{test}$ , so the incorporation of DWD into the classification screening methods could potentially enhance the hit rate achieved by *k*NN models.

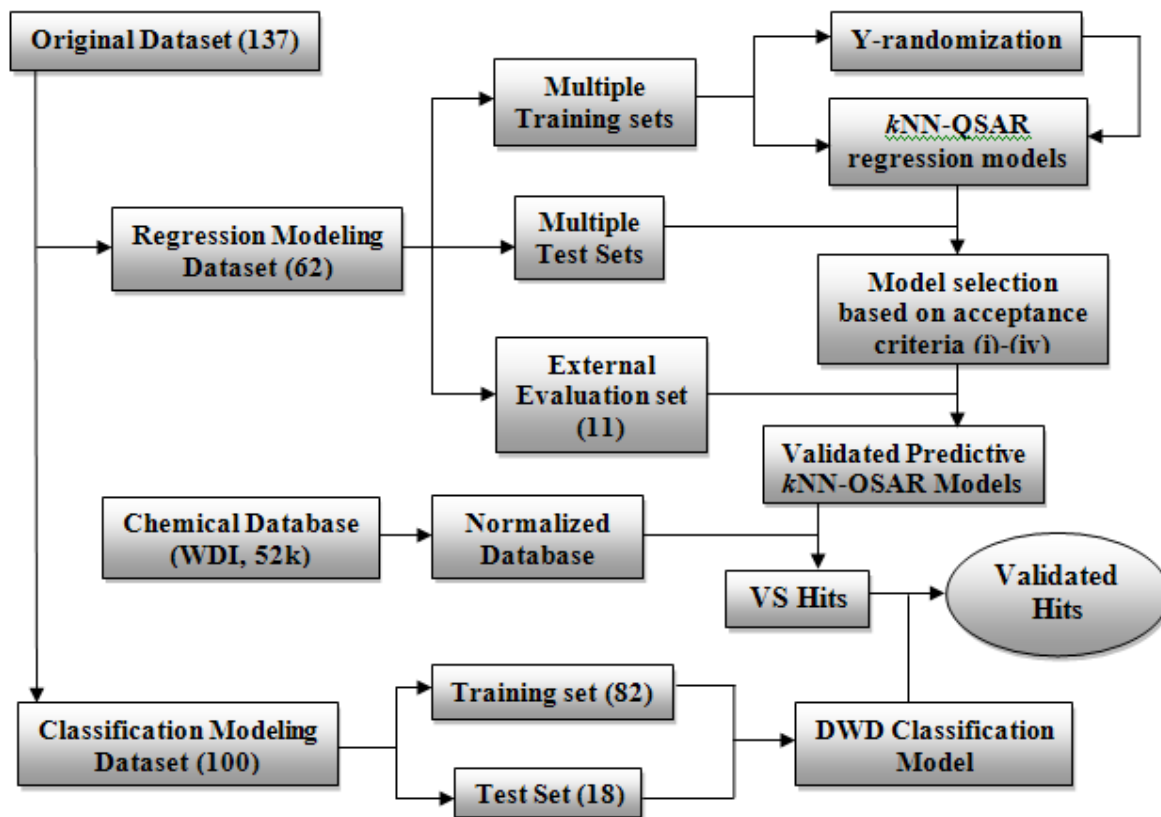
We subsequently used the rigorously validated models to screen the WDI database with over fifty thousand compounds, and selected 43 consensus hits that were predicted as potent 5-HT<sub>7</sub> binders. Seven commercially available hits were submitted for experimental validation. Among the seven, two were identified as potent 5-HT<sub>7</sub> receptor binders, two were confirmed to have medium binding affinity, one had weak binding affinity, and the remaining two were confirmed to be false positives. As shown in Table 2.4, droperidol had the lowest  $K_i$  value as 3.50 nM. Furthermore, all of the five confirmed binders were tested in a functional assay. The two best binders, droperidol and perospirone, completely inhibited 5-HT<sub>7</sub> activity (cAMP production) at the concentration of 10  $\mu$ M, and thus were identified as potent 5-HT<sub>7</sub> antagonists. In addition, we have found no evidence in literature that these drugs have been tested against the 5-HT<sub>7</sub> receptor, and therefore we believe that we have found novel 5-HT<sub>7</sub> binders.



Additional six 5-HT<sub>7</sub> binders confirmed by recent publications were also tested by our models. All were predicted as binders, in which amisulpride, DMT, fluspirilene, and fendiline were predicted quite close to the experimental results, with exception of raloxifene and DO-897.

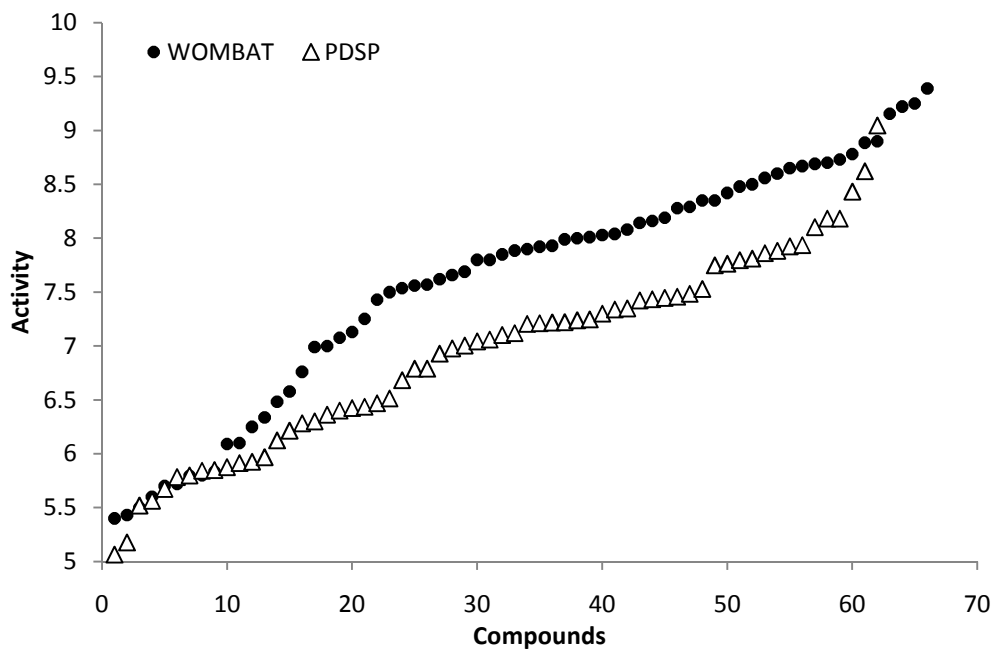
It is valuable that the two 5-HT<sub>7</sub> antagonists identified by our study, droperidol and perospirone, are both marketed drugs. Droperidol was approved by FDA before 1984 (FDA cannot verify dates on drugs approved before 1984). It is used in conjunction with an opioid analgesic such as fentanyl (e.g. innovar) to maintain the patient in a calm state of neuroleptanalgesia before the surgery[100]. It is also used as an antiemetic[101] and for the control of agitation in acute psychoses[102]. Droperidol binds strongly to postsynaptic GABA receptors[103] and selectively block alpha-adrenergic receptors[104]. It also binds to dopaminergic receptors such as D2 and D4 receptors potently[105]. However, the exact mechanism of action is still unknown. Our study revealed that droperidol might act through the antagonism of the 5-HT<sub>7</sub> receptor. In addition, since the 5-HT<sub>7</sub> receptor has high correlation with psychotic diseases such as schizophrenia, droperidol may be repurposed as a drug to treat schizophrenia. Perospirone is a novel atypical antipsychotic drug approved in 2001[106]. It was considered to act through antagonizing 5-HT<sub>2A</sub> and D2 receptors[107,108]. Since the 5-HT<sub>7</sub> receptor pharmacologically resemble 5-HT<sub>2</sub> receptors[78], perospirone may also target the 5-HT<sub>7</sub> receptor in addition to the well-known 5-HT<sub>2A</sub> and D2 receptors. Overall speaking, our findings may unveil both the physiological roles of the 5-HT<sub>7</sub> receptor and unknown mechanisms of action for certain drugs and chemicals.

## Tables and Figures

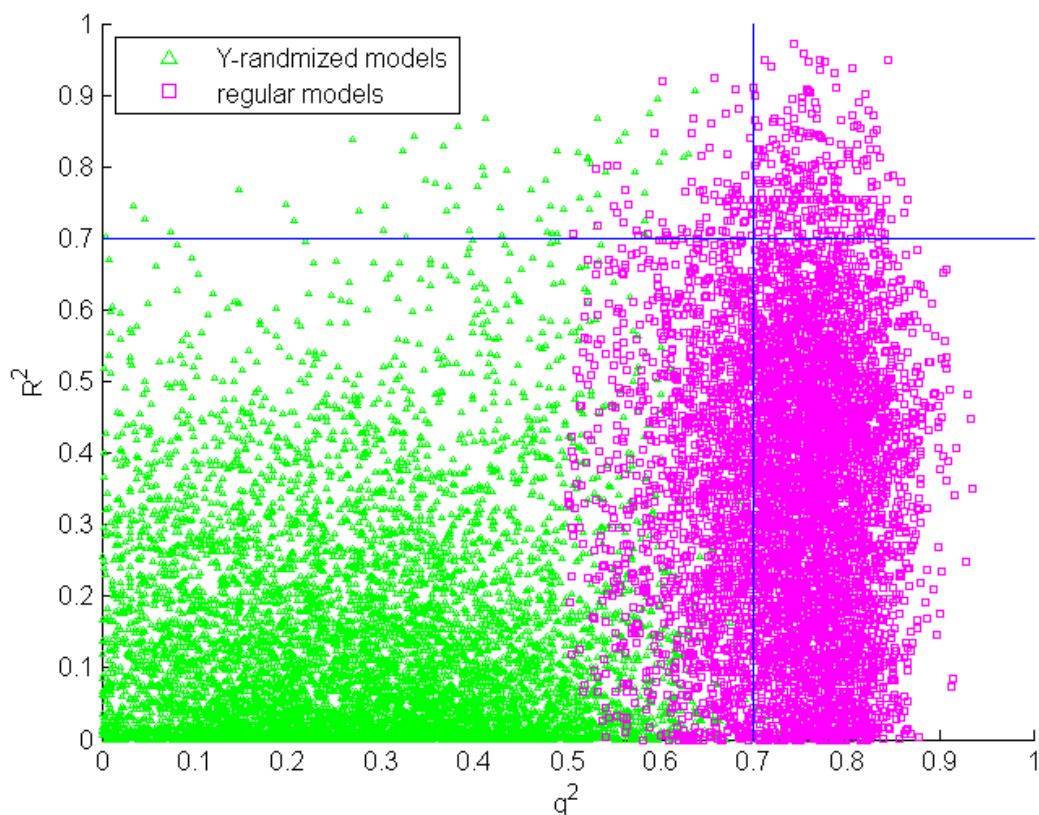


**Figure 0.1. The workflow of QSAR model building, validation and virtual screening applied to the 5-HT7 dataset and WDI database.**

Number of compounds in each step is bracketed.

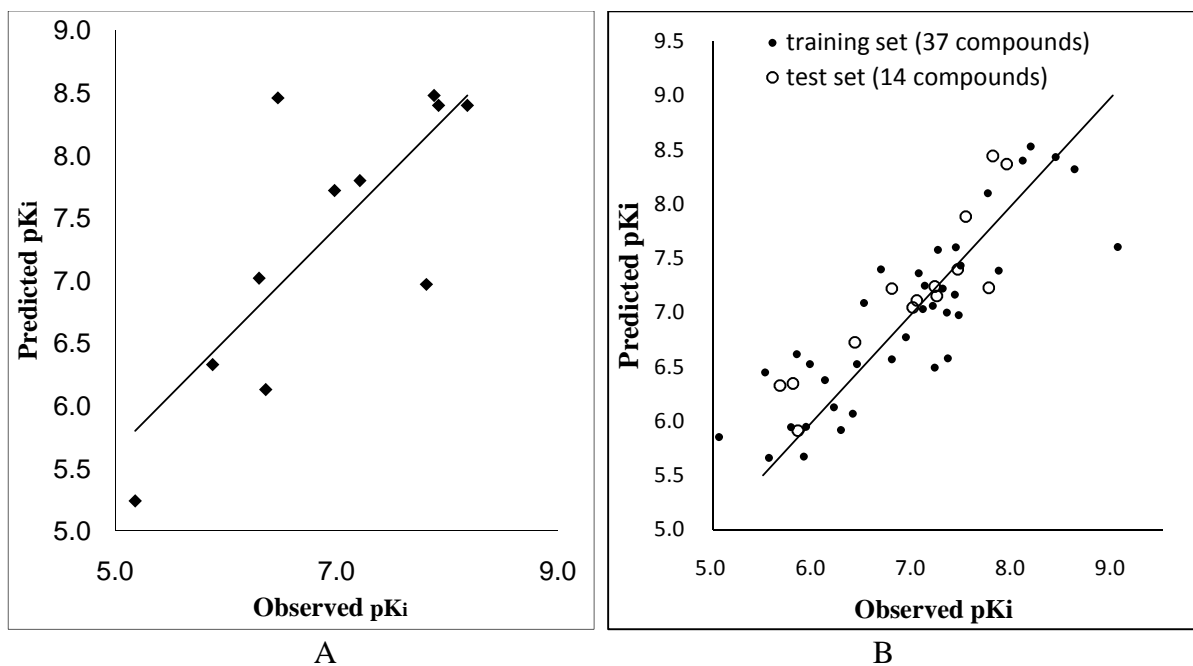


**Figure 0.2. Distribution of pKi values of each dataset extracted from PDSP and WOMBAT.**



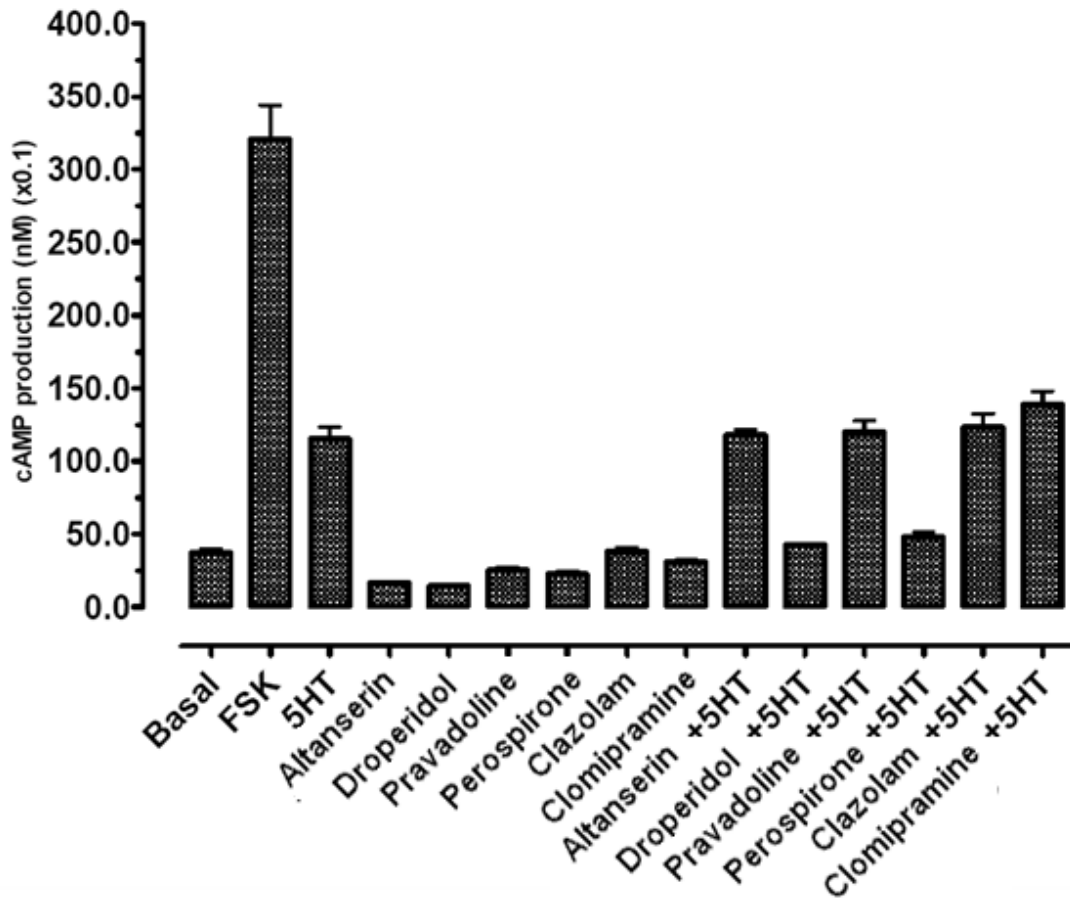
**Figure 0.3. The  $q^2$  and  $R^2$  distribution for 5-HT<sub>7</sub> kNN QSAR models built with actual activity data versus models generated with randomized data (Y-randomization).**

For each case, a total of 6090 models were generated using the cutoff of 0 for  $q^2$ . The standard one-tail hypothesis test was conducted to evaluate the statistical significance of QSAR models for the actual data set. The Z score that corresponds to the  $q^2$  value is 2.88, indicating that the level of significance  $\alpha < 0.01$  ( $Z_c = 2.33$ ).



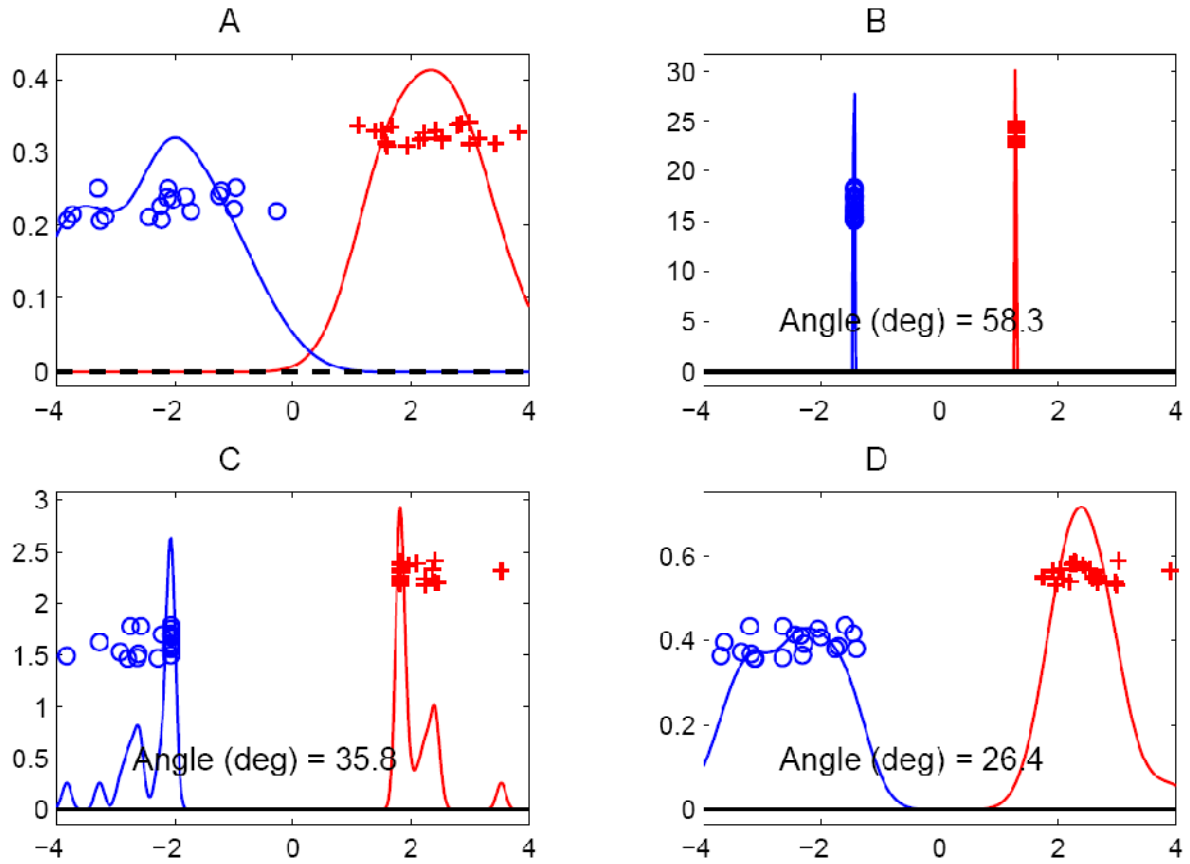
**Figure 0.4. Comparison of actual and predicted  $pK_i$  for the external evaluation set using the best kNN models.**

A: External evaluation set ( $R^2=0.61$ ,  $R_0^2=0.59$ ); B: Training and test sets ( $q^2=0.73$ ,  $R^2=0.81$ ).



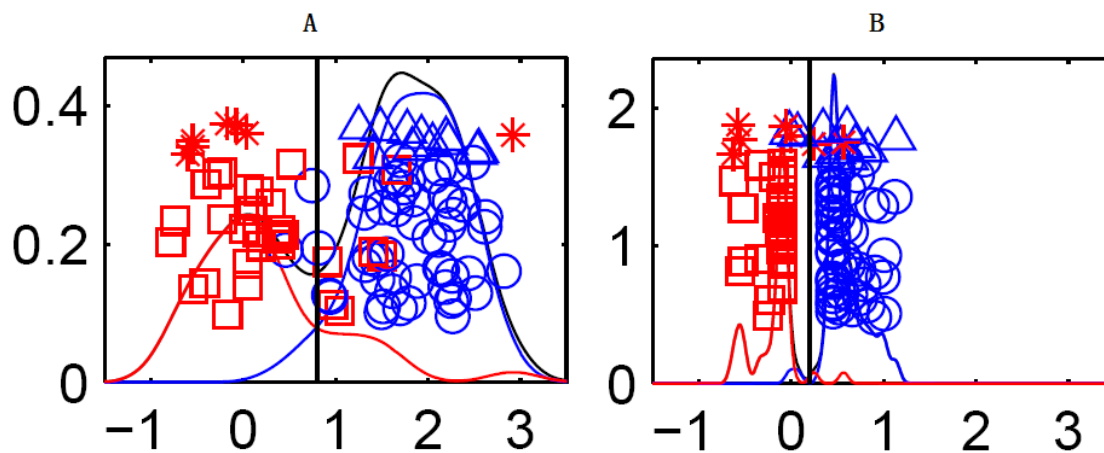
**Figure 0.5. Functional assay of hit compounds against the 5-HT<sub>7</sub> receptor.**

Droperidol and perospirone are receptor antagonists indicated by the reduced expression level of cAMP production.



**Figure 0.6. Toy Example, illustrating potential for “data piling” problem in HDLSS settings, for discrimination using SVM.**

Figure A shows projection of the data on the theoretically optimal classification direction. Figure B is projection on the MDP direction (with poor generalizability, reflected by large, 58°, angle to the optimal). Figure C is projection on the SVM direction (with some piling resulting in inferior generalizability, reflected by 36° angle). Figure D is projection on the DWD direction (with improved generalizability, reflected by a smaller, 26°, angle).

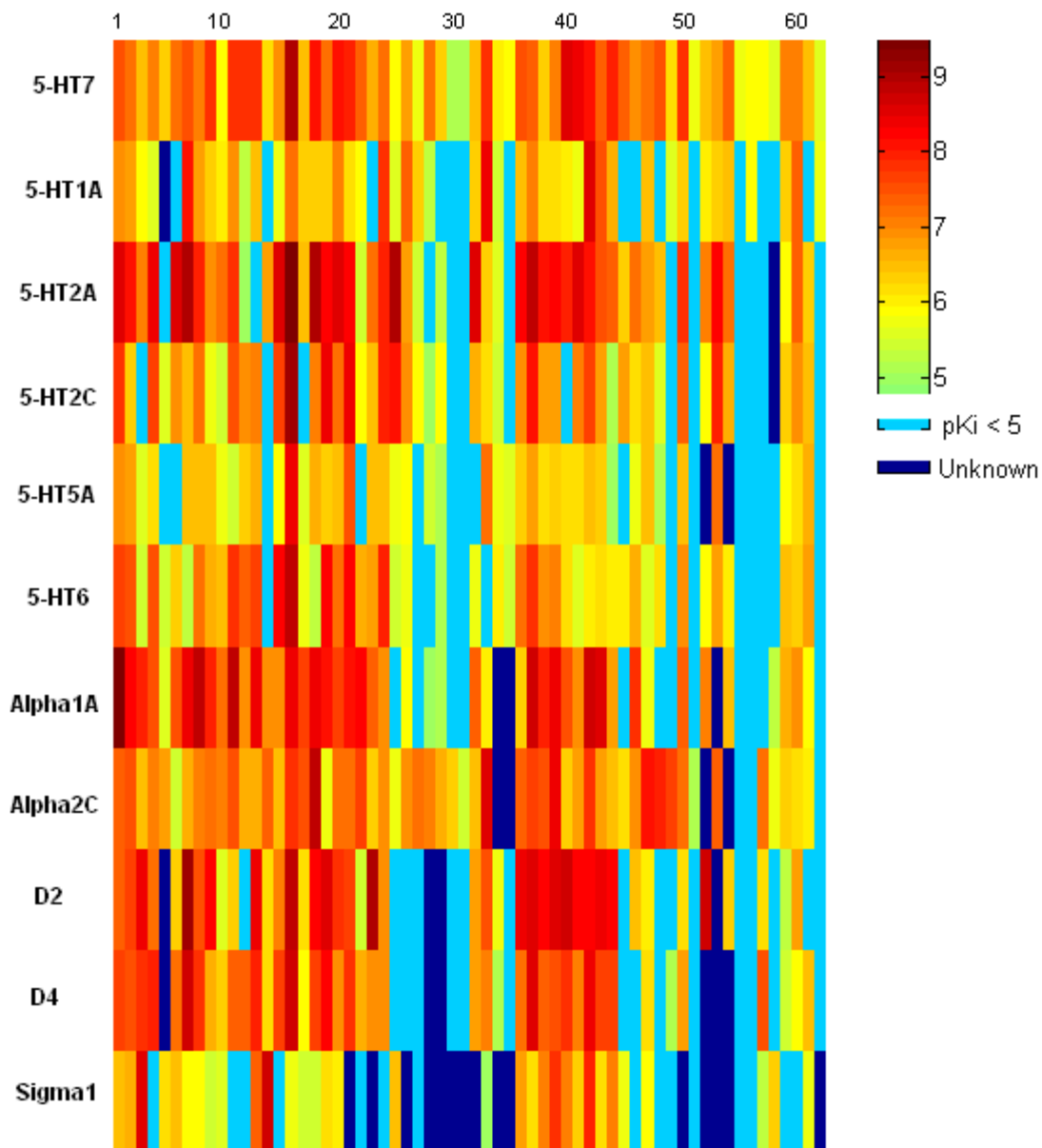


**Figure 0.7. DWD and SVM classification models based on the 5- HT<sub>7</sub> dataset.**

A: DWD classification model; B: SVM classification model.

□ non-binders in training set; ○ binders in training set; \* non-binders in test set; △ binders in test set.





**Figure 0.8. The 5-HT<sub>7</sub> dataset promiscuous binding matrix on 11 receptors including 5-HT<sub>7</sub>.**

The compounds are numbered 1 to 62 (left to right in this figure).

**Table 0.1. Frequently Used R Values and the Corresponding Critical Values of Zc for One-Tail Test.**

$$\alpha = \frac{\bar{y}}{\sigma\sqrt{2\pi}} e^{-\frac{Z_c^2}{2}} \quad \text{for } Z \geq 4$$

$\alpha$	$Z_c$
0.10	1.28
0.05	1.64
0.01	2.33
0.001	3.10

**Table 0.2. Confusion matrices for DWD and SVM predicted on 18 external compounds**

		Predicted			
		DWD		SVM	
Actual		Active	Inactive	Active	Inactive
	Active	10	0	8	3
	Inactive	1	7	2	5

**Note:** External compounds contain 11 active and 7 inactive compounds.

**Table 0.3. Consensus prediction of external test sets by *k*NN Models passed acceptance criteria based on PDSP and WOMBAT datasets.**

Dataset	Modeling set size	External set size	No. of Descriptors	No. of Models	$R^2$	k	$R_0^2$	$k_0$
PDSP	51	11	294	216	0.61	0.89	0.59	1.06
WOMBAT	53	13	377	240	0.72	0.80	0.68	0.98

**Table 0.4. Results of the experimental assays and drug information of seven hit compounds.**

Name	*Predict $K_i$ (nM)	#Predict $K_i$ (nM)	$K_i$ (nM)	Function	Therapeutic Category
<b>Altanserin</b>	3.39	6.87	143.0	N/A	Human neuroimaging
<b>Droperidol</b>	3.24	22.76	3.50	Antagonist	Butyrophenone antiemetic and antipsychotic
<b>Pravadoline</b>	9.55	14.08	3184.0	N/A	Cannabinoid analgesic
<b>Perospirone</b>	7.08	14.31	8.60	Antagonist	Atypical antipsychotic
<b>Clazolam</b>	6.46	37.76	>10000	N/A	N/A
<b>Clomipramine</b>	13.80	2.44	46.00	N/A	Tricyclic antidepressant; antiobsessional
<b>Sulazepam</b>	14.13	13.54	>10000	N/A	Sedative and anxiolytic

\*Predictions by the models based on the PDSP dataset; #Predictions by the models based on the WOMBAT dataset.

**Table 0.5. Examples of activity variation for the 5-HT<sub>7</sub> binders extracted from public PDSP K<sub>i</sub> database.**

Name	Radioligand	Species	Binding affinity (nM)	References
Ergotamine	<sup>125</sup> I-LSD	Rat	17.37	Boess FG, et al., Neuropharmacology, 1994; 33(3-4): 275-317.
	<sup>3</sup> H-5CT	Guinea Pig	138.03	Z.P.To, et al., Br J Pharmacol., 1995; 115(1): 107-16.
	<sup>3</sup> H-LSD	Human	1291	PDSP certified data.
Ketanserin	<sup>125</sup> I-LSD	Rat	206	Shen Y, et al., J Biol Chem., 1993; 268(24): 18200-4.
	<sup>3</sup> H-5CT	Human	794.33	Thomas DR, et al., Br J Pharmacol, 1998; 124(6): 1300-6.
	<sup>3</sup> H-5HT	Human	1334	Bard Ja, et al., J Biol Chem, 1993; 268(31): 23422-6.
	<sup>3</sup> H-5HT	Rat	>7500	Ruat M, et al., Proc Natl Acad Sci, 1993; 90(18): 8547-51.
	<sup>3</sup> H-5HT	Rat	>7943.28	Eglen, RM, et al., Trends Pharmacol Sci, 1997; 18(4): 104-7.
NAN-190	<sup>125</sup> I-LSD	Rat	79.1	Lovenberg TW, et al., Neuron. 1993; 11(3): 449-58.
	<sup>125</sup> I-LSD	Rat	<1000	Boess FG, et al., Neuropharmacology. 1994; 33(3-4): 275-317.
	<sup>3</sup> H-LSD	Rat	>1000	Shen Y, et al., J Biol Chem., 1993; 268(24): 18200-4.
Risperidone	<sup>3</sup> H-5HT	Human	4.3	Fernandez J, et al., J Med Chem, 2005; 48(6): 1709-12.
	<sup>3</sup> H-LSD	Human	6.6	PDSP certified data
	<sup>3</sup> H-LSD	Rat	0.93	Kongsamut S, et al., Eur J Pharmacol, 1996; 317(2-3): 417-23.

## CHAPTER 3

### DEVELOPMENT OF ALGORITHM ECONOMIC RATIO (ER) AS BOTH A COST FUNCTION AND A VALIDATION MERIT FOR CLASSIFICATION QSAR MODELS

#### 3.1. Introduction

The pharmaceutical industry must rationally design programs in consideration of huge experimental costs at multiple stages in the drug development pipeline and also substantial potential benefits of successful drug discovery[1,109]. Safety tests both in animals and in humans as part of clinical trials place severe constraints on choosing therapeutic areas for drug development, which is part of the reason that Big Pharma tends to avoid projects targeting rare, or orphan, or third world diseases[110]. Computational methods such as Quantitative Structure-Activity Relationships (QSARs) have thus become increasingly important[111].

Novel QSAR modeling methods continue to emerge focusing on one or another conventional figure of classifier metric. For example, increasing sensitivity at the expense of specificity can be achieved with the Gaussian processes regression method of Obrezanova and Segall[112]. Alternatively, Bruce et al. compared Support Vector Machine (SVM), decision tree, and several ensemble decision tree methods by the percentage of correctly classified molecules[113]. Truchon and Bayly, instead, presented an analysis of several ranking metrics used to evaluate virtual screening methods[114]. Among important metrics

in their study was the area under the receiver operating characteristic curve (ROC) and the enrichment factor (EF). To our knowledge, none of these approaches connects directly with the cost/benefit ratio of a testing strategy.

By contrast, we established a new path called “Economic Ratio” (ER) that does not seek or reward high values of conventional metrics. Rather, we strive to build and evaluate classifiers that astutely consider only a subset of compounds and then predict hits with very few false positives in proportion to true positives. Thus we discount correct prediction of negatives. In some circumstances, our approach might be closely related to some of the actual decision-making processes of drug discovery.

In summary, given the performance of a classifier constructed from historical experiments (represented by its prior probabilities), the cost of further bioactivity experiments, and the assumed benefit of a discovered hit, we propose a new procedure to decide whether or not additional testing guided by the classifier is likely to be cost-effective. As shown below, the models based on the cost/benefit ratio (applying the Economic Ratio we will define) yielded decision tree models with higher positive hit rates in some applications.

### 3.2. Materials

To compare our new procedure called Economic Ratio (ER) with current popular methods to build or choose QSAR classifiers, we employed fragment descriptors and historical experimental data. The decision tree learning technique was used because of its interpretability and its insensitivity to imbalanced datasets having few hits and many misses in the historical experiments, which is a common aspect of biological screening campaigns[115]. We also relied on accommodation of sparse descriptors by decision trees,

that is, accommodation of data sets in which most fragments are absent from most compounds.

### 3.2.1. Datasets

From five datasets (see Table 3.3) we built decision trees with branch variables selected with three target functions, namely, a new ER metric together with the traditional metrics CCR[5] and Gini impurity measure[29] (defined or introduced *infra*).

All datasets were curated following the “Trust but Verify” procedures[77]. The Drug Bank database[116,117] was employed to evaluate chemical diversity of the virtual screening hits.

**3.2.1.1. P-Glycoprotein dataset (PGP).** PGP is a member of the ABC transporter family implicated in intestinal transport, blood-brain barrier function, and multi-drug resistance of tumor cells[118,119]. We collected a dataset containing PGP substrates and non-substrates from the literature[5,120,121]. The activity classes depended on whether there are existing publications that document activity (substrate) or inactivity (non-substrate). The derived dataset was used for early calibration of our methods and also in 5-fold external cross-validation.

**3.2.1.2. Antimalarial dataset (ATM).** Antimalarial activities of 3133 compounds were tested in St. Jude Children's Research Hospital[122]. Active inhibitors were defined to be compounds that had reproducible potency at concentrations less than 2  $\mu$ M, while the remaining compounds were considered inactive. After curation, 3123 compounds remained for use in this study.

**3.2.1.3. T. pyriformis dataset (TPY).** The growth inhibition of the ciliated protozoan *T. pyriformis* is a toxicity screening tool developed and implemented by Schultz and co-



workers[123]. The *T. pyriformis* toxicity dataset used by us was compiled in our previous study [124] from several publications of the Schultz group[125–129] as well as from data available at the Tetratox database website (<http://www.vet.utk.edu/TETRATOX/>). The *T. pyriformis* toxicity of each compound was expressed as the logarithm value of 50% growth inhibitory concentration in mg/L (IGC50). For the purpose of this study, IGC50 values less than zero were considered toxic. Our final dataset included 1085 unique compounds.

**3.2.1.4. 5HT2B dataset.** The Psychoactive Drug Screening Program (PDSP) of the National Institute of Mental Health (NIMH) reported activity of roughly 800 FDA-approved drugs and drug-like molecules against 5-HT2B receptors[130]. After curation of the compounds, the final dataset consisted of 37 binders ( $pK_i \geq 5.0$ ) and 573 non-binders ( $pK_i < 5.0$ ) to 5HT2B receptors. Detailed PDSP protocols were published online (<http://pdsp.med.unc.edu/>) and in Huang et al[130]. All chemical structures were obtained from PubChem as SDF files.

**3.2.1.5. Acute toxicity estimate (ATE) dataset.** The acute toxicity data collection was described in detail elsewhere[131]. Briefly, 7385 distinct organic compounds were used with rat LD<sub>50</sub> dose expressed as mg/kg bodyweight. The endpoint selected for our study was 24 hours following a single, oral dosage. Based on the categories of acute toxicity by the Globally Harmonized System of Classification and Labeling of Chemicals[132], we defined Category 1 as toxic (LD<sub>50</sub> < 50 mg/kg) and Category 4 and Category 5 as non-toxic (LD<sub>50</sub> > 300 mg/kg).

### 3.2.2. Descriptors

**3.2.2.1. Simplex representation of molecular structures (SiRMS).** Two-dimensional (2D) simplex descriptors (tetraatomic fragments with fixed composition and topological

structure) are used for molecular structure representation. These fragment descriptors differentiate atoms based on atom type and other physical-chemical characteristics of an atom, e.g., partial charge[133], lipophilicity[134], refraction[135], and the ability of an atom to be a donor or acceptor in hydrogen bond formation[136,137]. The main advantages of SiRMS are the opportunity of analysis of molecules with pronounced structural differences and the revelation of individual molecular fragments (simplex combinations) promoting or suppressing investigated activity. SiRMS methodology was described earlier[138,139].

**3.2.2.2. XCHEM fragments.** Connected chains of atoms of variable length and branching are the basis of XCHEM descriptors. Prior to fragmentation, atomic labels are defined to include a desired subset of chemical and topological properties. Examples of applications in QSAR studies can be found elsewhere[140,141]. In the present study we have generated linear fragments from 2 to 8 atoms in length with 1 possible branch; the following features were used in the atomic labels: nuclear charge, valence, hybridization, number of hydrogen atoms, aromaticity, resonance, and membership in ring systems (3-membered cycles, 4-membered cycles, cyclic junctions, etc.).

After fragmentation of the modeling set, the fragments were filtered to require: (1) a fragment must occur in at least in five molecules of the training set; and (2) the mean activity of these host molecules must differ from the dataset's average by at least by 0.1 (z-score). The remaining fragments were sorted by occurrence frequency and the top 200 of were then used as binary descriptors (i.e., with 0/1 values for absence/presence of a fragment in a molecule). Datasets were provided along with both SiRMS and XCHEM descriptors. Although comparison of descriptor selections was not among the goals of this study, no significant difference was observed between SiRMS and XCHEM applications.

### 3.3. Theoretical basis

#### 3.3.1. Evaluation of QSAR prediction performance

Let  $P_{TP}$ ,  $P_{FP}$ ,  $P_{FN}$ , and  $P_{TN}$  denote the prior probabilities derived from application of a classifier to past experiments with designations TP (true positives), FP (false positives), FN (false negatives), and TN (true negatives). Suppose the four prior probabilities of two classifiers called Q1 and Q2 are known to be as in the following Table 3.1. We see from Table 3.1 that if Q1 is presented with 1000 novel compounds, it will, on average, correctly declare one of 13 actual hits to be a hit. Within the same 1000 compounds we expect nine incorrect declarations of hits. Applied to 1000 novel compounds, classifier Q2 will on average correctly declare two of 13 actual hits to be hits—but incorrectly will declare 45 misses to be hits. On average, Q1 finds one hit and Q2 finds two hits per 1000 compounds in which a total of 13 actual hits are included.

Conventional wisdom would evaluate Q1 and Q2 on the basis of sensitivity, specificity, or their average (called Correct Classification Ratio (CCR)[5]), all calculated from the prior probability tables. By definition,  $\text{sensitivity} = TP/(TP+FN)$  and  $\text{specificity} = TN/(TN+FP)$ [142]. Another figure of merit for a classifier is the Right Fisher Exact Test (RFET)[143], yielding the probability that guessing with a certain ratio of hit guesses to miss guesses would be at least as accurate as the classifier at hand. All these measures for Q1 and Q2 are shown in Table 3.2. They and other calculations in this paper can be conveniently observed in spreadsheets in the Supporting Information of this paper. But how might the testing strategist use these classifier performance values to decide which of Q1 and Q2 to use (if either)?

### 3.3.2. Evaluating QSAR classifiers using cost/benefit ratio

Tests are costly, but finding a hit is beneficial. Let us assume the cost per test is a constant  $C$  and the benefit per hit is a constant  $B$ . For a given classifier with known performance relative to historical data, the expected number of future tests needed to find a hit (including the test that yields the first hit) within a novel set of compounds is the reciprocal of the classifier true positive rate  $P_{TP}$ . This assumes that the set of possible compounds is large and that for practical purposes, removing a tested compound from the list does not significantly affect its size. Each false positive and also the first true positive cause a costly test. Thus the total cost of finding the first hit with cost  $C$  per test is  $C$  multiplied by the ratio of the total predicted positives to true positives. Therefore the cost/benefit ratio, determining whether or not the gamble of investing in costly tests is worthwhile, is

$$\text{Cost/benefit ratio} = \left[ \frac{C}{B} \right] \times \left( \frac{P_{TP} + P_{FP}}{P_{TP}} \right) \quad (3.1)$$

The positive likelihood ratio (also called Positive Predictive Power[144]) is by definition the ratio of true positives to all declared positives[145]; this is equivalent to the inverse of the ratio:

$$\frac{P_{TP} + P_{FP}}{P_{TP}}$$

used inside equation (3.1).

Thus from equation (3.1) we define the *economic ratio (ER)*:

$$ER = \left( \frac{P_{TP} + P_{FP}}{P_{TP}} \right) \quad (3.2)$$

Note ER can also be calculated directly using the number of TP and FP instances from the historical data that yield the prior probabilities. The importance of ER is that for any

values of cost  $C$  per test and benefit  $B$  per hit, the cost/benefit ratio of classical economics is simply  $C$  multiplied by  $ER$ , then divided by  $B$ . Among other factors, a test strategist will desire the cost/benefit ratio to be as small as possible, certainly less than one.

Suppose we can choose between classifiers  $Q1$  and  $Q2$  with past performance as described in Table 3.1. Conventional classifier analysis provides the values for sensitivity, specificity, etc. in Table 3.2. Suppose the benefit of a hit is 15 units of money and the cost of a test is one unit. Suppose also that we can neglect both the cost of populating the prior probability table (Table 3.1) and the ongoing cost of computationally applying the classifier to novel compounds. Table 3.2 in itself does not directly inform us of an economically preferable choice. However, applying equation (1) reveals the cost/benefit ratios of  $Q1$  and  $Q2$  to be 0.67 and 1.57, respectively. Thus the cost/benefit ratio of  $Q1$  is much better than that of  $Q2$  even though  $Q1$  has much lower sensitivity and about the same specificity,  $CCR$ , and  $RFET$  (Table 3.2). We see instantly that  $Q1$  is the better choice.

In summary, once a valid prior probability table has been populated, the subsequent economic value as cost/benefit ratio of the classifier is readily expressed in equation (3.1); the same is not obviously reflected in some conventional figures of classification metric.

### 3.4. Expanding $ER$ as a Target Function

It would seem reasonable to hypothesize that using  $ER$  as a target function during the growth of a decision tree would result in models with superior  $ER$  in prediction. To distinguish the same ratio used as target function and validation function of classification metric, we use  $ER_c$  to denote calculation of  $ER$  as a target function and  $ER_v$  to denote calculation of  $ER$  as a validation function. Likewise we define  $CCR_c$  and  $CCR_v$ . The validity of the hypothesis is clarified *infra*.

### 3.4.1. Decision tree construction

Specifically, our goal was to build a classifier using experimental data that maps combinations of descriptors of a compound (for instance, chemical fragments) to a binary state: active or inactive, often emphasizing highly imbalanced data sets with far more inactive compounds than active. This goal can be described as follows. We start with a large number of compounds and descriptors, such as 1000 each. We seek a small number of descriptors, such as 10, that become decision nodes in a tree. Each node uses one descriptor and a certain threshold value to partition compounds by comparing their values for that descriptor with the threshold. In case of fragment descriptors, such comparison becomes a test for absence or presence of a structural fragment in a molecule. In our study, the leaves of the tree are designated: active, inactive, or inconclusive. As will be shown *infra*, we disregard correct prediction of inactive compounds, although some leaves are populated almost entirely by inactives.

### 3.4.2. Tree construction algorithm

A tree construction algorithm selects a descriptor for a branch node that meets an optimal splitting criterion such as minimal  $ER_c$  ( $ER$  as branch target function), maximal  $CCR_c$ , or maximal decrease of Gini impurity. That is, the consequence of choosing a binary (0 or 1) descriptor is:  $P_R, N_R$  = respectively the number of positive (*active*) or negative (*inactive*) compounds *with* the fragment and placed in the right side of a branch; and  $P_L, N_L$  = respectively the number of positive (*active*) or negative (*inactive*) compounds *without* the fragment and placed in the right side of a branch.

$ER_c$  could be optimized (minimized) directly by equation (3.2) using only the rate of compounds with the fragment that are actually hits ( $P_R$ ) and the rate of compounds with the

fragment that are actually not hits ( $N_R$ ). However, we have found that such simplistic use of  $ER_c$  can create a highly skewed tree like the one shown in Figure 3.1 based on the PGP dataset. Descriptor instances of this dataset were only 7% of all possible instances, yielding a sparse QSAR matrix. To a depth of four tests, a total of 63 compounds were allocated in the active leaves. All of the 63 actually were hits. Consequently  $ER = 1.00$  (perfect) with coverage (the ratio of total predicted positives to modeling compounds) = 0.17. This tree predicted a hit by a sequence of tests for absence or presence of four fragment descriptors.

To achieve a more balanced tree with optimization of  $ER_c$ , we instead can select the descriptor that maximizes the following function  $\Delta ER$  over all branch choices.

$$\Delta ER_c = w \times \left| \frac{1}{ER_R} - \frac{1}{ER_L} \right| \quad (3.3)$$

Here  $w$  is a *weight function* defined at the branch as the smaller of two numbers:  $P_R+N_R$  and  $P_L+N_L$ , i.e.,  $w = \min (P_R+N_R, P_L+N_L)$ .  $ER_R$  is the  $ER_c$  of the right child node, and  $ER_L$  is that of the left child node. The employment of  $w$  here biases descriptor choice toward those descriptors that allow many *active* cases and also have low  $ER_c$ , not those that simply minimize  $ER_c$ . We designate the target function defined by equation (3.3) as  $WER_c$  (Weighted  $ER_c$ ), which can generate a nearly balanced tree using the same PGP data. In Figure 3.2, to a depth of four tests, a total of 122 compounds were allocated in the active leaves. 103 actually were hits. Consequently  $ER$  was 1.18 with coverage of 0.33. This tree (Figure 3.2) was more balanced than the tree built with  $ER_c$  (Figure. 3.1).

The tree grows from the root node by repeatedly applying the following steps to each node (based on CART algorithm[29]). A descriptor with small  $w$  (such as five or less, as in this paper) is considered to be *unsatisfactory*. Also unsatisfactory is any descriptor with an

undesirable target function value (such as  $ER_c$  of a leaf  $> 1.7$ , as in this paper). A flowchart for  $WER_c$  application is shown in Figure 3.3.

The Correct Classification Rate (CCR) is simply the average of sensitivity and specificity. Thus:

$$CCR_c = \frac{1}{2} \times \left( \frac{P_R}{P_R + N_R} + \frac{N_L}{N_L + P_L} \right) \quad (3.4)$$

Again addressing the PGP dataset, at each branch a descriptor can be chosen to maximize the  $CCR_c$  function, resulting in the tree in Figure 3.4. To a depth of four tests, a total of 107 compounds were predicted to be hits, of which 88 actually were hits. Consequently  $ER = 1.22$  with coverage = 0.29.

Gini impurity is a measure of the cost of misclassifying a randomly chosen compound from the set. It is used as the default target function in many popular decision tree methods such as Random Forest[30]. A detailed explanation of Gini impurity can be found in Breiman, *et al.*[29], where Gini impurity is calculated by:

$$i(t) = \sum_{i,j} C(i|j) p(i|t) p(j|t) \quad (3.5)$$

The Gini splitting criterion is choice of a descriptor with maximum decrease of impurity:

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R) \quad (3.6)$$

where in equations (3.5) and (3.6),  $i(t)$  is the impurity measure of node  $t$ , and  $C(i|j)$  is the cost of mis-classifying a compound in class  $j$  as a class  $i$  compound. Obviously,  $C(i|i)=0$ .  $p(i|t)$  is the probability of classifying a compound to class  $i$  given that it arrives at node  $t$ ;  $P_L$  and  $P_R$  are probabilities of sending a compound *with* the descriptor to the left child node  $t_L$  and *without* the descriptor to the right child node  $t_R$  respectively.



Our goal in tree construction can be thought of as a tree without excessive depth and with several leaves that are entirely or mostly hits, all with acceptable coverage of compounds. For example, designers at a certain stage of drug discovery might require at least 10% of compounds to be hits in addition to low  $ER_v$  values.

### 3.4.3. The advantage of application of ER as a target function

The analysis *supra* is most useful when ER is applied to classifiers with prior probability tables unlike textbook cases that have high  $P_{TP}$  and  $P_{TN}$  values[113]. In those textbook cases the usual criteria of classifier metric (sensitivity, specificity, CCR, accuracy, and RFET) are all nearly optimal. That is, in those cases we see no distinguished conventional criterion that can be directly and confidently used in a management decision to deploy resources. Coverage is universal and  $ER_v$  is so close to 1 that only the cost/benefit ratio C/B matters. By contrast, the cost/benefit ratio approach shines when: (1) a high rate of actual hits among predicted hits is absolutely essential; (2) the frequency of actual hits among all compounds is known to be low; and (3) low coverage (proportion of compounds that are predicted to be hits) is acceptable. These factors are determined by the pipeline stage, the current capacity to execute experiments, and so on. Our goal is a classifier that cautiously chooses which compounds can be predicted to be hits. This policy can be acceptable, for example, when a company faces test protocols that are expensive, possesses a large compound file (such as one million compounds), and seeks a much smaller number of hits (such as 100) because each actual hit automatically has excellent profit potential. Importantly, some data profiles seem to preclude construction of classifiers with high coverage, specifically data profiles in which most compounds include few or very few instances of

nonzero descriptor values, that is, when the descriptor matrix is sparse (almost entirely filled with 0 entries).

## 3.5. Results and Discussion

### 3.5.1. Model construction

As an example a classifier with low  $P_{TP+P_{FP}}$  rate, we considered binary data from a testing program for PGP (p-glycoprotein) inhibitors[5]. The model employed 599 descriptors and 371 compounds. In the 599-by-371 matrix of descriptor instances, only 7% of entries had nonzero values (presence of descriptor). The complete matrix is in the Supporting Information. We constructed decision trees in which each branch descriptor was chosen to minimize  $ER_c$ , or maximize  $WER_c$  and  $CCR_c$  (Figures. 3.1, 3.2, 3.4).

The point of these figures is that sparse QSAR tables can yield useful predictions provided the economic ratio is deemed important and a low coverage is deemed acceptable.

### 3.5.2. 5-fold external cross-validation

Five datasets were employed to test the predictive power of the decision tree growing algorithm described in section 2.4.2. Target functions included  $WER_c$ ,  $CCR_c$ , and best decrease of Gini impurity. The external test analysis followed a 5-fold cross-validation procedure, which means 20% of compounds from each dataset were extracted randomly for five times to compose an external validation set, i.e., a subset which is not involved in model construction, thus building a different model each time on the remaining 80% of compounds.

As shown in Table 3.3, four datasets have about 5% of compounds active, but the PGP has over 50% of compounds active. Decision trees were built on respective modeling sets for the five datasets. The minimum satisfactory size of a node was set at five compounds. Partitions were defined to be satisfactory (for continuing tree building) as follows: (1) TP

divided by (TP+FP) is greater than 0.6 for active leaves; and (2) TN divided by (TN+FN) is greater than 0.6 for inactive leaves. Otherwise, the leaf was defined to be inconclusive and was viewed as terminal.

Table 3.4 shows the performance of predictions for 5-fold external cross-validation of trees built with the three target functions. Here if a leaf (node) was inconclusive, then compounds in that leaf were omitted for the statistical calculations. The external  $ER_v$  values yielded by optimizing  $WER_c$  were the lowest for ATM, ATE and TPY datasets containing only about 5% of compounds as actives; however, this is not true for the 5HT2B dataset which also contains 5% of the compounds as actives, but has a much lower absolute number of active compounds. Furthermore,  $WER_c$  nearly resulted in the worst  $ER_v$  values for the PGP dataset with more than half of compounds as actives.

The primary hit rate of high throughput screening (HTS) usually does not exceed 5% [146,147], so  $WER_c$  is indeed worth considering in some cases. However, in the case of the 5HT2B dataset, the coverage of external prediction was very low. It points to the concern that too small a number of actives (37 in total for 5HT2B dataset) may produce inferior  $WER_c$  models. Indeed, results in Table 3.4 showed that models built by  $WER_c$  were not better than those optimally decreasing Gini impurity.

In order to avoid over-training problems, our decision trees were usually pruned using procedure implemented CART algorithm[29] to exclude deep leaves. However, this led to excessively small coverage for constructed trees. In case of external prediction for the ATM dataset, there were only ~20% positives captured by each target function.

### 3.5.3. Chemical similarity analysis

We hypothesized that decision tree models would find diverse hits, not just clusters of similar structures. We thus used the DrugBank database to virtually screen the PGP dataset whose active leaves tend to contain enough hits for diversity analysis. We found the screened hits were quite dissimilar to the modeling set compounds. The 69 compounds were captured by the biggest active leaf in the WER<sub>c</sub>-based tree model of PGP dataset. Figure 3.5 presents the distribution of Tanimoto coefficients (Tc) in the MACCS key fingerprints[148,149] descriptor space for the modeling set and for the hit compounds. The noticeable shift to the left indicates high divergence of the hits from the modeling compounds. The diversity among the hits alone was also superior to the diversity among the modeling set compounds, even though these hits came from only one active leaf. The reason for this is that, when evaluated by the full descriptors, hits identified by a few fragment descriptors may contain structural features not held by the modeling set. This analysis demonstrates that the decision tree method is to some extent capable of finding structurally diverse hits which are substantially different from the existing structures in the modeling set.

### 3.5.4. Discussion

Competent construction of a QSAR model requires knowledge of numerous pitfalls as explained by Dearden et al.[150]. Among other errors, they pointed to data errors including failure to consider heterogeneity of data, failure to use sufficiently varied data, redundancy of data, and limited domains of data that imply limited coverage. Some descriptor types might be mechanistically impossible to use or interpret. Model construction might fail to prevent over-fitting, fail to use statistics correctly, or fail to include adequate validation. Regarding statistical errors, Golbraikh et al.[38] warned against incorrect interpretation of  $q^2$  and

emphasized the need to employ an external validation set. Thus construction of a QSAR model is hardly a simple procedure[39].

In a recent publication, Swamidass, et al.[151] proposed a strategy of using models based on the economic supply-demand curve to enable an HTS test strategy. The optimal number of hits submitted for testing by confirmation tests was thus decided by economic rather than probabilistic analysis. Their work included consideration of economic factors of drug developers, and is consistent with our purpose to employ cost/benefit ratio and economic ratio in virtual screening.

### 3.6. Conclusions

We have devised a novel criterion for valuation of classifiers that in some circumstances captures the economic worth of bioactivity predictions, emphasizing prior probabilities of true positives ( $P_{TP}$ ) versus false positives ( $P_{FP}$ ). Our criterion, called "economic ratio", showed little relationship with conventional figures of classifier metric such as sensitivity, specificity, accuracy, etc. We also employed ER in weighted form as target function  $WER_c$  that seeks to avoid child nodes with very few compounds. Trees can be constructed by choosing each branch descriptor according to best values of  $WER_c$  or conventional target functions. We observed good  $ER_v$  values on external test sets in some cases as well as good diversity. However, successful application of our classifier methods might be restricted to highly skewed datasets containing only few active compounds.

Tables and Figures

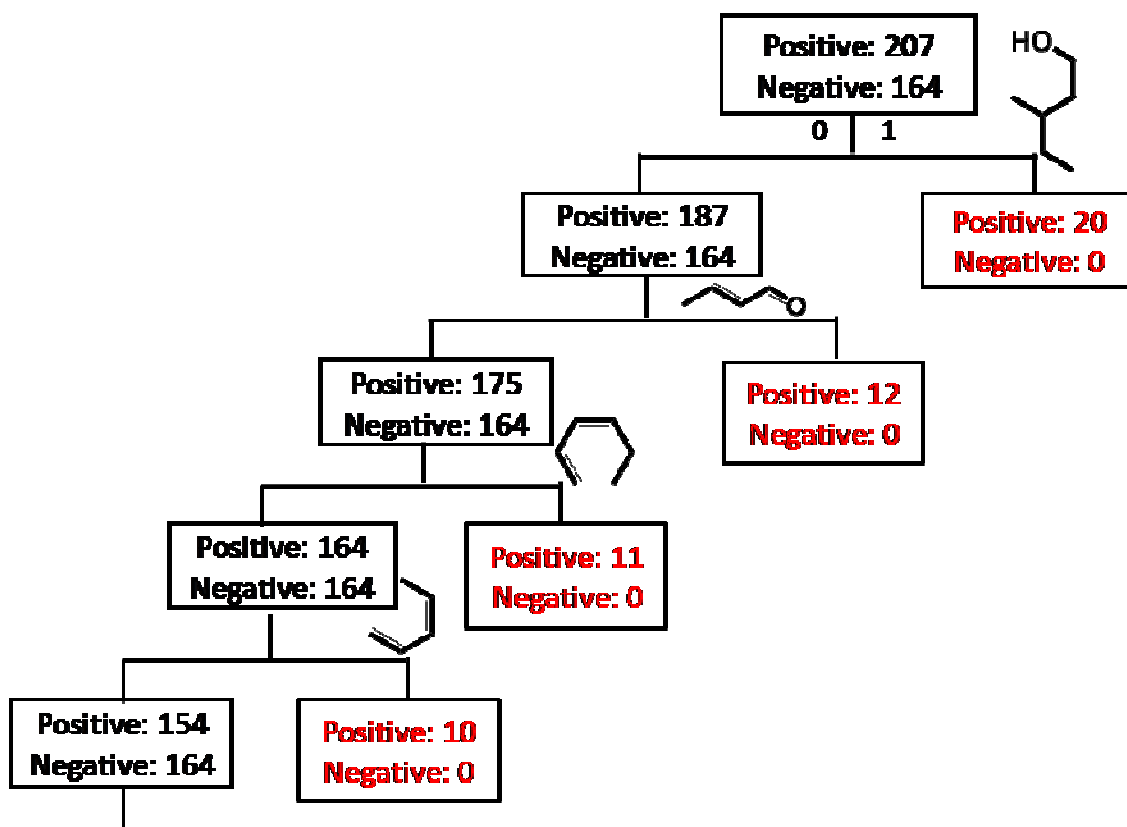


Figure 0.1. Decision tree for PGP dataset grown to minimize  $ER_c$ .

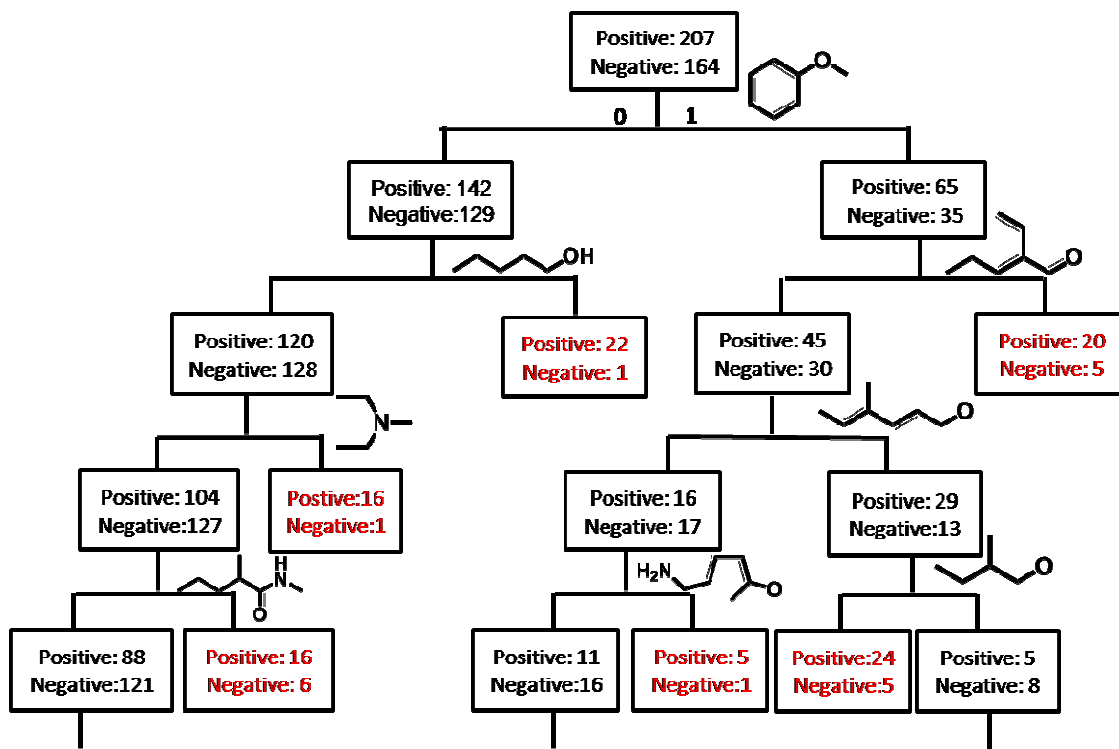
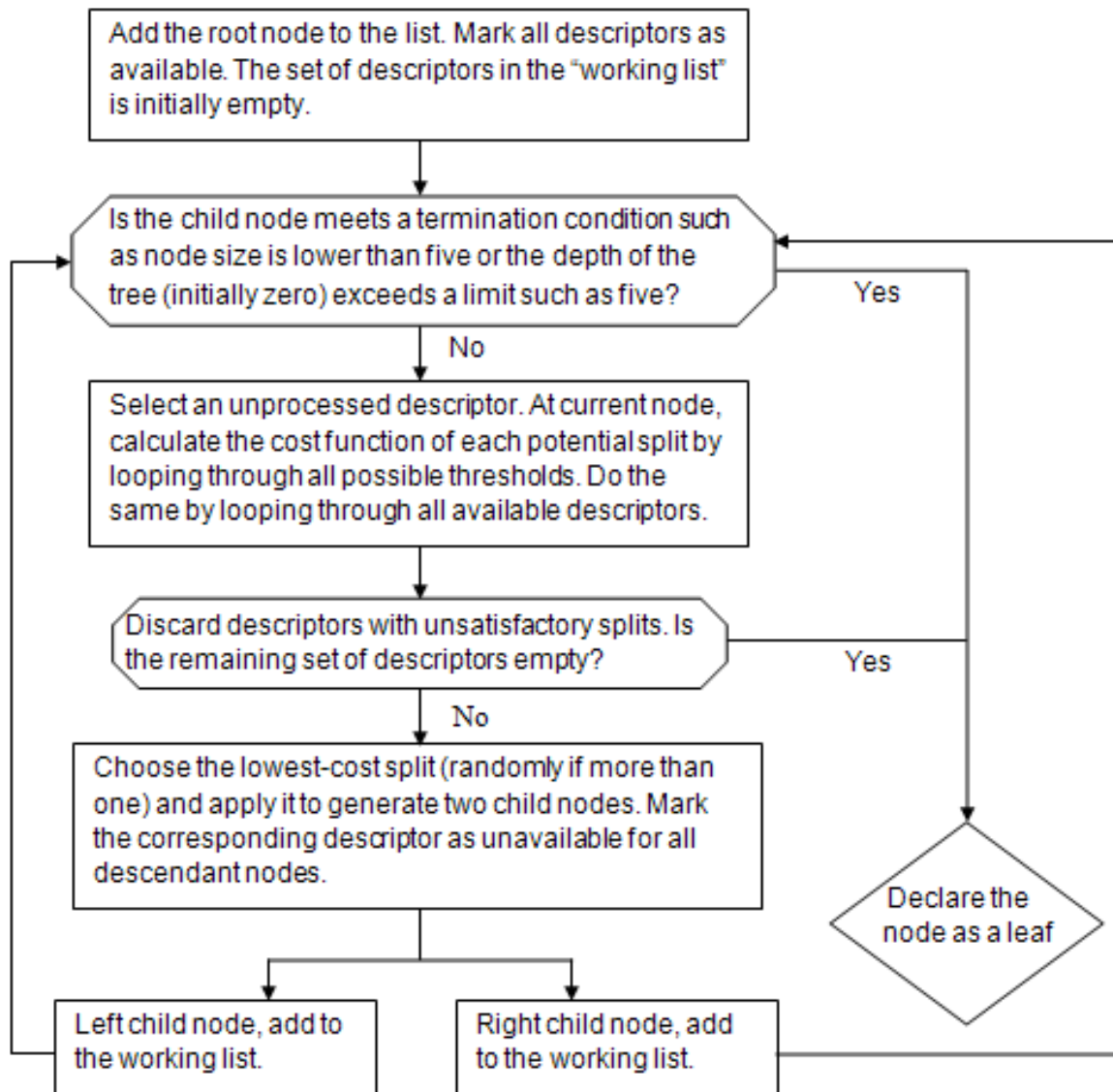


Figure 0.2. Decision tree for PGP dataset optimized to maximize WER.



**Figure 0.3. Main steps of ER-based Decision Tree Algorithm.**



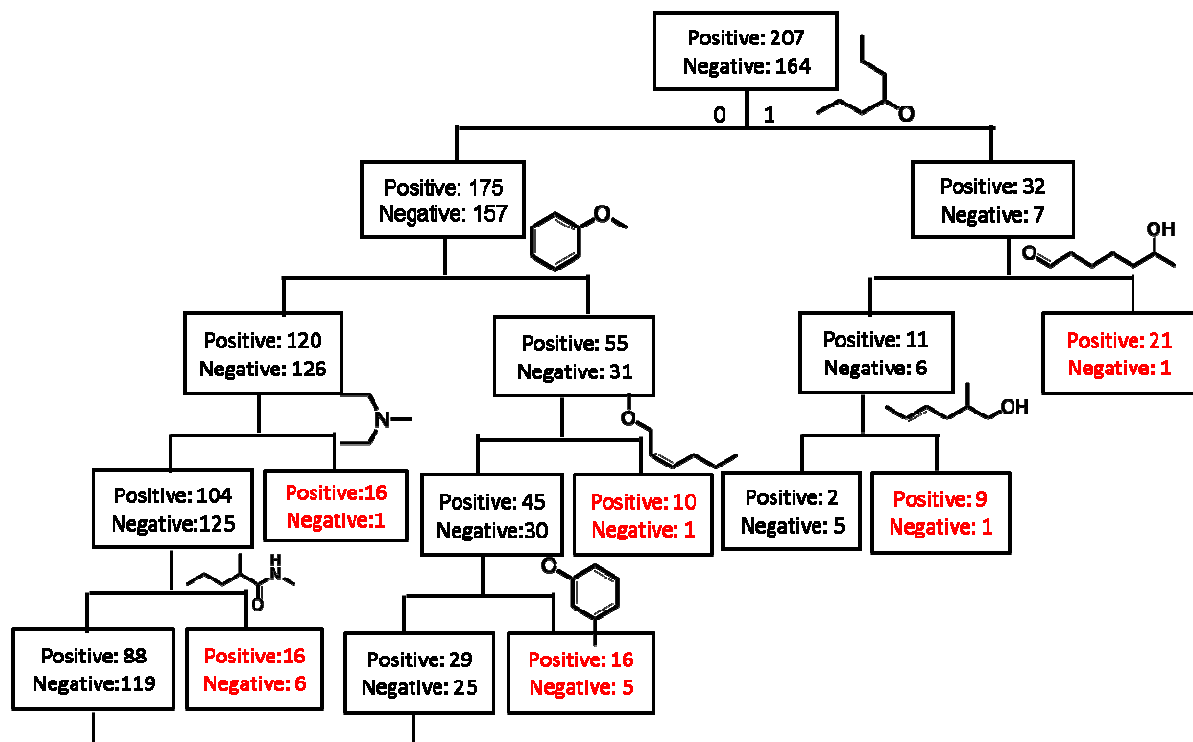
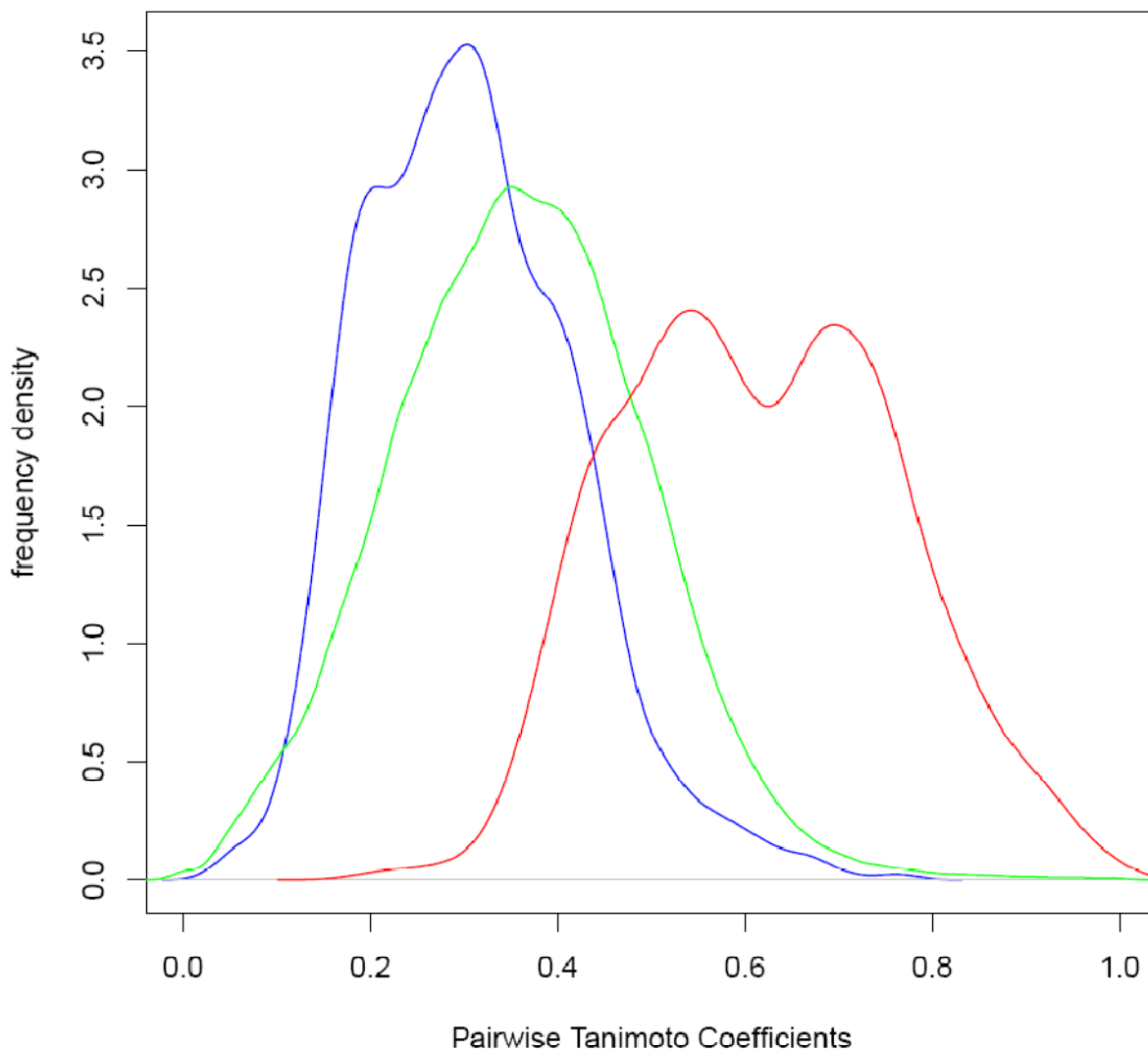


Figure 0.4. Decision tree for PGP dataset optimized to maximize weighted CCR.



**Figure 0.5. Distribution of Tanimoto coefficients ( $T_c$ ) for all pairs of compounds in PGP modeling set and virtual screening hits.**

Blue: pairs of hits and modeling set compounds; green: pairs of hits; red: pairs of compounds within the modeling set.

**Table 0.1. Examples of prior probability tables of two hypothetical QSAR classifiers called Q1 and Q2.**

The prior probabilities are  $P_{TP}$  = true positive rate,  $P_{FP}$  = false positive rate,  $P_{FN}$  = false negative rate, and  $P_{TN}$  = true negative rate.

	$P_{TP}$	$P_{FP}$	$T_{FN}$	$P_{TN}$
Q1	0.001	0.009	0.012	0.978
Q2	0.002	0.045	0.011	0.942

**Table 0.2. Conventional performance measurements of the two QSAR models derived from the prior probabilities.**

	Sensitivity	Specificity	ER	CCR	RFET
Q1	0.08	0.99	10	0.53	0.12
Q2	0.15	0.95	22.5	0.55	0.12

**Table 0.3. Datasets used for decision tree model construction and evaluation.**

	Descriptor		Dataset	
	Type	No.	Active	Inactive
<b>PGP</b>	<b>Simplex</b>	1128	258	204
<b>ATE</b>	<b>Simplex</b>	6631	284	6347
<b>ATM</b>	<b>Simplex</b>	2055	158	2975
<b>TYP</b>	<b>Simplex</b>	1404	58	1027
<b>5HT2B</b>	<b>Simplex</b>	1156	37	573

**Table 0.4. Performance statistics of 5-fold external cross-validation.**

<b>Dataset</b>	<b>Target Function</b>	<b>TP</b>	<b>FP</b>	<b>FN</b>	<b>TN</b>	<b>Sens.</b>	<b>Spec.</b>	<b>CCR<sub>v</sub></b>	<b>ER<sub>v</sub></b>
<b>PGP</b>	<b>CCR<sub>c</sub></b>	168	62	88	141	0.66	0.69	0.68	1.37
	<b>WER<sub>c</sub></b>	174	80	82	123	0.68	0.61	0.64	1.46
	<b>Gini</b>	167	60	89	143	0.65	0.70	0.68	1.36
<b>ATE</b>	<b>CCR<sub>c</sub></b>	144	41	6310	140	0.50	0.99	0.75	1.31
	<b>WER<sub>c</sub></b>	144	44	6305	142	0.51	0.99	0.75	1.28
	<b>Gini</b>	154	48	6301	132	0.54	0.99	0.77	1.31
<b>ATM</b>	<b>CCR<sub>c</sub></b>	37	103	117	2870	0.24	0.97	0.60	3.78
	<b>WER<sub>c</sub></b>	26	43	132	2922	0.16	0.99	0.58	2.65
	<b>Gini</b>	27	57	127	2916	0.18	0.98	0.58	3.11
<b>TPY</b>	<b>CCR<sub>c</sub></b>	18	18	40	1009	0.31	0.98	0.65	2.00
	<b>WER<sub>c</sub></b>	19	9	39	1018	0.33	0.99	0.66	1.47
	<b>Gini</b>	27	23	31	1004	0.47	0.98	0.72	1.85
<b>5HT2B</b>	<b>CCR<sub>c</sub></b>	7	21	30	552	0.19	0.96	0.58	4.00
	<b>WER<sub>c</sub></b>	7	15	30	558	0.19	0.97	0.58	3.14
	<b>Gini</b>	11	14	26	559	0.30	0.98	0.64	2.27

Notes:

CCR: Correct Classification Rate;

WER: Weighted Economic Ratio;

Sens.: sensitivity;

Spec.: Specificity.

## CHAPTER 4

### PILOT STUDY FOR THE QSAROME PROJECT: 5-HT<sub>1A</sub>

#### 4.1. Introduction

In the past decade, most major pharmaceutical and biotech companies have experienced a rapid growth of publicly available databases caused by the innovative technologies that allow rapid synthesis and high throughput screening. However, such technologies are limited to the academics, and thus the academic research community has been lack of the freedom to quickly screen large scale databases for novel drug candidates against new pathways or targets. The situation has been gradually changed since the adoption of several critical programs initiated by National Institute of Health (NIH) and a few other academic institutions. For example, the NIH Molecular Libraries Roadmap Initiatives launched on the national Molecular Library Screening Centers Network (MLSCN)[152] to encourage the collaboration of academics to form a public high-throughput biological screening resource. The NIMH Psychoactive Drug Screening Program[3] is another abundant resource to quickly assay and provide annotated information of large scale compounds with their pharmacological and functional activity at Central Neural System (CNS) receptors, channels, and transporters.

On the other hand, the explosive growth of publicly available large databases concurs with the plague of false annotations recorded in the databases[4]. While the percentage of

erroneous recordings in public databases is hard to estimate (error rates of the commercial databases, however, range from 0.1 to 3.4%, according to a recent study[153]), it is indeed a severe problem which would deteriorate the quality of QSAR modeling so badly that sometimes no valid QSAR models could be built from biological meaningful datasets[153]. Since the data generated by HTS experiments keep growing, it is more important than ever to address the issue of data quality that inherently affects the quality of models.

Unfortunately, QSAR modelers often consider data curation as a trivial step and are lack of due diligence to clean the data thoroughly. For example, one study from NCI AIDS Antiviral Screen included more than 40,000 chemicals with associated activity in its library. Even a brief examination of these records by our lab revealed that about 10% of them should be curated or even removed before launching any QSAR modeling[77]. Yet this database was frequently used by others without thorough descriptions of data curation (more than 57 citations by 2010). Another example is the *Tetrahymena pyriformis* aquatic toxicity data set which comprises 1093 compounds. Exactly this data set was used by at least eight research teams for QSAR modeling[124,137,154] including the organizers of CADAster toxicity challenge (<http://www.cadaster.eu>). Later it appeared that amongst 1093 compounds there were six pairs of duplicated with toxicity range up to one logarithmic unit, which was caused by the simultaneous presence of organic acids and their salts in the dataset [77]. Contrary to any above-mentioned cases, we do believe that chemical record curation should be viewed as a separate and critical component of any cheminformatics research. In this chapter, we developed logical basic steps for data curation which could create a foundation for the subsequent QSARome project, which is explained thoroughly in the next chapter. However, we should point out that this is not a universal approach for any dataset; instead, it is the least

curation workflow that should be exercised. The protocols used here have proved to be efficient, and have been successfully tested on several individual datasets. It may not work on some difficult or ambiguous cases appearing in other datasets, which requires additional special treatments. Moreover, personal participation in the process, i.e., manual inspection and curation at the last stage of curation process, is absolutely necessary, because some errors that are obvious for human eyes will be missed by a computer.

As the source of GPCRs ligands is promiscuous in several ways, e.g., species used to test the affinity (human or rat) and annotations for activity ( $K_i$  or  $IC_{50}$ ) are different, we intended to apply all data curation techniques to one receptor to verify which conditions will provide us predictive QSAR models. The example we choose from the target list is the 5-HT<sub>1A</sub> receptor. It is the most widespread receptor among the seven subtypes of serotonin (a.k.a. as 5-hydroxytryptamine or 5-HT) receptors[155]. As one of the critical protein targets that mediate inhibitory neurotransmission, the activation of 5-HT<sub>1A</sub> receptor will help to relieve some CNS disorders such as anxiety, depression, schizophrenia and Parkinson's disease[156–160]. Thus, discovery of new bioactive compounds targeting the 5-HT<sub>1A</sub> receptor (5-HT<sub>1A</sub> agonists) is attractive enough as an independent project. Here we not only made an individual effort to identify 5-HT<sub>1A</sub> binders, but also tried to establish a standardized routine for data curation and QSAR modeling of remaining GPCR targets (See Chapter 5 for details).

## 4.2. Materials and Methods

### 4.2.1. Dataset

The 5-HT<sub>1A</sub> dataset was extracted from ChEMBL[2] and PDSP[3], with the information provided at Table 4.1. Only compounds with  $K_i$  values were collected, and then



grouped based on the sources of tested species, indicating that most of the compounds were tested on human and rat tissues, with a few tested on other kinds of species such as mouse, pig, and even rabbit (Table 4.1). The  $K_i$  values were expressed as mol/L, which were converted to negative  $\log[1/(\text{mol/L})]$  values ( $pK_i$ ) according to standard QSAR practices.

#### 4.2.2. Chemical Data Curation

The curation procedures followed the protocol established by our lab. Figure 4.1 briefly describe the main steps for chemical curation. The entries with no recorded structural smiles (e.g., 114 entries for 5-HT<sub>1A</sub> dataset from PDSP) and compounds tested on tissues other than humans and rats (e.g., mice, pigs, rabbits, undefined, etc., as listed in Table 4.1 for details) were removed at the very beginning. Then the curation workflow could be divided into following steps:

- 1) Removal of inorganics, mixtures, and organometallics (compounds containing bonds between carbon and a metal). The reason to remove inorganics is because most molecular descriptors can only be computed for organic molecules. Consequently, most cheminformatics and QSAR software does not process inorganic molecules. For many mixtures, it is common that only the largest fragment possesses the experimentally determined biological activity. So retaining the component with the highest molecular weight or largest number of atoms is widely used, including the given study. However, such simple treatment can only work on mixtures formed by a relatively large organic molecule and small inorganic molecules (e.g., hydrochlorides, hydrates, etc.). In the case of two equally similar components, we must use biological expertise to determine which one to be retained or completely delete the entire record. In addition, for compounds containing metal atoms or rare elements, it would be better

to remove the entire record. Dragon software only computes descriptors for molecules containing the following 38 atoms: H, B, C, N, O, F, Al, Si, P, S, Cl, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Mo, Ag, Cd, Ln, Sn, Sb, Te, I, Gd, Pt, Au, Hg, Tl, Pb, Bi[161]. As a result, molecules containing other common atoms such as Na and Mg will be rejected by Dragon software. However, we only retain the compounds containing H, C, O, N, S, P, F, I, Br, and Cl, because they represent the majority of compounds for QSAR modeling and drug discovery.

2) Structural conversion and cleaning.

This step refers to the conversion of SMILES strings recorded in the databases into 2D molecular graphs. It is better to compute descriptors from the 2D level than from SMILES due to the presence of wrong initial SMILES strings in the database. The wrong SMILES strings may be caused by manual drawing errors or conversion errors from 2D structures to SMILES. ChemAxon Standardizer was used in this study to accomplish this task because of its fast and simplified graphical tools, although other software like MOE, Sybyl, and OpenBabel can also be used. In the case of compounds present as salts in the records, we removed the metal counterions first and neutralized the remaining carbocations (or carbanions).

3) Normalization of specific chemotypes.

The same function groups could be represented by different structural patterns in the same dataset. One of the most common examples is the nitro group which can be represented by five different patterns as shown in the Figure 4.2.

In such cases, we used ChemAxon Standardizer to convert all possible forms into one consistent chemotype so that two identical compounds represented by different

patterns would be recognized as duplicates by conventional similarity metrics. Without such conversion these compounds will have differences in calculated descriptors. The conversions in this study included classical settings of neutralization, aromatization, 2D cleaning, and tautomerization (explained later).

4) Treatment of tautomeric forms.

Both ChEMBL and PDSP have compounds existing in several tautomeric forms. The most common ones are the keto-enolic tautomers. It was revealed by Young et al. that choosing different tautomeric forms would result in different prediction performance of QSAR models[153]. Ideally speaking, choosing the one based upon the mechanism of action is recommended. In this study, it is time-consuming and unrealistic to identify such information for each compound, so we chose the most stable tautomeric form at the most possible chemical system (e.g., neutral pH).

5) Analysis/removal of duplicates.

This is the most important issue to compile a clean dataset for QSAR modeling, since it is often observed that even published QSAR models were based on datasets in the presence of structural duplicates. Presence of duplicates is very common for datasets extracted directly from even the same database. However, identification of duplicates is risky if only use SMILES strings because often they are not recorded as canonical SMILES. In our case, all SMILES strings were treated by aforementioned steps, and then duplicates were identified based on the 2D molecular graphs by HiT QSAR software[138]. However, to ensure higher level of confidence we also confirmed the resulting datasets with ISIDA/Duplicates software, which is complementary to HiT QSAR but based on different algorithm of structural comparison.

#### 4.2.3. Biological Data Curation

Once the duplicates are identified, analysis of their biological activities is mandatory. In this dataset, we found that majority of duplicates (often more than two structures are identical) have remarkably different  $pK_i$  values. As a consequence,  $pK_i$  values for the duplicated structures were determined by following procedures:

- 1) Calculate the standard deviations (SD) of  $pK_i$  values for the identical structures. If SD was greater than 0.5, examine the underlying  $pK_i$  values to confirm which scenarios it belonged to: a) one of the  $pK_i$  values was significantly different from the others. In this case, this value should be excluded and the retained  $pK_i$  values would be averaged to afford a determined  $pK_i$  value for the identical structures. b) The associated  $pK_i$  values varied from each other and the range of the  $pK_i$  values were greater than 0.5 log units. In this case, there were no significant outliers and the reason to cause the variations was hard to identify (manual errors when the database was built, tested by different laboratories under different experimental conditions, variations in the protocol, etc.). We excluded all duplicates under such scenario.
- 2) In some cases the duplicates were introduced by the aforementioned curation steps, for example, the removal of counterions in salts. Sometimes, a neutral compounds and its salt might be different in experimental properties. We would exclude both records if it belonged to such case. Otherwise, use the averaged  $pK_i$  value if the properties showed no significant differences.
- 3) For classification datasets (binders and non-binders determined by the threshold of  $pK_i$  values), we examined the excluded duplicates due to high SDs again. If all  $pK_i$  values

associated with the identical structures were higher (or lower) than the threshold, then one of the identical structures was added back to the class of binders (or non-binders).

#### 4.2.4. Generation of Descriptors

Three different types of descriptors were used in this pilot study based on the curated SD file of the structures. All kinds of descriptors were processed as follows. First, we removed all descriptors that had zero values or zero variance for all compounds. Furthermore, redundant descriptors were identified by analyzing correlation coefficients between all pairs of descriptors; if the correlation coefficient between two descriptor types for all modeling set compounds was higher than 0.95, one of them was randomly chosen and removed. As a result, the final numbers of descriptors depended on the size of each data set. Finally, the descriptors were range scaled to 0~1 based on the maximal and minimal values of each descriptor type in whole dataset.

- 1) Dragon descriptors. These descriptors were calculated by the Dragon software v5.5[162]. Only 0D, 1D, and 2D descriptors were considered in this study. The initial number of Dragon chemical descriptors was as high as 2442, but reduced significantly by aforementioned process.
- 2) 2D MOE descriptors that include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors[82,83,88,163–166].
- 3) Simplex fragment descriptors. Generation of Simplex representation of molecular structure (SiRMS) was the same as what we used in Chapter 3. To stress the details, 2D simplex descriptors (tetraatomic fragments with fixed composition and topological

structure) are used for molecular structure representation. These fragment descriptors differentiate atoms based on atom type and other physical-chemical characteristics of an atom, such as partial charge[133], lipophilicity[134], refraction[135], and the ability of an atom to be a donor or acceptor in hydrogen bond formation[136,137]. The main advantages of SiRMS are the opportunity of analysis of molecules with pronounced structural differences and the revelation of individual molecular fragments (simplex combinations) promoting or suppressing investigated activity.

#### 4.2.5. *k*NN Modeling Algorithm

Initially, a subset of descriptors is selected randomly. The model developed with this set of descriptors is validated by leave-one-out (LOO) cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the weighted average activity of its *k* (*k*= 1 to 9) nearest neighbors in the subspace of descriptors (Equation 4.1). The weights of neighbors, *w<sub>i</sub>*, decrease with distance, thus closer neighbors contribute to the calculated activity more:

$$y_{pred} = \frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i}; w_i = \exp(-d_i) \dots \dots \dots 4.1$$

Here *y<sub>pred</sub>* is predicted activity; *d<sub>i</sub>*, *w<sub>i</sub>* and *y<sub>i</sub>* are Euclidean distance, weight and actual activity for the nearest neighbor *i*, respectively. A genetic algorithm was used to optimize the variable selection.

#### 4.2.6. Support Vector Machines (SVM) Modeling Method

The SVM algorithm employed by this study was originally implemented in the open-source LibSVM package (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). We used a windows version of this algorithm named WinSVM (developed by our group at UNC, available upon

request)[167] which provide users with a convenient graphical interface to prepare input data, to split compounds into training, test and validation sets, to set up parameters for SVM grid calculations (iterative and simultaneous grid optimization of SVM parameters), to launch and follow calculation progress in a powerful graphical interface, to select models with the best prediction performances on both training and internal test sets, and then apply them for the external test set as an ensemble consensus model with its defined applicability domain. The program also allows one to visualize molecular structures and various plots, making the use of SVM easier and more appropriate for QSAR modeling at Windows environment, in order to obtain robust and predictive models and apply them to virtual libraries as well.

The core of SVM algorithm[33] is to search for the optimal hyper-plane separating the two classes in the descriptor space by maximizing the margin between the closest points of the two classes as shown by following equation:

$$\min \frac{1}{2} W^T W + C \sum_{i=1}^l \xi_i \dots \dots \dots (4.2),$$

subject to

$$y_i (W^T \Phi(X_i + b)) \geq 1 - \xi_i \dots \dots \dots (4.3)$$

where C is the penalty parameter and  $\xi_i$  is the slack parameter. To make the dataset linearly separable, the data points are projected to a higher dimensional space by Radial Basis Kernel Function (RBF):

$$K(x, x_j) \equiv \Phi(X_i)^T \Phi(X_j) = \exp(-\gamma \|x_i - x_j\|), \gamma > 0 \dots \dots (4.4)$$

where  $\gamma$  is the kernel parameter. We used the in-house software WinSVM to determine the optimal parameters C. The search range of C and  $\gamma$  were from -3 to 16 and -15 to 1, respectively, with the step of 1.

#### 4.2.7. Five-fold External Cross-Validation

Five-fold external cross-validation (5FECV) has been used for the estimation of predictive power of developed models. During that the dataset was randomly split into five even subsets, each as an external test set predicted by selected models built upon the rest of four subsets (modeling set), i.e., each compound appeared in an external test set for once. Internal five-fold CV was applied for each of the five modeling sets. Selected models should meet the acceptance criteria defined by certain statistical performance metrics for the modeling sets such as  $q^2$  for continuous models and Correct Classification Rate (CCR) for classification models.

- a) Continuous Modeling. The goal of continuous modeling is to obtain regression models capable of predicting the exact  $pK_i$  values of untested compounds correctly. To select predictive models,  $q^2$  for internal training sets and  $R^2$  for internal test sets should be both greater than 0.6. If none of the generated models met the acceptance criteria, then no predictions should be made; otherwise, consensus prediction was used to determine the predicted  $pK_i$  values for compounds in the external test sets.
- b) Classification Modeling. Similarly, classification modeling aims to obtain predictive classification models capable of separating compounds from potent binders to weak or non-binders. The threshold to distinguish 5-HT<sub>1A</sub> binders and non-binders for the data set was determined as  $pK_i = 7$  (i.e.,  $K_i = 100$  nM). Models selected for prediction should have sensitivity (SE), specificity (SP), and CCR for both internal training sets and internal test sets greater than 0.7. Consensus prediction was used to determine which category (binders or non-binders) the compounds in the external test sets should belong to.



#### 4.2.8. Virtual Screening (VS)

Using developed models based on all kinds of descriptors and data tested on human species, VS was applied to the curated DrugBank database[116,117]. This is not a large database considering it only contains 6629 compounds, but is valuable for the purpose of drug repositioning. Applicability domain (AD) was applied as  $Z_{\text{cutoff}}=0.5$ .

### 4.3. Results and Discussion

#### 4.3.1. Curated Datasets

First, we discarded 15 compounds tested on mouse tissues from ChEMBL, and 161 compounds tested on species other than on human and rat from PDSP. Since the main aim of this pilot study was to compare modeling results under several different conditions, for example, sources of data and protocols for experimental properties, we compiled four datasets based on human and rat data collected from ChEMBL and PDSP (Table 4.2).

**Dataset I** was composed of compounds tested on human tissues from ChEMBL. After structural curation and activity analysis, 155 entries were discarded due to identical structure and varied activities; 135 entries were discarded due to different isomeric forms and varied activities. Final dataset consisted of 1048 unique compounds.

**Dataset II** was composed of compounds tested on rat tissues from ChEMBL. After structural curation and activity analysis, 81 entries were discarded due to identical structure and varied activities; 131 entries were discarded due to different isomeric forms and varied activities. Final dataset consisted of 804 unique compounds.

**Dataset III** was composed of compounds tested on human tissues from PDSP. After structural curation and activity analysis, 294 entries were discarded due to identical structure

and varied activities; 36 entries were discarded due to different isomeric forms and varied activities. Final dataset consisted of 344 unique compounds.

**Dataset IV** was composed of compounds tested on rat tissues from PDSP. After structural curation and activity analysis, 249 entries were discarded due to identical structure and varied activities; 4 entries were discarded due to different isomeric forms and varied activities. Final dataset consisted of 255 unique compounds.

#### 4.3.2. Activity Analysis

After the duplicates were determined, we treated one set of compounds with identical structures as one duplicate case. Activity analysis of the multiple  $pK_i$  values for the same duplicate case was conducted. Figure 4.2 showed the distribution of numbers of duplicate cases at various ranges of  $\Delta pK_i$  values (the difference between the maximal and minimal  $pK_i$  values for the same duplicate case). It is obvious that most duplicate cases had  $\Delta pK_i$  values within 0.5 log units, so we excluded all data records associated with the cases having  $\Delta pK_i$  values greater than 0.5 log units, and assigned the averaged  $pK_i$  values as the activity values for the cases kept in the modeling sets (one unique structure for each duplicate case is kept).

There were some overlaps between Datasets I-IV. Figure 4.3 illustrated the numbers of overlapped compounds between each two of these four datasets (Venn diagram). The correlation of recorded  $pK_i$  values for the overlapped compounds (Figure 4.4) has been investigated. From no more than 30 pairs of overlaps, such correlation coefficients between compounds tested on human and rat tissues were 0.64 and 0.84 for ChEMBL and PDSP, respectively. Thus, we did not merge human and rat data together because the correlation between them were not high enough. On the contrary, the correlation coefficients between compounds from different databases were as high as 0.89 and 0.94 for human and rat data,

respectively, indicating that most data from different databases are identical (probably taken from the same literature).

In summary, activity analysis for the duplicated cases and correlation analysis for the overlaps helped us determine that QSAR models should be built separately on each datasets. However, compounds tested on the same species yet from different databases could be merged. To compare with a recent publication[168], we found that the number of compounds in 5-HT<sub>1A</sub> dataset after carefully curation is much smaller (less than 1,500 vs. over 4,000 in their study). Their dataset may be mixed with all sources of data. A similarity-based method was used to make predictions, but such practice is in doubt to be regarded as rigorous QSAR modeling.

#### 4.3.3. QSAR Modeling

As illustrated by Figure 4.5, QSAR models were built using three types of descriptors (Dragon, MOE, Simplex), two types of modeling techniques (WinSVM, *k*NN), and two types of endpoints (continuous, classification). The number of compounds used to build continuous models was different from that used to build classification models. As mentioned earlier, compounds with variations of activity values for the same duplicate case were excluded from continuous modeling sets. However, if the multiple values for the same duplicate case were not greater (or less) than the threshold to classify the two classes of binders and non-binders, they would be assigned as non-binders (or binders), and should be added back to the classification modeling sets. Table 4.3 showed the number of compounds for each kind of modeling sets.

Prediction accuracies in terms of  $R^2$  for continuous models and CCR for classification models are given in Tables 4.4-4.8. Tables 4.4-4.7 showed the prediction accuracies for each

fold, whereas Table 4.8 showed cumulative prediction performance of consensus WinSVM classification models. *k*NN models used innate local applicability domain (AD) while prediction accuracies of WinSVM models were provided both without and with AD applied. In each case, AD were applied as  $Z_{cutoff} = 0.5$ .

- a) Continuous models. It appeared that continuous models based on Datasets I-IV were so poor that no models were obtained for some folds. While the numbers of accepted models were limited for other folds,  $R^2$  values for the prediction accuracies of external test sets hardly exceeded 0.6, even with AD applied which typically decreased the coverage of predictable external compounds yet no significant effect on improving the  $R^2$  values (Table 4.4 and 4.5 for WinSVM and *k*NN models, respectively). Regardless of the universal poor performance, models based on Simplex descriptors seemed to outperform those built on Dragon or MOE descriptors.
- b) Classification models. Contrary to continuous models, classification models showed excellent prediction accuracies with CCRs ranging from 0.71 to 0.82. Here we should point out the observation that the application of AD did not improve the prediction accuracies significantly. CCRs increased by 0~0.3, and sometimes even decreased by 0.1. The coverage of predictable external compounds decreased to 70% to 81%. Even though AD did not show any advantage of improving prediction accuracies, previous studies stressed the importance of using AD when applying the models to much larger databases to afford higher level of prediction confidence. Finally, consensus prediction (Table 4.8) improved both CCRs and coverage of predictions, indicating advantage over models based upon individual fold or type of descriptors.

c) Cross prediction between the data obtained from different species (human or rat). From Figure 4.3, we know there were always certain compounds not included by other datasets among Datasets I-IV. It would be helpful to predict the absent compounds by models built on another dataset to determine whether these datasets could be merged if the prediction accuracies were high. For example, models based on Dataset I were used to predict the 775 compounds from Dataset II. It could help to determine whether models based on compounds tested on human tissues can predict the properties of compounds tested on rat tissues correctly. Cross prediction results were summarized in Table 4.9, which indicated that cross-prediction accuracies between the models built upon data from different species (human or rat) were so poor that CCRs were below 0.7. On the contrary, cross-prediction accuracies between the models built upon data from same species (but different databases, ChEMBL or PDSP) were reasonably as high as 0.82.

#### 4.3.4. Virtual Screening (VS)

Consensus VS hits were obtained by application of all developed models based on three types of descriptors (Dragon, MOE, and Simplex), two modeling methods (kNN and WinSVM), and human data only. A list of 15 VS hits were prioritized for further experimental testing (Appendix II).

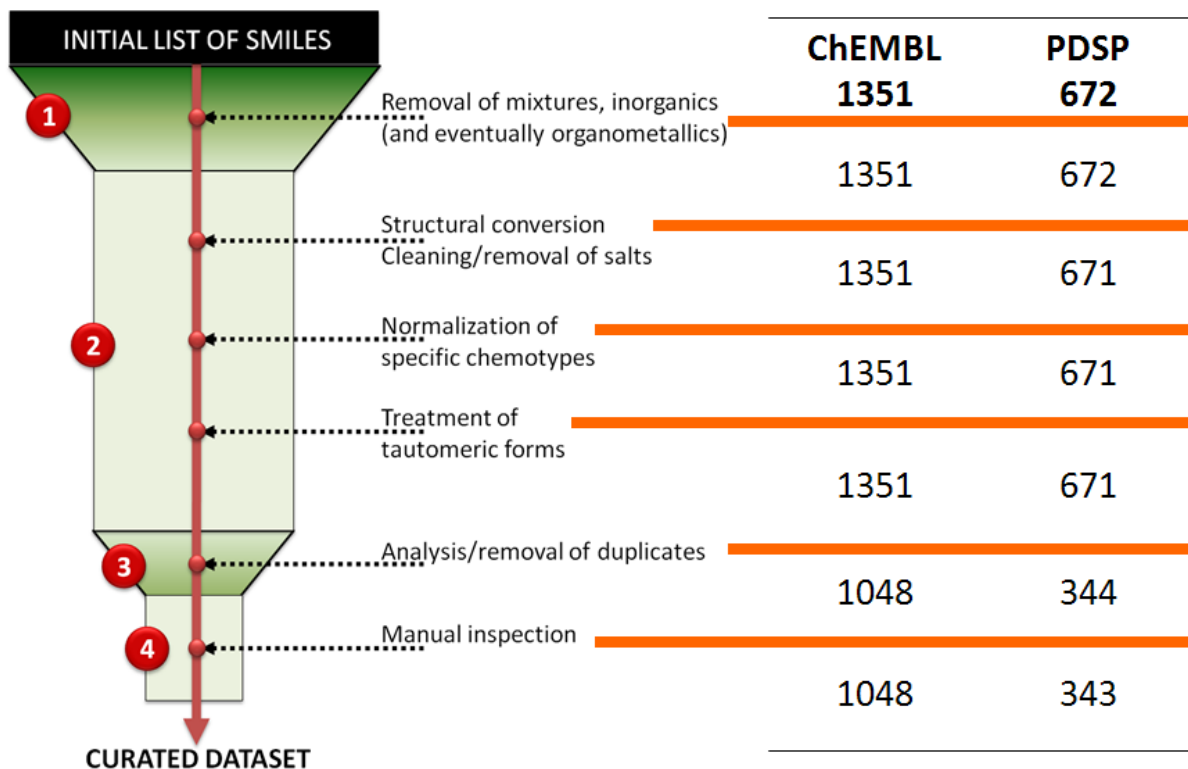
#### 4.4. Conclusions

We applied rigorous chemical and biological data curation for 5-HT<sub>1A</sub> dataset, one of investigated GPCR targets. Rigorous external validation allowed us to develop robust and predictive classification QSAR models based on various types of descriptors, modeling

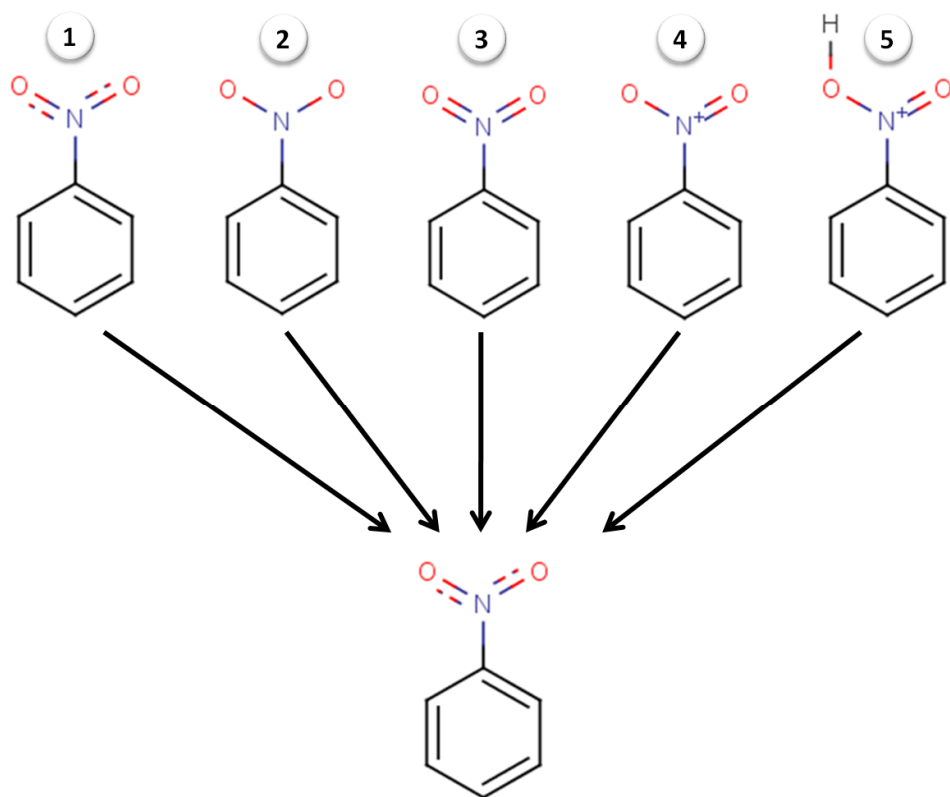
techniques, types of endpoint, and chosen species. At the same time, no continuous models passed the external validation that, probably, caused by low quality of biological data.

Cross-prediction of the models based on data tested on different species reinforced our decision to merge data from different databases but tested on the same species only. Since we are most interested in discovering drugs applying to humans, we will only collect data tested on human tissues.

## Figures and Tables

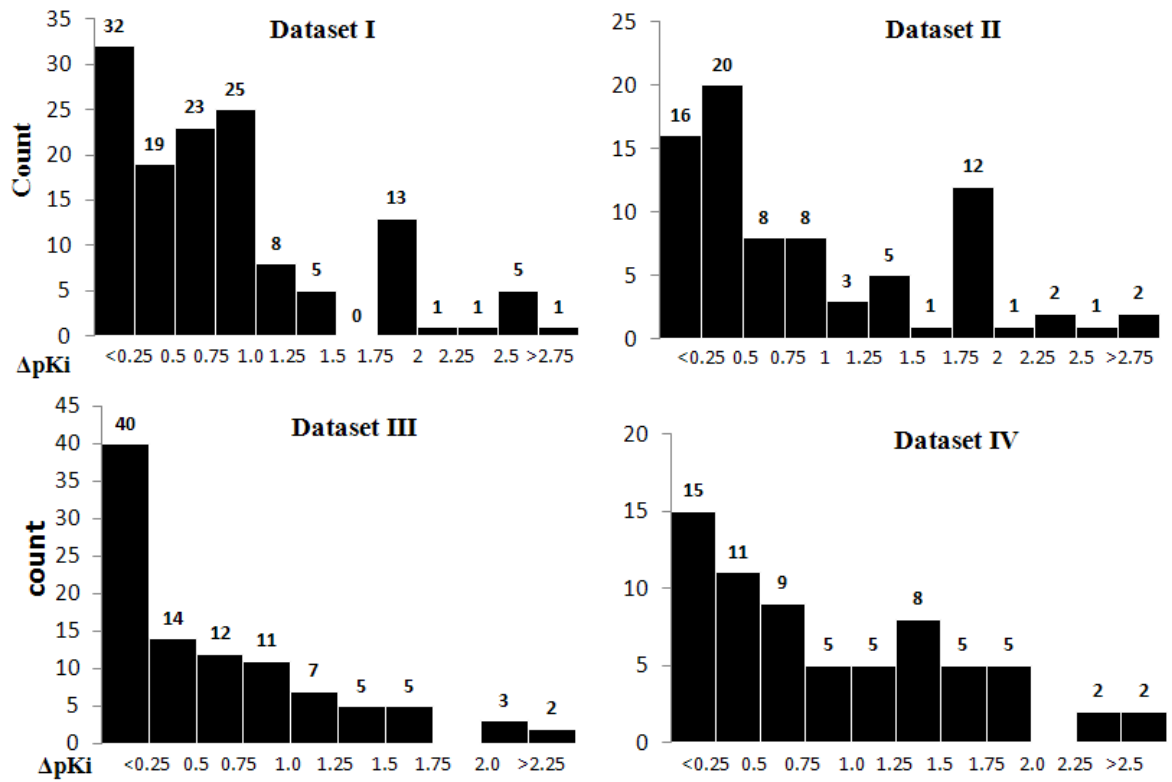


**Figure 0.1. General dataset curation workflow and number of compounds kept after each step.**

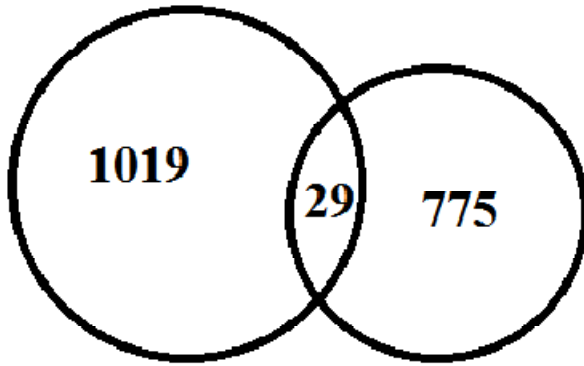


**Figure 0.2. The nitro group which can be represented by five different patterns.**

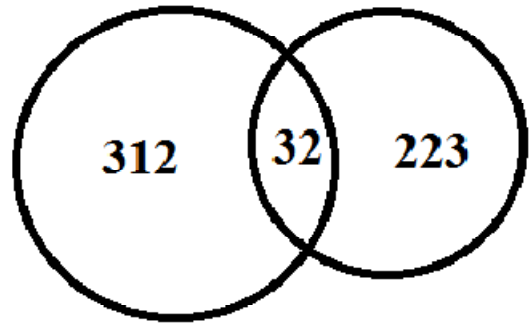




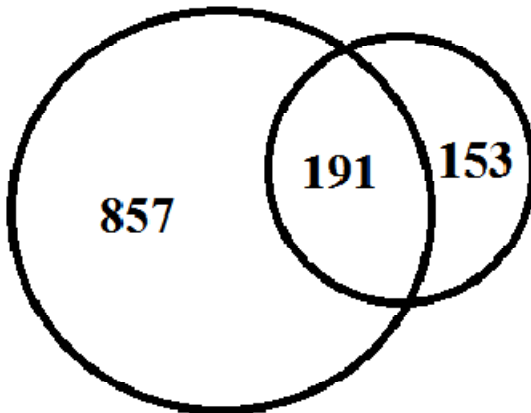
**Figure 0.3. Distribution of activity variation for the duplicate cases.**



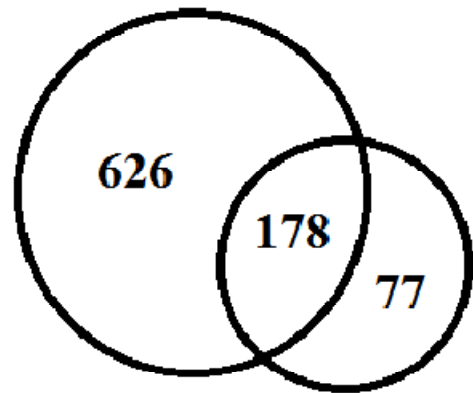
**A**



**B**

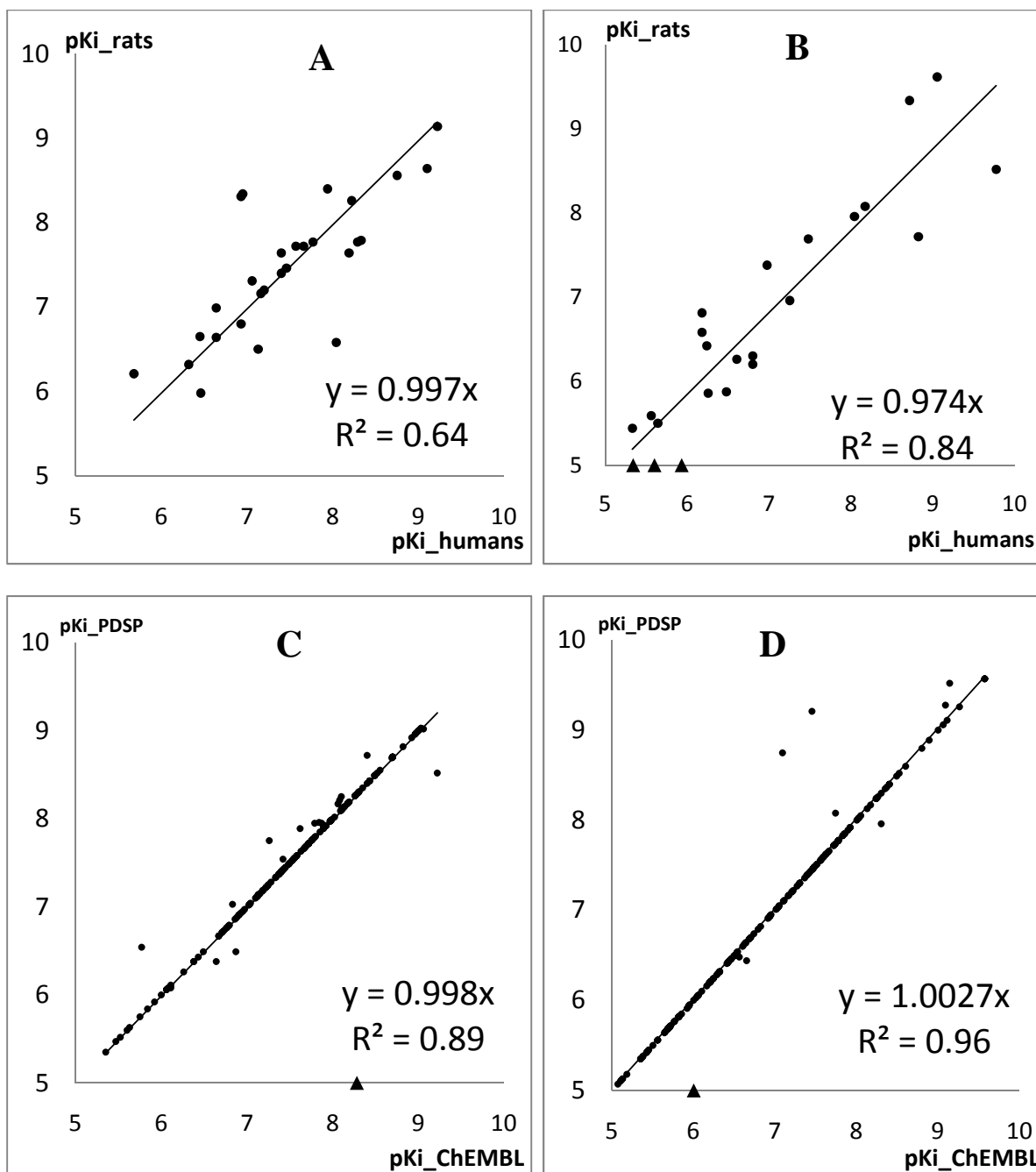


**C**



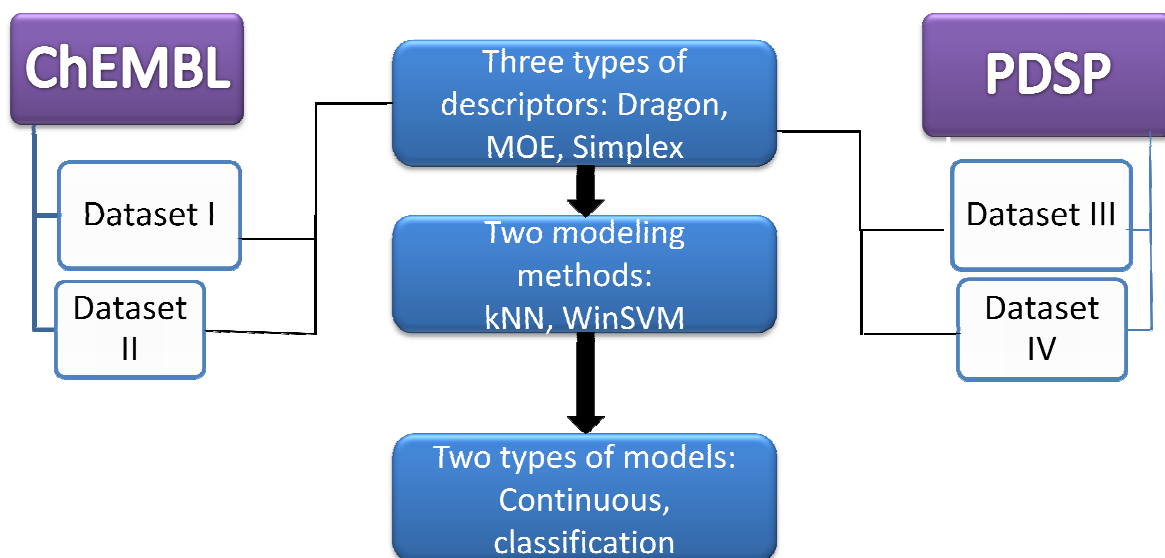
**D**

**Figure 0.4. Overlaps between the datasets.**



**Figure 0.5. Correlation of  $pK_i$  values for the compounds from Datasets I-IV.**

▲ The chemicals that tested as  $pK_i < 5$  on rat organisms but  $> 5$  in human organisms.



**Figure 0.6. Types of descriptors, modeling methods, and variable properties of QSAR models based upon Datasets I-IV.**

**Table 0.1. 5-HT<sub>1A</sub> dataset.**

<b>Resource</b>	<b>Organism</b>	<b>Number of Compounds</b>
<b>ChEMBL</b>	Humans	1351
	Rats	1002
	Mice	15
	<b>Total</b>	<b>2368</b>
<b>PDSP</b>	Humans	672
	Rats	508
	?Humans	14
	Bovine	17
	COS-7	16
	Mice	1
	Pigs	49
	Pigeons	5
	Porcine	1
	Guinea pig	1
	Rabbits	8
	UNDEFINED	49
<b>Total</b>	<b>1341</b>	

**Table 0.2. Four different datasets based on the resources and tested organisms.**

	<b>Humans</b>	<b>Rats</b>
<b>ChEMBL</b>	Dataset I 1048	Dataset II 804
<b>PDSP</b>	Dataset III 343	Dataset IV 255

Note: numbers represent the counts of compounds in each dataset.

**Table 0.3. Numbers of compounds for all continuous and classification modeling sets.**

<b>Dataset</b>	<b>Continuous Modeling Set</b>	<b>Classification Modeling Set</b>		
		<b>Total</b>	<b>Binders</b>	<b>non-binders</b>
<b>Dataset I</b>	<b>1048</b>	<b>1100</b>	713	387
<b>Dataset II</b>	<b>804</b>	<b>843</b>	430	413
<b>Dataset III</b>	<b>344</b>	<b>373</b>	159	214
<b>Dataset IV</b>	<b>255</b>	<b>285</b>	115	170

**Table 0.4. Prediction performance ( $R^2$ ) of developed WinSVM continuous models.**

<b>Datasets</b>	<b>Dragon</b>			<b>MOE</b>			<b>Simplex</b>		
	w/o AD	AD	Cov.	w/o AD	AD <sup>1</sup>	Cov.	w/o AD	AD	Cov.
<b>Dataset I</b>	NA	NA	NA	0.51	0.52	71%	0.51	0.53	72%
<b>Dataset II</b>	0.49	0.53	71%	0.51	0.51	79%	0.49	0.53	71%
<b>Dataset III</b>	NA	NA	NA	0.43	0.49	68%	0.47	0.51	62%
<b>Dataset IV</b>	NA	NA	NA	0.49	0.53	73%	0.53	0.65	71%

**Table 0.5. Prediction performance ( $R^2$ ) of developed  $k$ NN continuous models.**

<b>Dataset</b>	<b>Dragon</b>	<b>MOE</b>	<b>Simplex</b>
<b>Dataset I</b>	N/A	0.09	0.44
<b>Dataset II</b>	0.52	0.06	0.03
<b>Dataset III</b>	0.24	0.31	0.35
<b>Dataset IV</b>	0.34	0.28	0.40

N/A: no models passed the acceptability criteria



**Table 0.6. Prediction performance (CCR) of developed WinSVM classification models.**

<b>Datasets</b>	<b>Dragon</b>			<b>MOE</b>			<b>Simplex</b>		
	w/o AD	AD	Cov.	w/o AD	AD <sup>1</sup>	Cov.	w/o AD	AD	Cov.
<b>Dataset I</b>	0.75	0.77	73%	0.72	0.74	74%	0.76	0.76	73%
<b>Dataset II</b>	0.82	0.82	75%	0.75	0.75	81%	0.77	0.77	70%
<b>Dataset III</b>	0.80	0.81	76%	0.80	0.79	72%	0.79	0.80	75%
<b>Dataset IV</b>	0.77	0.80	74%	0.71	0.70	74%	0.79	0.80	75%

**Table 0.7. Prediction performance (CCR) of developed *k*NN classification models.**

<b>Dataset</b>	<b>Dragon</b>	<b>MOE</b>	<b>Simplex</b>
<b>Dataset I</b>	0.78	0.75	0.77
<b>Dataset II</b>	0.80	0.76	0.77
<b>Dataset III</b>	0.79	0.78	0.80
<b>Dataset IV</b>	0.75	0.73	0.76

**Table 0.8. Prediction performance of consensus WinSVM classification models.**

<b>Dataset</b>		<b>Sensitivity</b>	<b>Specificity</b>	<b>CCR</b>	<b>Coverage</b>
<b>Dataset I</b>	w/o AD	0.86	0.68	0.77	100%
	AD	0.88	0.70	0.79	88%
<b>Dataset II</b>	w/o AD	0.80	0.76	0.78	100%
	AD	0.82	0.77	0.80	89%
<b>Dataset III</b>	w/o AD	0.77	0.82	0.80	100%
	AD	0.80	0.81	0.81	86%
<b>Dataset IV</b>	w/o AD	0.71	0.86	0.79	100%
	AD	0.75	0.85	0.80	87%

**Table 0.9. Prediction accuracies (CCR) of cross prediction between Datasets I-IV.**

<b>External Modeling</b>	<b>Dataset I</b>	<b>Dataset II</b>	<b>Dataset III</b>	<b>Dataset IV</b>
<b>Dataset I</b>	N/A	0.50 (775)	<b>0.75 (153)</b>	NA
<b>Dataset II</b>	0.53 (1019)	NA	NA	<b>0.82 (77)</b>
<b>Dataset III</b>	<b>0.63 (857)</b>	NA	NA	0.67 (223)
<b>Dataset IV</b>	NA	<b>0.71 (626)</b>	0.74 (312)	NA

**Note:**

1. Numbers in the brackets are the counts of compounds being predicted.
2. "N/A" means no prediction was performed between the crossed datasets.

## CHAPTER 5

### QSAROME OF THE RECEPTOROME: QSAR MODELING OF MULTIPLE LIGAND SETS ACTING AT MULTIPLE RECEPTORS

#### 5.1. Introduction

G protein-coupled receptors (GPCRs) are transmembrane proteins functioning as the media to transduce external stimuli into intracellular signals. The biology and physiology of GPCRs have been considered for a long time as intriguing areas for scientific studies, and the excitement still persists. Their importance is highlighted by the fact that at least one third of currently marketed drugs target GPCRs[13]. The number of drugs that target GPCRs is still expected to increase considering the intense research effort for the elucidation of unknown functions of GPCRs and that only 10% of GPCRs are established drug targets so far[169]. As the panel of GPCRs grows sufficiently and functions are comprehensively characterized, a systematic analysis of the ‘receptorome’ (the portion of the proteome encoding receptors) is viable for important discoveries[10]. This approach has successfully helped to discover the molecular mechanisms underlying serious drug side effect, e.g., phen/fen-induced heart disease[170] and weight gain triggered by atypical antipsychotics[171].

In the QSAR field, the counterpart of receptorome summarized above could be referred to as the ‘QSARome’. Due to the recent advances in high-throughput screening and rapid growth of publicly available databases of biologically tested compounds[2,3], we are motivated to develop this efficient modeling protocol to help elucidate more complex actions

of chemicals and unknown functions of drug targets. In response to the disclosure of the poly-pharmacological nature of antipsychotic drugs[172], it is suggested that drugs selectively targeting a combination of GPCRs might provide more benefits than single-action agents in many CNS disorders. This implication challenges the conventional QSAR practices where a specific target was often emphasized by virtue of focusing on relatively simple datasets. On the contrary, we should use sophisticated computational methods to develop a compendium of predictors based on available complex data. The compendium of predictors will in turn shed a light on the interaction of agents with their targets in a systematic way. Overall speaking, we expect QSARome to be an efficient complement to the approach of receptorome in the sense of both elucidating the actions of natural compounds and validating molecular targets for drug discovery[173].

To establish reliable QSARome predictors, we should notice that the explosive growth of publicly available large databases[4] concurs with the plague of false annotations recorded in the databases. It was even critiqued that more than half of the published studies from academic laboratories fell into the dilemma of reproducibility by subsequent confirmation experiments[174–176]. While the percentage of erroneous recordings in public databases is hard to estimate (error rates of the commercial databases, however, range from 0.1 to 3.4%, according to a recent study[153]), it is indeed a severe problem which would deteriorate the quality of QSAR modeling so badly that sometimes no valid QSAR models could be generated from biological meaningful datasets. Unfortunately, even experienced QSAR practitioners often make less effort on data curation and jump to the statistical quality of models too hastily. In the previous chapter, we designed clear and basic steps for data curation which could form a foundation for the subsequent QSARome project, and used the

5-HT<sub>1A</sub> receptor as an example to demonstrate the importance of data curation for QSAR modeling (Chapter 4). The established approach was employed as a universal approach for all datasets involved in the QSARome project, but should be adapted for specific issues when applying to other data.

## 5.2. Datasets and Methods

### 5.2.1. Datasets

The datasets containing compounds which interact with 34 mostly studied GPCR targets were extracted from the databases of ChEMBL[2] and PDSP[3]. We focused on the compounds with  $K_i$  values tested on the human tissues, excluding the ones tested on rats, mice, or any other animal models. The  $K_i$  values were expressed as mol/L, which were converted to negative  $\log[1/(\text{mol/L})]$  values ( $pK_i$ ) according to standard QSAR practices. The structures of the data were verified following the protocol established by our group[77]. If there were different  $pK_i$  values for one molecule with exactly the same type of interaction (either within the same database or across databases), the standard deviations were calculated. Then we excluded the molecules with the deviation greater than 0.5 and assigned average  $pK_i$  values for the remaining molecules. Finally, unique compounds with interactions with the 34 GPCRs were employed as the dataset for further modeling and validation studies. Chemical structures of all compounds and their experimental  $pK_i$  values used in the study are available from the authors upon request.

The above 34 GPCR targets included 10 serotonin receptors (5-HT<sub>1A</sub>, 5-HT<sub>1B</sub>, 5-HT<sub>1D</sub>, 5-HT<sub>1E</sub>, 5-HT<sub>2A</sub>, 5-HT<sub>2C</sub>, 5-HT<sub>3</sub>, 5-HT<sub>5</sub>, 5-HT<sub>6</sub>, and 5-HT<sub>7</sub>), 7 adrenoceptors ( $\alpha_{1A}$ ,  $\alpha_{1B}$ ,  $\alpha_{2A}$ ,  $\alpha_{2B}$ ,  $\alpha_{2C}$ ,  $\beta_1$ , and  $\beta_2$ ), 5 dopamine receptors (D<sub>1</sub>, D<sub>2</sub>, D<sub>3</sub>, D<sub>4</sub>, and D<sub>5</sub>), 5 muscarinic receptors (M<sub>1</sub>, M<sub>2</sub>, M<sub>3</sub>, M<sub>4</sub>, and M<sub>5</sub>), 4 histamine receptors (H<sub>1</sub>, H<sub>2</sub>, H<sub>3</sub>, and H<sub>4</sub>), and 3 neurotransmitter

transporters (serotonin (SERT), norepinephrine (NET), and dopamine (DAT)). Recently, the affinity profiles ( $pK_i$  values) of these protein targets to 13 antipsychotic drugs were released[172] and brought into our attention by Vidal and Mestres[168]. The 13 antipsychotic drugs, which were utilized to constitute the external validation set, contained six first generation antipsychotics (chlorpromazine, fluphenazine, haloperidol, loxapine, thioridazine, and thiothixene) and seven second generation antipsychotics (clozapine, olanzapine, quetiapine, risperidone, zotepine, ziprasidone, and aripiprazole).

ChEMBL is a database integrating literature reported bioactivities of drug-like small molecules. It contains 2D structures, calculated chemical properties (e.g., logP, Molecular Weight, etc.) and abstracted bioactivities (in our case, the binding constant  $K_i$ ). The most recent version of ChEMBL\_11 contains about 1.2 million compound records against 8,603 targets, abstracted from more than 42,000 publications[2]. Compounds extracted from this database represented over 90% of the whole datasets we collected.

The NIMH Psychoactive Drug Screening Program (PDSP) is a unique national resource devoted to discovering new treatments for mental illness. Dr. Roth has implemented many biological assays and made it a reliable source of antipsychotic drug-target interactions. Although PDSP provides a less portion of extracted compounds, it provides additional assay values to most existed compounds in ChEMBL, and certain values are labeled as PDSP certified activity, indicating higher reliability for the annotations.

### 5.2.2. Descriptor Generation

The QSARome models for all final unique compounds were developed with the chemical descriptors calculated by the Dragon software v5.5[162] based on the curated SD file of the structures. Only 0D, 1D, and 2D descriptors were considered in this study. The



initial number of Dragon chemical descriptors was as high as 2442, which was processed as follows. First, we removed all descriptors that had zero values or zero variance for all compounds. Furthermore, redundant descriptors were identified by analyzing correlation coefficients between all pairs of descriptors; if the correlation coefficient between two descriptor types for all modeling set compounds was higher than 0.95, one of them was randomly chosen and removed. As a result, the final total number of Dragon descriptors used for model building was reduced to 762, and the descriptors were normalized to 0~1 based on the maximal and minimal values of each descriptor type in whole dataset. A detailed explanation of descriptor generation and preparation procedures can be found elsewhere[18].

### 5.2.3. Building Models with Support Vector Machines (SVM)

The Support Vector Machines (SVM) algorithm[33] packaged in the R program (e1071) was used in this study in the light of its popularity for machine learning. It serves 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. Briefly, an SVM model finds a separating hyper-plane with a maximal margin in the feature space by minimizing the special loss function. To cope better with different classification tasks, e.g., linear vs. nonlinear correlations, a handful of kernel functions were developed to map the original descriptor space to a higher dimensional feature space for modeling purpose. In this study, we built models with the linear kernel. The cost was assigned from 0 to 10 with a step of 1 to search for the optimal separating margin.

### 5.2.4. Applicability Domain (AD)

If a query compound is highly dissimilar to all other compounds in the modeling set, its predicted activity by the developed models could be unreliable. Thus, it is critical to define a

proper Applicability Domain (AD) of a model. In this study, AD for each of the models was determined by a threshold Euclidean distance  $D_T$  between each query compound and its nearest neighbors in the modeling set.  $D_T$  was represented by all chemical descriptors (referred as the global AD) and calculated as follows:

$$D_T = \bar{y} + Z\sigma \dots \dots \dots (5.1)$$

where  $\bar{y}$  is the average Euclidean distance between each compound and its k nearest neighbors (k=1) in the modeling set,  $\sigma$  is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Usually, the default Z value is 0.5, that is, the AD for a model places its boundary at one-half of the standard deviation calculated for the distribution of distances between each compound in the modeling set and its k nearest neighbors (k=1) in the same set. If the distance of a test compound from any of its nearest neighbors exceeds the threshold, the prediction is considered unreliable. However, by applying AD for each model, only a certain fraction of compounds in any external dataset is expected to fall within the AD. The fraction is therefore referred to as the coverage and should be considered for external prediction results. Detailed description of AD definition and calculation could be found in our previous publications[9,39].

#### 5.2.5. Data Division for Model Building and Validation

A set of 13 antipsychotic drugs (chlorpromazine, fluphenazine, haloperidol, loxapine, thioridazine, thiothixene, clozapine, olanzapine, quetiapine, risperidone, zotepine, ziprasidone, and aripiprazole)[168,172] which had bioactivities tested on all 34 GPCRs was set aside as the external set for evaluation of the prediction performance of built models. Although this external set is pretty small in comparison with the modeling datasets, it is a

valuable source to validate the obtained models because external compounds were not involved in any modeling procedure and were tested against all the investigated targets.

Five-fold external cross-validation (5FECV) has been used for the assessment of predictive power of the developed models. For each of the 34 datasets, 20% of compounds were extracted randomly for five times (already chosen compounds were out of selection for each next time) to compose an external validation set. In this case, every compound will be set aside in the external sets for once. An SVM model was obtained each time on the remaining 80% of compounds. The overall CCR was then calculated based on the integrated predictions of the whole modeling set composed of the five external sets, with error bars calculated from the five corresponding CCRs (Figure 5.2). The importance of a sufficient external validation set was stated by our previous publications[38] and has been integrated long ago as a part of our predictive QSAR modeling workflow (Figure 2.1, Chapter 2).

#### 5.2.6. Y-Randomization test

Y-Randomization test is a widely used validation technique to ensure the absence of chance correlations during obtaining of a QSAR model[177]. This test includes (i) randomly shuffling the dependent-variable vector, Y-vector of training sets (class labels in this study), and (ii) rebuilding models with the randomized activities (class labels) of the training sets. This procedure was applied to each of the five splits of each dataset. It is expected that the resulting QSAR classification models, built with randomized activities for the training set, should generally have low CCRs for training, test. It is likely that sometimes, though infrequently, high CCR values may be obtained because of a chance correlation or structural redundancy of the training set. However, if some QSAR classification models obtained in the

Y-randomization test have relatively high CCR, it implies that an acceptable QSAR classification model cannot be obtained for the given data set.

#### 5.2.7. Gap Filling and Virtual Screening (VS)

The matrix of investigated activities used for modeling was dramatically sparse (only 6.75% were filled), so developed and externally validated models were applied to fill up the gaps within the matrix. Because the descriptors were calculated and linearly normalized based on the whole data matrix, no further treatment of descriptors was needed. However, we did restrict ourselves to the most conservative applicability domain using  $Z_{cutoff} = 0.5$  when making predictions. Overall, 83.5% of the gaps were filled by exercising this AD criterion on QSAR predictions.

### 5.3. Results and Discussion

#### 5.3.1. Curated Data Matrix

It has been shown that an unclean dataset would lead to unsuccessful QSAR practices [77]. At the current stage of explosive growth of publicly available bioactivity databases, data curation becomes the most critical prerequisite for building any QSAR models, because all the databases contained certain percentage of different errors[4]. According to the protocol established earlier by our lab[77], we performed a thorough examination of the original data for inorganics, organometallics, mixtures, tautomers, and duplicates (See Table 5.1 for examples). Furthermore, activity analysis was undergone by calculating the variations of these multiple values if the duplicates identified by the structures were associated with multiple  $pK_i$  values to the same protein target (See Table 5.2 for example). In total, there were 1175 cases with standard deviations greater than 0.5 that should not be considered for

QSAR modeling unless all reported  $pK_i$  values were above (or below) the corresponding cutoff values.

In addition, we expected that activity variation between the datasets were biased to be smaller, because there must be a great number of the same bioactivities recorded by both databases, which would reduce the observed deviations. A complete identification of the duplicate records in both databases simultaneously, however, would be exhaustive and unrealistic as one has to examine all the corresponding references cited in both sources. Another issue found by our analysis was the wrong translation from literature to databases (See an example shown in Figure 5.1). The record was wrongly taken from the article[178] as 398 instead of 5.8 for  $pK_i$ . Due to the scale of our study, we were incapable of removing such errors completely, but alert it here for database managers to beware of this issue.

The final data matrix, constructed from 22,633 ChEMBL and 11,243 PDSP records, is composed of 9,088 unique compounds with cleaned bioactivities against 34 GPCR targets. However, only 6.75% of the matrix is filled, i.e., the data matrix is very sparse (Figure 5.7).

### 5.3.2. Validation of SVM Models

To streamline the modeling process, we applied the simple rule of cutoff adjustment to balance each dataset for classification modeling. Thus, dataset-specific cutoff value to distinguish binders and non-binders was determined based on the balance of these two classes within a given dataset. For most datasets,  $pK_i = 7$  was used as the cutoff values since the ratios of these two classes were restricted within  $[1/3, 3]$ . The exceptions were: 5-HT<sub>1E</sub>, 5-HT<sub>3</sub>, 5-HT<sub>5</sub>, D<sub>5</sub>, H<sub>2</sub>, H<sub>3</sub>, M<sub>4</sub>, and M<sub>5</sub>. Composition and cutoff values for all 34 datasets are depicted in Figure 5.2.

The 5FECV results for the 34 SVM models developed for investigated GPCRs are shown in Figure 5.3. Our SVM models reached CCR above 0.70 for 33 out of 34 GPCR targets. For nine targets (5-HT<sub>2A</sub>, 5-HT<sub>3</sub>, 5-HT<sub>5</sub>,  $\alpha_{1A}$ ,  $\alpha_{1B}$ , M<sub>3</sub>, M<sub>4</sub>, SERT, and NET), CCRs were even greater than 0.80. The only CCR below 0.70 is referred to 5-HT<sub>1E</sub>, which has the smallest size and contained only 37 defined binders. As a consequence, it is not surprised to observe a high deviation (0.16) of CCRs from 5FECV. Overall, we consider our approach is capable of assessing the GPCR binding profile of a query compound given it is within models' applicability domains.

### 5.3.3. Prediction Performance for the External Matrix

As shown in Figure 5.4, the predicted GPCR bioprofiles of the 13 drugs were colored based on the prediction accuracy. The overall prediction accuracy was 70.6%. It is noted that thiothixene and quetiapine were out of AD for most models, and, expectedly, their predictions were as poor as 42.4% and 51.5%, respectively. The overall accuracy reached 77.3% after exclusion of these two molecules. One should keep in mind that as an external set, 13 drugs for any individual model provide limited information. However, taking the external matrix as a whole gave us a certain level of confidence for the models' predictive power.

### 5.3.4. Filling the Gaps in the Matrix

Developed predictive models were used for gap-filling in the initial data matrix. At this step, we excluded the compounds that were already tested and were active on more than 3 GPCRs because they will not be selective, which resulted in 7,267 compounds that were expected to be valuable for further investigation. This yielded a mixed matrix with even less experimental bioactivities and the rest of them with predicted bioactivities. Figure 5.5 shows

the distribution of number of compounds that interact with the 34 GPCRs. The 7,267 compounds were then ranked by the number of bound GPCR targets (either experimentally tested or predicted). Among those compounds, we identified 146 compounds that bind to only one or two GPCR targets (the “magic bullet”). In addition, it is also viable to select compounds with preferred binding profiles, which referred to as the “magic shotgun” that interact with a preferred class of GPCRs hypothesized to improve therapeutic effects and avoid some serious side effects[172].

The applicability domain was determined by  $Z_{cutoff} = 0.5$  when making predictions, and was an important element to consider when prioritizing compounds for further experimental testing, no matter for selective binders or nonselective binders that target a specific combination of GPCR targets. The coverage of overall predicted matrix, excluding the ones used in the modeling matrix, was 83.5%, varying from 59.5% to 95.0% for each target.

#### 5.3.5. Experimental Testing

Based on the mixed matrix with both experimental and predicted  $pK_i$  values, we ranked the compounds by the number of targets it will bind. Figure 5.6 shows the distribution of compounds targeting various numbers of GPCRs. As illustrated, most compounds interact with 7~15 targets. This is consistent with a recent study about binding profiles of a virtual molecule library GPB-13[179]. In their pure computational study, the active group of compounds averagely bound to 8 protein targets[15]. Taking applicability domain into consideration, a list of 148 selective ligands (binders for only 1 or 2 GPCRs) was provided for experimental testing (Appendix III).

## 5.4. Conclusions

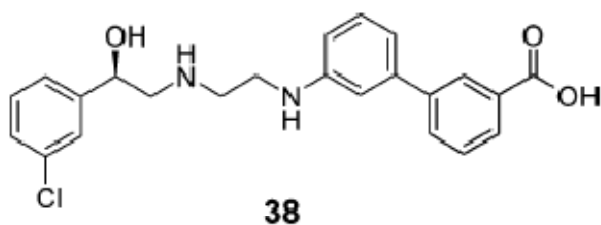
In this Chapter, we collected a full-scale interaction matrix of various chemicals tested against 34 well-studied GPCRs. Rigorous structural curation and activity analysis were performed to yield clean structures and standardized activity values, which constituted the foundation for our QSARome modeling. We thus built a compendium of SVM models based on each of the 34 GPCR datasets. The predictive power of each model was first examined by 5-fold external cross-validation. Among our 34 predictors, 33 had CCRs greater than 0.70 and nine of them greater than 0.80, indicating high predictive power for individual models. Furthermore, an external set of 13 drugs was utilized to assess the ability of these predictors as a whole to profile a given chemical's polypharmacological characteristics. The accuracy of predicting the external matrix reached 70.6%. This strategy, as validated by various levels of external validation, is very meaningful to help discover antipsychotic drug candidates with favorable binding profiles, which will enhance therapeutic effects but avoid serious side effects. According to that, list of selective compounds (binders for only 1 or 2 GPCRs) was provided for experimental testing.

Currently, the dimension of the external set (13 drug  $\times$  34 targets) and target diversity (31 aminergic GPCRs and 3 neurotransmitter transporters) are both limited. In order to obtain a higher level of confidence for QSARome models, we should rely on the experimental tests and generate even more well-rounded data for validation purpose. Nevertheless, to the best of our knowledge, this is the first robust QSAR modeling exercise that operates on multiple drug targets interacting with a large scope of rigorously curated compounds. In order to make our QSARome practices publicly available to general scientific community, all curated datasets and developed models would be posted on-line on our ChemBench server[180].



QSAR modeling has proved to be an efficient complement to *in vitro* screening. Most of previous QSAR studies including earlier works from our laboratory have focused on relatively “simple” datasets against a specific target *in vitro*. Nowadays, multiple biological responses have been measured for a set of compounds, thus the QSARome approach established by this study will again provide an efficient avenue to rapidly reveal drug-target interactions, unravel mechanisms of action, and evaluate chemical toxicity when expanded to other protein targets.

## Figures and Tables



$\beta_3$ AR activity	$\beta_2$ AR <sup>a</sup> binding	$\beta_3$ functional/ $\beta_2$ binding <sup>b</sup>	$\beta_1$ AR binding <sup>a</sup>	$\beta_3$ functional/ $\beta_2$ binding <sup>b</sup>
$8.4 \pm 0.2$	$5.8 \pm 0.5$	398	$6.4 \pm 0.5$	100

<sup>a</sup> The binding constant  $pK_i$  of compound **38** ( $n = 3$ ) against  $\beta_2$  or  $\beta_1$  ARs; see Experimental Section. <sup>b</sup> The ratio of the  $pIC_{50}$  of the compound for  $\beta_3$  AR relative to the binding constant for  $\beta_2$  or  $\beta_1$  ARs.

### Figure 0.1. An example of wrong translation from literature to ChEMBL.

This is a piece of Table 8 from *J. Med. Chem.* 2006, 49, 2758-2771. Observed error is the recorded value “398” instead of the authentic value “5.8” for  $\beta_2$ .

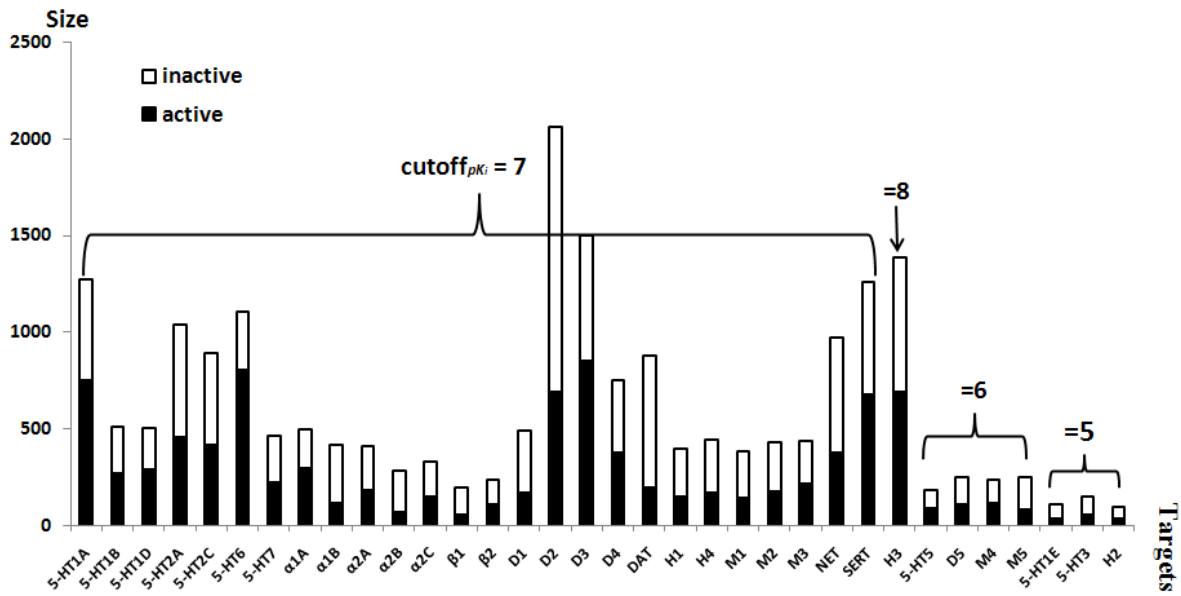


Figure 0.2. Data distribution and  $pK_i$  cutoff values for each dataset.

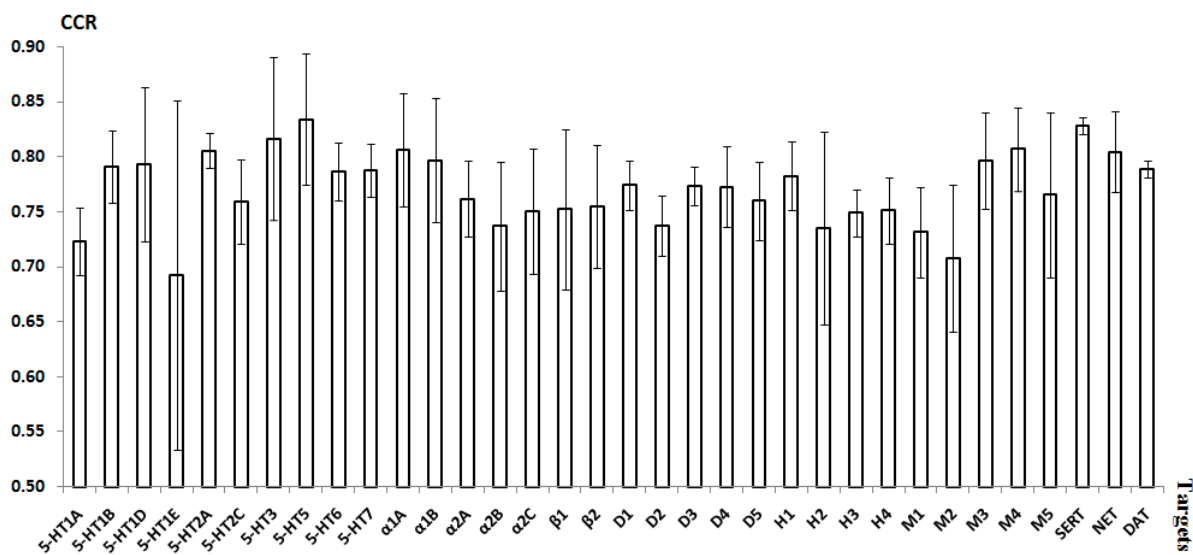


Figure 0.3. Prediction performance of SVM models. CCR is the cumulative CCR values of 5 external folds.

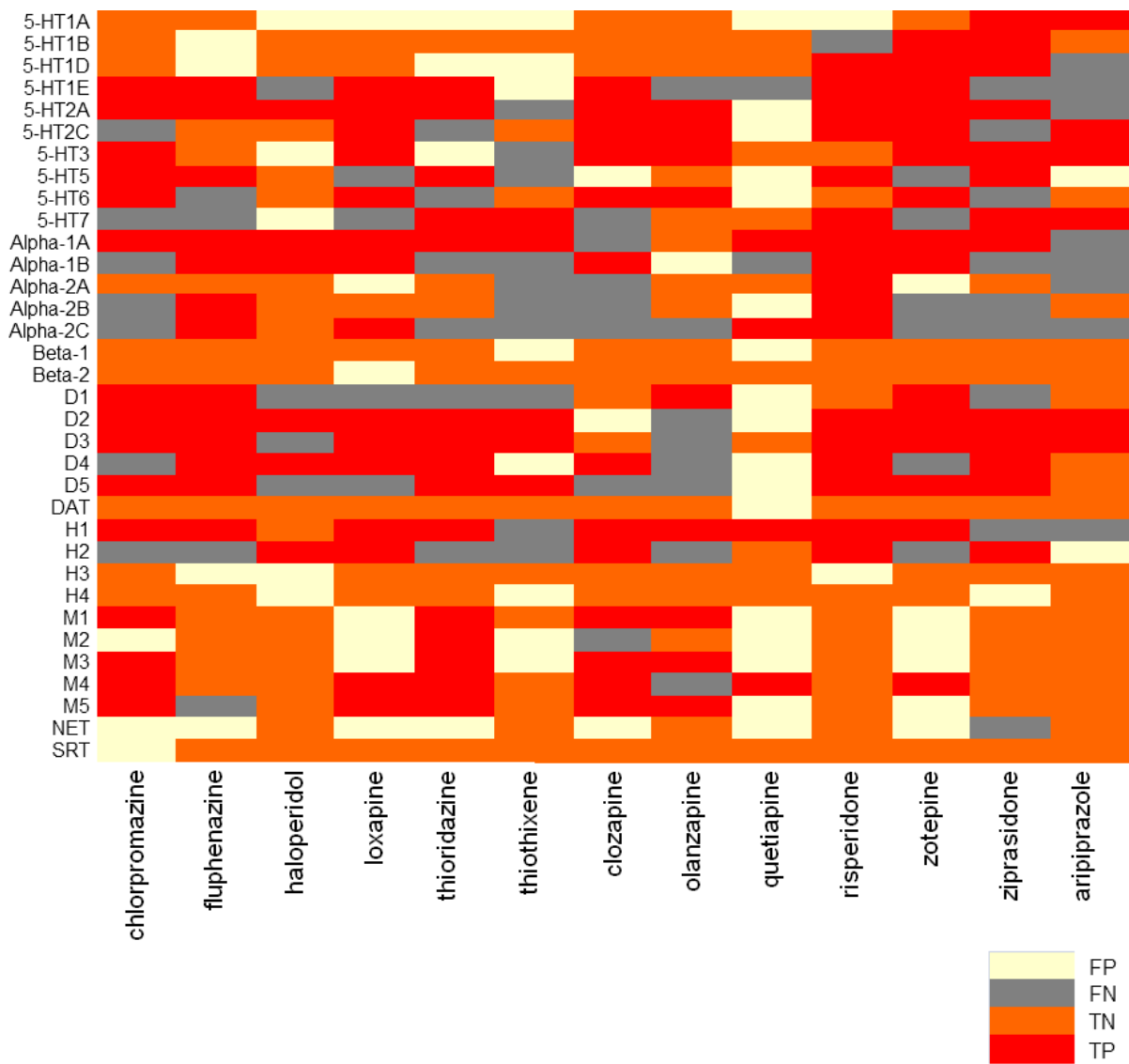
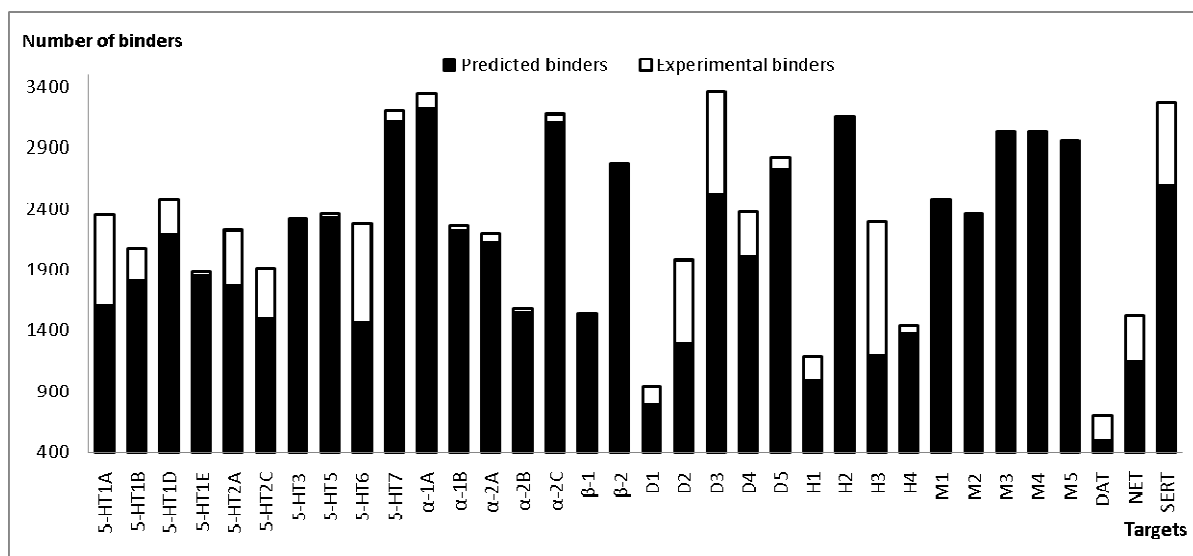
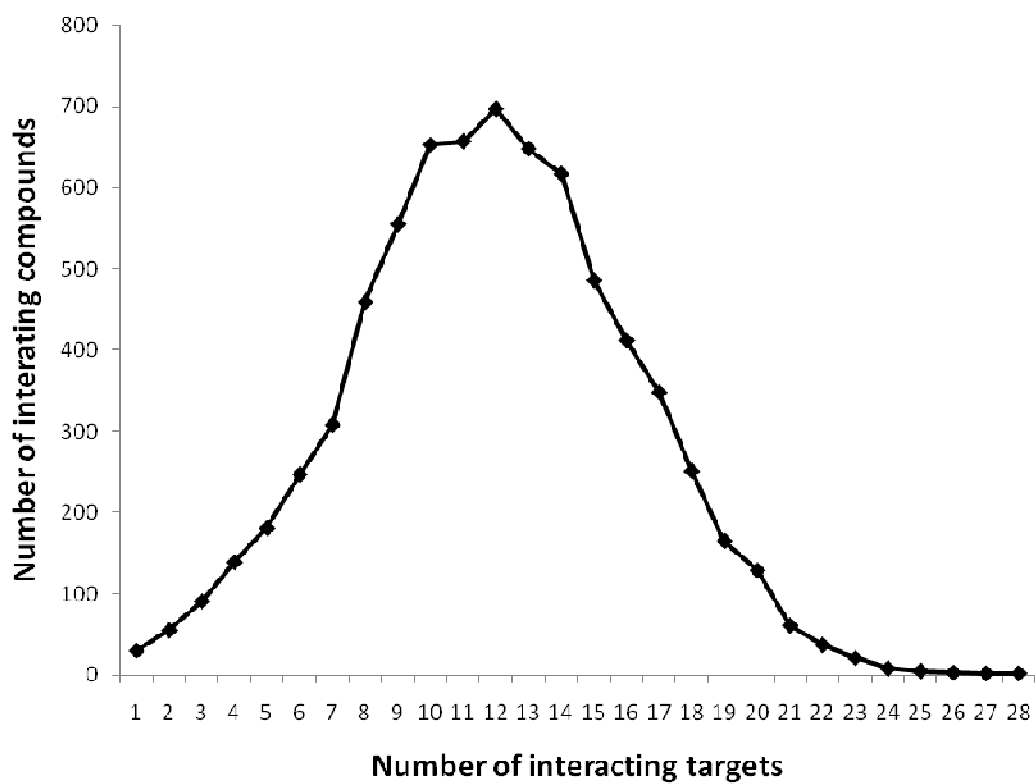


Figure 0.4. Prediction performance for the external matrix composed of 13 drugs.



**Figure 0.5. Number of binders (either experimental or predicted) to each GPCR target.**



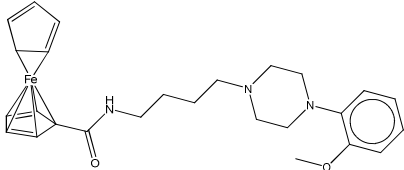
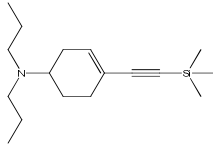
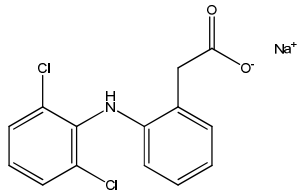
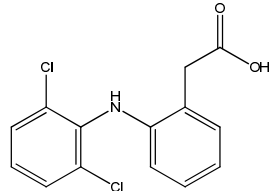
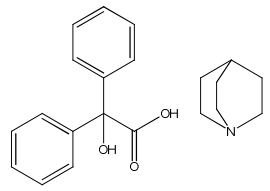
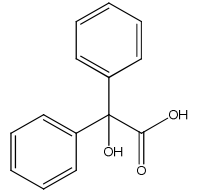
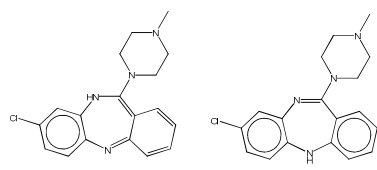
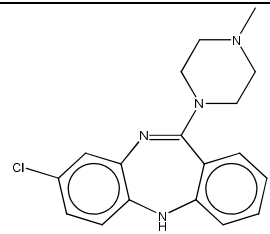
**Figure 0.6. Distribution of compounds targeting various numbers of GPCRs.**

	5-HT3	5-HT5	5-HT6	5-HT7	$\alpha$ -1A	$\alpha$ -1B	$\alpha$ -2A	$\alpha$ -2B	$\alpha$ -2C	$\beta$ -1	$\beta$ -2
COMP1			5	5							
COMP2				5						8.96	9.04
COMP3											
COMP4			5.5	5.5	5.74	5.74					
COMP5			8.33	7.13							
COMP6			8.68	7.62							
COMP7											
COMP8					8.39	7.32	8.07	7.82	7.59		
COMP9	5									9.34	9.68
COMP10							8.6	7.96	8.68		
COMP11		7.7	6.29	9.22							
COMP12			8.87	6.59							
COMP13		5	5	6.67							
COMP14		5	5	5							
COMP15		5	5.19	6.15							
COMP16		5	5.35	6.39							
COMP17		5	5.13	6.99							
COMP18		5	5	5							
COMP19		5	5.15	6.5							
COMP20		5	5	6.47							
COMP21		5	5	6.38							
COMP22				6			8.23	7.64	7.72		
COMP23			8.9	5.3							
COMP24			9	5.7							
COMP25			9	5.6							
COMP26			9.2	5							
COMP27			8.9	5.3							
COMP28			9	5.8							
COMP29			8.7	5.1							
COMP30			8.9	5.3							
COMP31				6.82			7.26		6.88		
COMP32				7.46			6.39		6.13		
COMP33				8.01			7.06		7.13		
COMP34				7.89			6.9		7.16		
COMP35		5.3	5.3	6.15							
COMP36			5	6.9							

**Figure 0.7. The heat map of final curated matrix (part).**

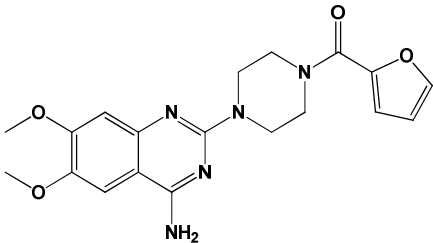
Grey shows gaps without experimental assay values. Sparsity degree = 93.25%.

**Table 0.1. Examples of chemical structure processing during basic data curation.**

Issues	Source	Before curation	After curation
Organometallics	ChEMBL		Deleted
	PDSP		Deleted
Salts or Mixtures	PDSP		
	PDSP		
Tautomers	ChEMBL PDSP		



**Table 0.2. Prazosin – an example of excluded compound with multiple records and high activity deviation in both ChEMBL and PDSP.**

<p><b>Prazosin</b></p>			
<p><b>Targets</b></p>	<p><b>5-HT<sub>2A</sub></b></p>	<p><b>α-1A</b></p>	<p><b>D2</b></p>
<p><b>Maximal deviation</b></p>	<p>5</p>	<p>2.06</p>	<p>0.95</p>
<p><b>Assay records (pK<sub>i</sub>)</b></p>	<p>5.15 5.45 10.15</p>	<p>9.16 10.22 8.74 8.14 9.29 9.23 9.23</p>	<p>7.24 7.51 7.84 7.97 7.02</p>

## CHAPTER 6

### CONCLUSIONS AND FUTURE DIRECTIONS

6.1. Application of current cheminformatic techniques to human 5-HT<sub>7</sub> datasets to build validated and predictive QSAR classification models for drug repurposing.

The 5-HT<sub>7</sub> receptor is the newest member in the serotonin receptor family. However, it has recently become investigated more and more since many studies indicated that it is involved in a large number of psychological and behavioral functions. In Chapter 2 we used *MolConnZ* descriptors to generate continuous *k*NN models for 62 receptor binders. In addition, we obtained a classification DWD model for the same dataset enriched by additional 38 non-binders. All developed QSAR models were rigorously validated and possessed high external predictive power. Seven compounds, all of which were known drugs, with high predicted activity were identified by virtual screening of WDI database. Subsequent experimental testing confirmed that five of them were potent 5-HT<sub>7</sub> receptor binders. Droperidol and perospirone were the most potent and had affinity to 5-HT<sub>7</sub> receptor at nanomolar level. They were later identified as 5-HT<sub>7</sub> antagonists by functional assays.

The most critical contribution of our work is a new strategy to identify novel antipsychotic drugs from the source of FDA-approved drugs. Our findings led to potential repositioned drugs to treat psychotic disorders, especially for treating schizophrenia, since 5-HT<sub>7</sub> malfunction has been implicated in the disorder's etiology. The computational strategy

attested in this study thus could be referred to as QSAR-aided Drug Repurposing. This is a good example of drug repositioning using *in silico* methods, which has garnered attention in light of the advantages they provide in accelerating drug discovery for neglected, rare, and orphan diseases[181]. Our strategy could be used along with other *in silico* approaches (e.g., Bayesian classification methods[182]) to help achieve this goal in the future studies.

## 6.2. Development of algorithm Economic Ratio (ER) as both a cost function and a validation merit for classification QSAR.

In this Chapter, we developed a new fitness metric, the Economic Ratio (ER), and applied this metric for QSAR modeling and Virtual Screening (VS). This was motivated by the fact that rapid growth of High Throughput Screening (HTS) databases concurs with the phenomenon that the majority of tested compounds have inactive annotations. As a consequence, most classification models face the issue of balancing the modeling set or otherwise have to accept the truth that VS results would be biased towards the class encompassing the larger number of compounds in the modeling set. ER, on the other hand, makes use of the imbalances in available datasets and enhances hit enrichment of VS when integrated with Decision Tree methods.

Using five various binary datasets, we tested this ER in two different but complementary ways: 1) as an evaluation metric to assess the prediction performance of QSAR classifiers and 2) as a target function for model building using the Decision Tree algorithm. In the latter case, we used the weighted form of ER to avoid the over-fitting problem. The results showed that successful application of weighted ER was observed on highly imbalanced datasets containing few active compounds, suggesting that ER would be a good complement to traditional classification methods (e.g., *k*NN and SVM). Future efforts

of applying the modified Decision Tree method should focus on such unbalanced datasets, but as an evaluation metric, we highly recommend considering ER in addition to the conventional metrics such as sensitivity, specificity, accuracy, and correct classification rate.

### 6.3. QSARome of the Receptorome: QSAR Modeling of Multiple Ligand Sets Acting at Multiple Receptors

Chapter 4 and 5 focus on the development of QSARome models for predicting chemical binding profiles instead of binding affinity to a specific target. Conventional QSAR studies, including previous works from our laboratory, have dedicated their efforts to relatively simple datasets with a smaller number of similar compounds tested against a specific target. However, as more complex datasets (e.g., multiple biological responses measured for a set of compounds) emerge, we are now capable of systematically evaluating the binding profiles of untested chemicals. In this dissertation, we collected chemicals that were tested on 34 GPCRs associated with therapeutic or side effects of antipsychotic drugs, and developed highly predictive QSARome models after thorough curation of the original raw data. Most models achieved high CCRs when verified by individual 5-fold external cross-validation. Prediction accuracy reached 70.5% when applying the consensus of these models to predict the activity of 13 drugs with known binding affinities to the same 34 GPCRs.

Developed strategy, as proven by various levels of external validation, is very meaningful to help discover antipsychotic drug candidates with favorable binding profiles, which will enhance therapeutic effects but avoid serious side effects. According to that, list of selective compounds (binders for only 1 or 2 GPCRs) was provided for experimental testing. Moreover, to the best of our knowledge, this is the first robust QSAR modeling

exercise that operates on multiple drug targets interacting with a large scope of rigorously curated compounds. In order to make our QSARome practices publicly available to general scientific community, all curated datasets and developed models would be posted on-line on our ChemBench server[180].

One of the foundations of developing successful QSARome models is the rigorous curation of the underlying datasets. The extensive curation effort was verified by the 5-HT<sub>1A</sub> dataset in Chapter 4.  $K_i$  data tested on human tissues and rigorous structural curation protocols proved to be essential for the subsequent success of QSARome modeling. Nevertheless, erroneous annotations are ubiquitous in large databases, and efforts of curation must be maintained continuously. We should strive to control these errors below a certain threshold to minimize their influence on model quality. Although the universal protocol established for the QSARome project seems to meet this goal, we recommend reasonable scrutiny, including manual inspection, of collected data regardless of how large the dataset might be.

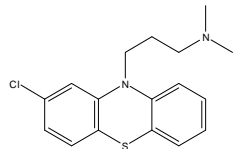
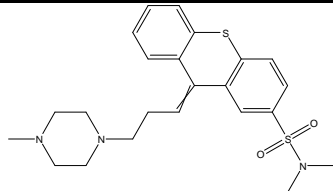
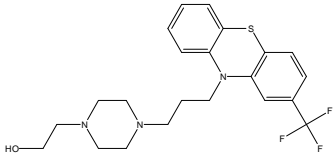
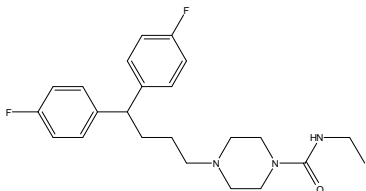
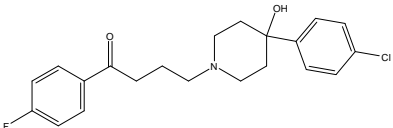
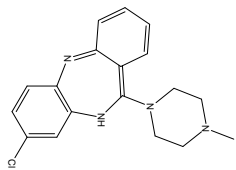
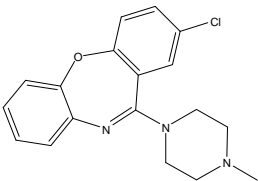
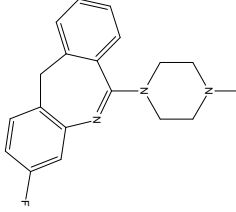
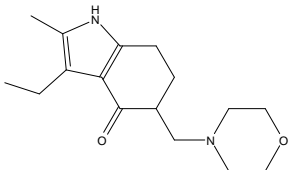
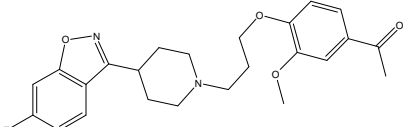
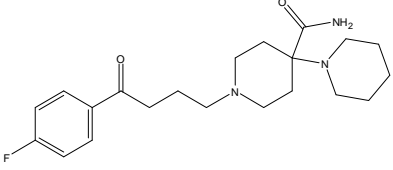
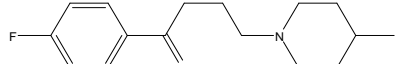
Our earlier work based on relatively simple datasets has helped us gain abundant experience on combinatorial QSAR modeling and rigorous validation schemes. However, the previous QSAR modeling strategies are ill-suited for handling data with multiple endpoints. Thanks to the launch of the ChemBench server[180], it will be much easier to build combinatorial QSARome models with the prompt generation of various types of chemical descriptors (e.g., Dragon, MolConnZ, and CDK[183]) and prepared modeling algorithms (e.g., Random Forest, SVM, and  $k$ NN). ChemBench can thus serve as a convenient one-stop platform for combinatorial QSARome modeling involving various descriptor types and machine learning methods.

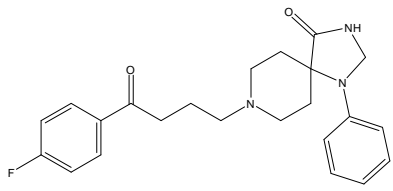
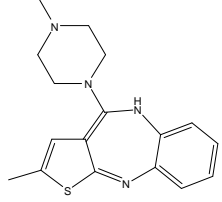
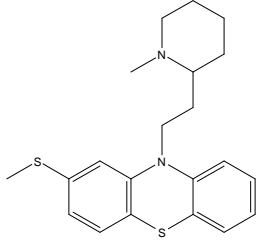
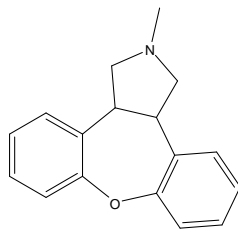
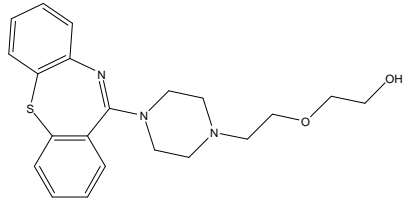
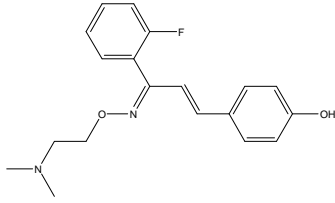
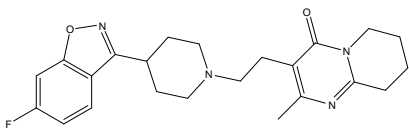
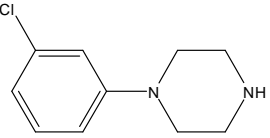
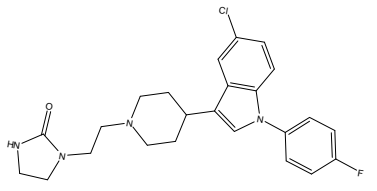
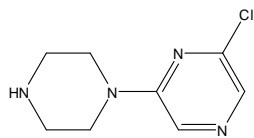
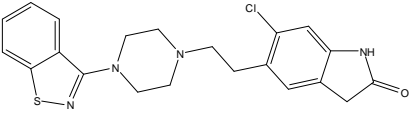
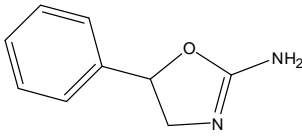
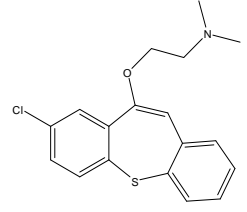
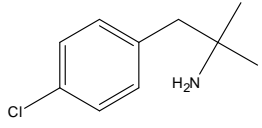
The QSAR field has long been focusing on providing models with good statistics that fit known experimental data. However, we emphasize more on the predictive power of resulting models, e.g., capable of capturing virtual screening hits valid in real experiments. In view of our focus on experimental validation, we collaborate closely with Prof. Bryan Roth at the UNC Department of Pharmacology. Prof. Roth directs the NIMH Psychoactive Drug Screening Program (PDSP), a unique national resource devoted to discovering new treatments for mental illness. His laboratory has assay protocols ready to validate *in silico* predictions made by any QSAR models. In addition, it would be more meaningful if we could apply our approach to find drug-orphan GPCR pairs to help elucidate the functions of orphan GPCRs, which currently account for 40% of known GPCRs. Finally, we should continue to seek collaborations which could provide complementary expertise in hit validation against other protein targets, since our ultimate aim is to broaden the strategy to cover other valuable protein targets in addition to GPCRs.

## Appendices

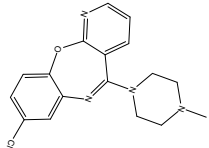
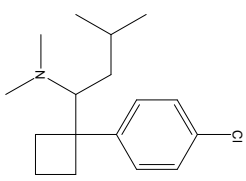
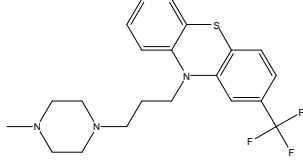
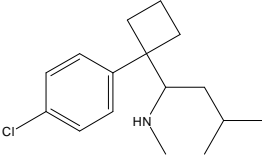
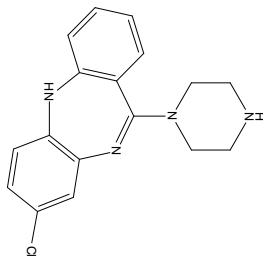
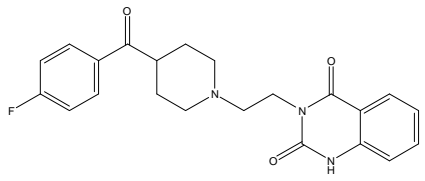
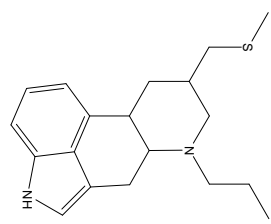
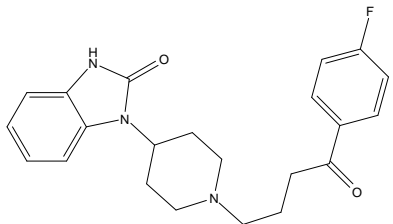
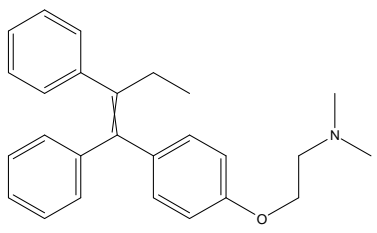
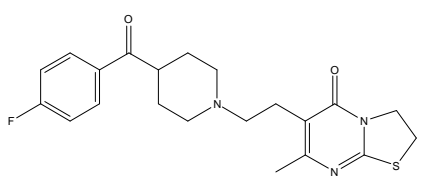
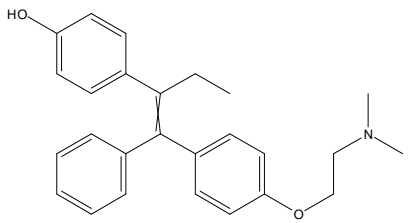
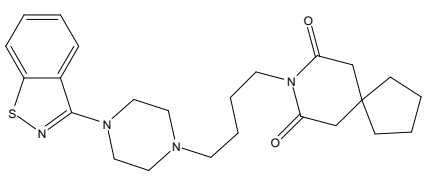
### Appendix I. Chemical structures and measured $pK_i$ values for 5-HT<sub>7</sub> binding affinity of 62 5-HT<sub>7</sub> binders used in QSAR model building and validation.

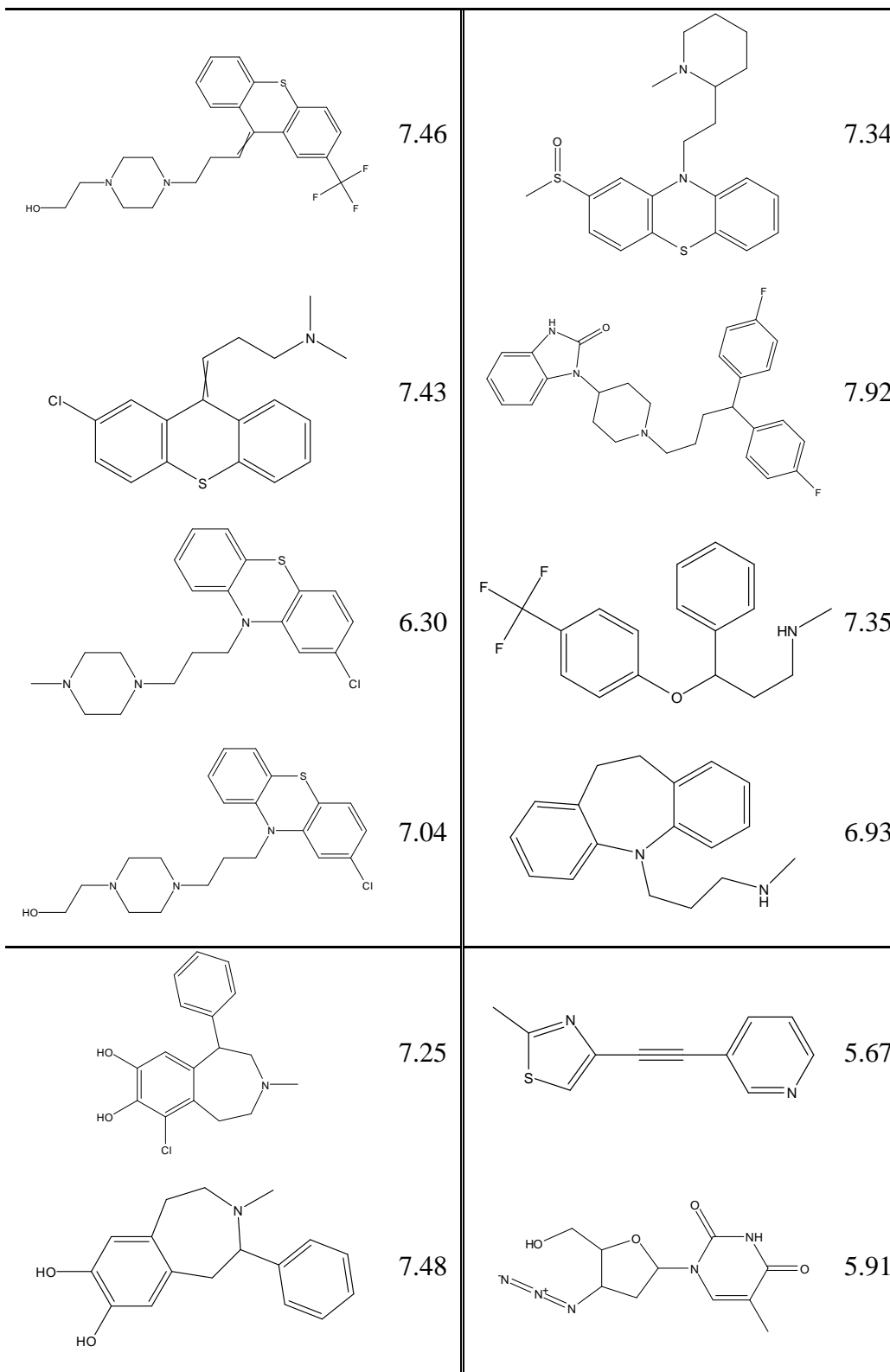
$pK_i$ : experimentally measured activity (binding affinity).

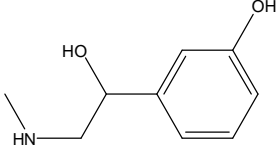
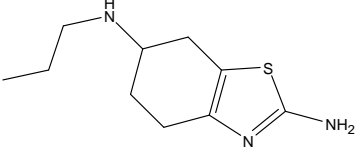
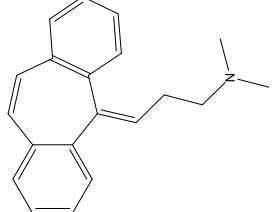
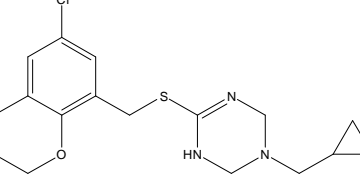
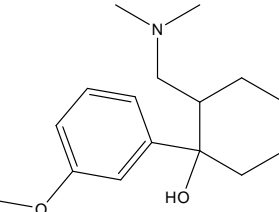
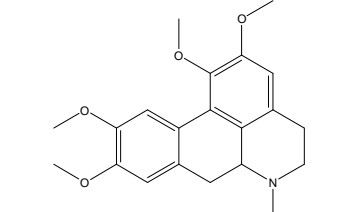
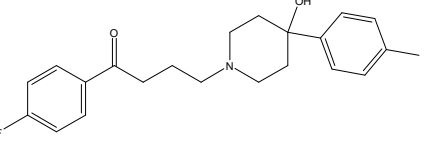
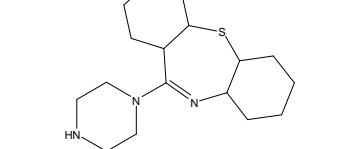
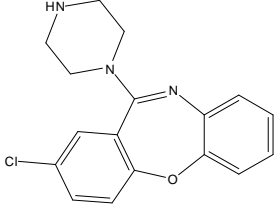
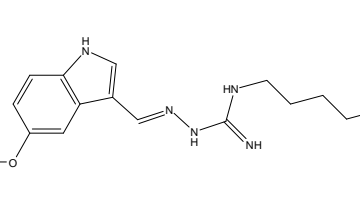
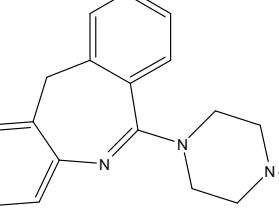
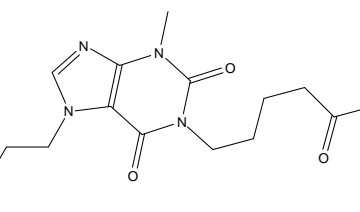
Mol.	$pK_i$	Mol.	$pK_i$
	7.45		7.81
	7.24		5.84
	6.42		7.75
	7.06		7.76
	6.36		7.88
	7.20		6.21

 <p>7.53</p>	 <p>6.98</p>
 <p>7.00</p>	 <p>9.05</p>
 <p>6.51</p>	 <p>5.88</p>
 <p>8.18</p>	 <p>6.79</p>
 <p>7.21</p>	 <p>5.80</p>
 <p>8.18</p>	 <p>7.22</p>
 <p>7.94</p>	 <p>6.28</p>

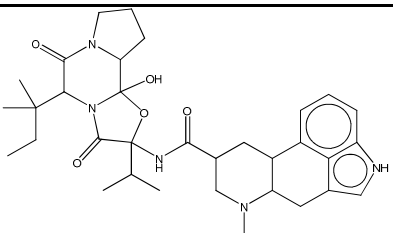
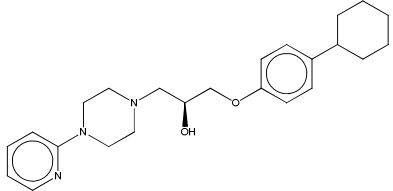
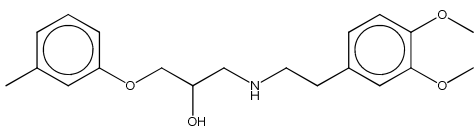
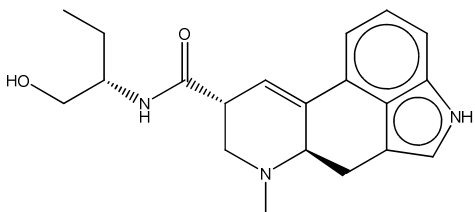
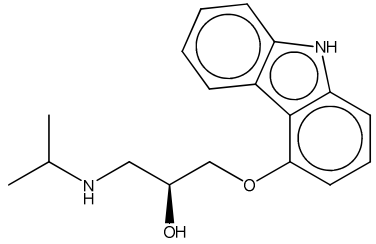
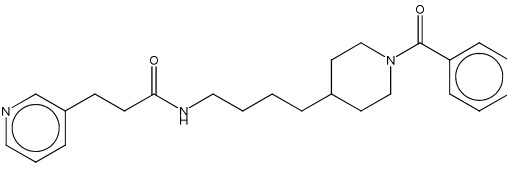
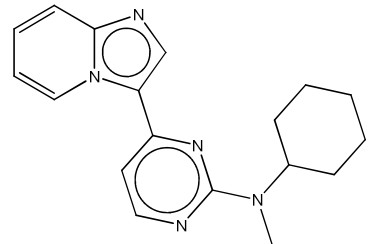


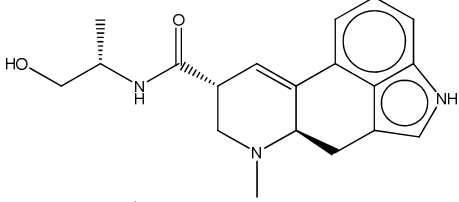
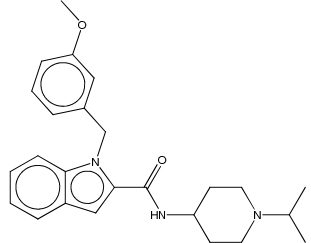
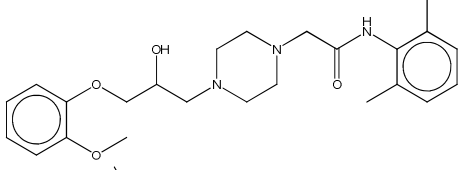
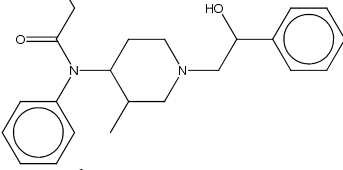
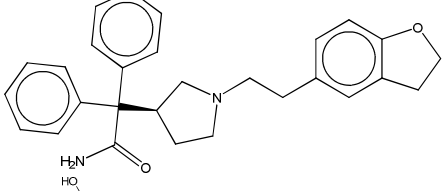
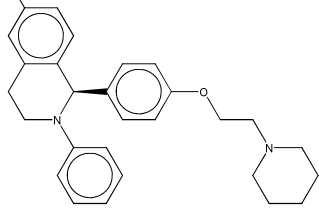
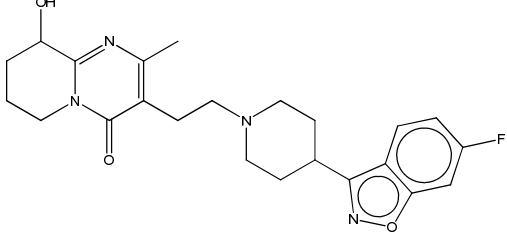
	<p>7.30</p>  <p>5.18</p>
	<p>6.68</p>  <p>5.06</p>
	<p>7.22</p>  <p>6.47</p>
	<p>7.86</p>  <p>8.62</p>
	<p>5.97</p>  <p>8.43</p>
	<p>5.85</p>  <p>8.10</p>



	6.12		5.93
	7.80		5.52
	5.78		7.10
	6.44		7.12
	6.79		6.40
	7.42		5.56

Appendix II. Chemical structures of VS hits yielded by 5-HT<sub>1A</sub> models (DrugBank).

DB_ID	Prediction score	Structure	Name
DB01049	1		Ergoloid
DB08543	1		1-[2-hydroxy-3-(4-cyclohexyl-phenoxy)-propyl]-4-(2-pyridyl)-piperazine
DB01295	1		Bevantolol
DB00353	1		Methylergonovine
DB07543	1		(2S)-1-(9H-Carbazol-4-yloxy)-3-(isopropylamino)propan-2-ol
DB07656	1		n-[4-(1-benzoylpiperidin-4-yl)butyl]-3-pyridin-3-ylpropanamide
DB08023	1		N-cyclohexyl-4-imidazo[1,2-a]pyridin-3-yl-N-methylpyrimidin-2-amine

<b>DB01253</b>	1		<b>Ergonovine</b>
<b>DB07973</b>	1		<b>n-(1-isopropylpiperidin-4-yl)-1-(3-methoxybenzyl)-1h-indole-2-carboxamide</b>
<b>DB00243</b>	1		<b>Ranolazine</b>
<b>DB01570</b>	0.8		<b>Beta-hydroxy-3-methylfentanyl</b>
<b>DB00496</b>	0.6		<b>Darifenacin</b>
<b>DB04471</b>	0.6		<b>2-Phenyl-1-[4-(2-Piperidin-1-Yl-Ethoxy)-Phenyl]-1,2,3,4-Tetrahydro-Isoquinolin-6-Ol</b>
<b>DB01267</b>	0.6		<b>Paliperidone</b>

**Appendix III. Chemical structures and predicted binding targets for untested compounds within the QSARome data matrix.**

Name or ChEMBL Molregno	Targets	Smiles
392109	5-HT <sub>2B</sub>	<chem>O(C(=O)C=1CCCC=1c1cc(N)ccc1)C</chem>
416709	5-HT <sub>1E</sub>	<chem>Clc1cc(ccc1Cl)C=1CCCC=1C(=O)N(OC)C</chem>
331114	5-HT <sub>1E</sub>	<chem>P(Oc1c2c(n(cc2CCN(C)C)C)ccc1)(O)(O)=O</chem>
CARBACHOL	5-HT <sub>1E</sub>	<chem>O(CC[N+](C)(C)C)C(=O)N</chem>
Quercetin	5-HT <sub>2c</sub>	<chem>O1c2c(C(=O)C(O)=C1c1cc(O)c(O)cc1)c(O)cc(O)c2</chem>
Prucalopride	5-HT <sub>3</sub>	<chem>Clc1cc(c2OCCc2c1N)C(=O)NC1CCN(CC1)CCCO</chem>
22826	5-HT <sub>3</sub>	<chem>Oc1cc2CCNc2cc1</chem>
422004	5-HT <sub>3</sub>	<chem>O(C)c1c2c(NC(=O)C(C)(C2=O)c2ccccc2)cc(OC)c1</chem>
7646	5-HT <sub>5</sub>	<chem>NCC(c1ccccc1)c1ccccc1</chem>
438991	5-HT <sub>5</sub>	<chem>Oc1ccccc1CCN1CCCC1</chem>
355471	5-HT <sub>6</sub>	<chem>S(=O)(=O)(Nc1cc(CCN)c(OC)cc1)c1ccccc1</chem>
400598	5-HT <sub>6</sub>	<chem>Clc1ccc(cc1)C=1C2NC(CC2)C=1C(OC)=O</chem>
2724	5-HT <sub>7</sub>	<chem>Oc1cc(ccc1)C1CCCN(C1)CCC</chem>
582997	$\alpha_{1A}$	<chem>Clc1cc(ccc1)C(=O)C(NCCC)C</chem>
583005	$\alpha_{1A}$	<chem>O=C(C(NC(C)(C)C)C)c1ccccc1</chem>
583013	$\alpha_{1A}$	<chem>Clc1cc(ccc1)C(=O)C(NC(C)C)C</chem>
565536	$\alpha_{2A}$	<chem>O(C)c1cc(N2CCN(CCN3CCN(CC3)C(C)C)C2=O)ccc1</chem>
422046	$\beta_2$	<chem>Brc1cc(ccc1OCc1ccccc1)C1(C)C(=O)c2c(NC1=O)cc(Cl)cc2Cl</chem>
ISOPRENALINE	D <sub>1</sub>	<chem>Oc1cc(ccc1O)C(O)CNC(C)C</chem>
365731	D <sub>2</sub>	<chem>Clc1ccc(cc1)C(C(OC)=O)C1NCCCC1</chem>
Cathinone	D <sub>3</sub>	<chem>O=C(C(N)C)c1ccccc1</chem>
Ephedrine	D <sub>3</sub>	<chem>OC(C(NC)C)c1ccccc1</chem>
Methcathinone	D <sub>3</sub>	<chem>O=C(C(NC)C)c1ccccc1</chem>
339502	D <sub>3</sub>	<chem>O(C)c1ccc(cc1)C(=O)C(N1CCCC1)CCC</chem>
L-DOPA	D <sub>3</sub>	<chem>Oc1cc(ccc1O)CC(N)C(O)=O</chem>
423	D <sub>3</sub>	<chem>S(C)c1ccc(cc1)CC(N)C</chem>
582926	D <sub>4</sub>	<chem>Fc1cc(ccc1F)C(=O)C(NC(C)(C)C)C</chem>
299511	D <sub>4</sub>	<chem>O1c2c(cccc2)C(=O)CC1CN1CCN(CC1)Cc1ccccc1</chem>
392134	H <sub>1</sub>	<chem>O(C(=O)C=1CCCCC=1c1cc([NH+](O)[O-])ccc1)C</chem>
499031	H <sub>2</sub>	<chem>FCCOc1ccc(C(=O)C2CCN(CC2)CCc2ccc(OC)cc2)c1OC</chem>

<b>439047</b>	H <sub>2</sub>	s1cccc1CCN1CCCC1
<b>D-cystine</b>	H <sub>3</sub>	S(SCC(N)C(O)=O)CC(N)C(O)=O
<b>D-methionine</b>	H <sub>3</sub>	S(CCC(N)C(O)=O)C
<b>L-cysteine</b>	H <sub>3</sub>	SCC(N)C(O)=O
<b>L-cysteic-acid</b>	H <sub>3</sub>	S(O)=(O)=(O)CC(N)C(O)=O
<b>338970</b>	H <sub>3</sub>	O(C(=O)c1ccc(cc1)C(=O)C(N1CCCC1)CCC)C
<b>206670</b>	H <sub>3</sub>	O(CCC1nc[nH]c1)c1ccc(cc1)C(=O)C
<b>438654</b>	H <sub>3</sub>	O(C)c1ccc(cc1)CCN1CCCC1
<b>Naproxen</b>	M <sub>1</sub>	O(C)c1cc2c(cc(cc2)C(C(O)=O)C)cc1
<b>ibuprofen</b>	M <sub>1</sub>	OC(=O)C(C)c1ccc(cc1)CC(C)C
<b>339144</b>	M <sub>1</sub>	Clc1cc(ccc1Cl)C(=O)C(NCCCC)CCC
<b>392103</b>	M <sub>1</sub>	O(C)c1ccc(cc1)C=1CCCC=1C(OC)=O
<b>392105</b>	M <sub>1</sub>	FC(F)(F)Oc1ccc(cc1)C=1CCCC=1C(OC)=O
<b>392132</b>	M <sub>1</sub>	Oc1ccc(cc1)C=1CCCC=1C(OC)=O
<b>392380</b>	M <sub>1</sub>	O(C(=O)C1CCCCC1c1ccccc1)C
<b>336399</b>	M <sub>1</sub>	BrCC(C(OCC)=O)c1ccccc1
<b>336505</b>	M <sub>1</sub>	O1CC(CC1=O)c1ccccc1
<b>336648</b>	M <sub>1</sub>	BrCCCC(C(OCC)=O)c1ccccc1
<b>400589</b>	M <sub>1</sub>	O(C(=O)C=1C2CC(CC2)C=1c1ccccc1)C
<b>PBR28</b>	M <sub>4</sub>	O(c1ccccc1N(Cc1ccccc1OC)C(=O)C)c1ccccc1
<b>Rutin</b>	M <sub>5</sub>	O1C(COC2OC(C)C(O)C(O)C2O)C(O)C(O)C(O)C1OC1=C(O)c2c(C1=O)c(O)cc(O)c2)c1cc(O)c(O)cc1
<b>422029</b>	M <sub>5</sub>	Clc1c2c(NC(=O)C(C)(C2=O)c2ccc(Cl)cc2)cc(Cl)c1
<b>438479</b>	M <sub>5</sub>	n1ccccc1CCN1CCCC1C
<b>METHYLPHENIDATE</b>	DAT	O(C(=O)C(C1NCCCC1)c1ccccc1)C
<b>273894</b>	SERT	N(CCC)(CCC)C1CCC(CC1)=CC#N
<b>66613</b>	SERT	O1c2cc(ccc2OC1)CC(NC)C
<b>TRAZODONE</b>	5-HT <sub>2A</sub> , H <sub>2</sub>	Clc1cc(N2CCN(CC2)CCCN2N=C3N(C=CC=C3)C2=O)ccc1
<b>nimesulide</b>	β <sub>2</sub> , D <sub>3</sub>	S(=O)(=O)(Nc1ccc([NH+](O)[O-])cc1O)c1ccccc1)C
<b>PSILOCYBIN</b>	D <sub>3</sub> , M <sub>4</sub>	P(Oc1c2c([nH]cc2CCN(C)C)ccc1)(O)(O)=O
<b>PARECOXIB</b>	D <sub>3</sub> , H <sub>4</sub>	S(=O)(=O)(NC(=O)CC)c1ccc(cc1)-c1c(noc1C)-c1ccccc1
<b>CELECOXIB</b>	β <sub>2</sub> , D <sub>3</sub>	S(=O)(=O)(N)c1ccc(-n2nc(cc2-c2ccc(cc2)C)C(F)(F)F)cc1
<b>498962</b>	5-HT <sub>2A</sub> , H <sub>2</sub>	Fc1ccc(cc1)CCN1CCC(CC1)C(=O)c1ccccc1OCCF)c1OC
<b>498964</b>	β <sub>2</sub> , H <sub>2</sub>	FCCOc1ccc(C(O)C2CCN(CC2)CCc2ccc(cc2)C)c1OC

<b>513010</b>	5-HT <sub>3</sub> , H <sub>2</sub>	<chem>O(C(=O)c1ccccc1)C1CC2N(C(CC2)C1C(OC)=O)C</chem>
<b>499028</b>	5-HT <sub>2C</sub> , M <sub>4</sub>	<chem>FCCOC1CCCC(C(=O)C2CCN(CC2)CCc2ccc(cc2)C)c1OC</chem>
<b>Diclofenac</b>	5-HT <sub>2C</sub> , M <sub>3</sub>	<chem>Clc1cccc(Cl)c1Nc1ccccc1CC(O)=O</chem>
<b>Acetylsalicylic-acid</b>	$\alpha_{2B}$ , H <sub>2</sub>	<chem>O(C(=O)C)c1ccccc1C(O)=O</chem>
<b>SalvinorinA</b>	5-HT <sub>5</sub> , $\beta_2$	<chem>O1C(CC2(C3C(CCC2C1=O)(C)C(CC(OC(=O)C)C3=O)C(OC)=O)C)c1ccoc1</chem>
<b>421822</b>	5-HT <sub>2C</sub> , M <sub>1</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2ccccc2)cc(Cl)c1</chem>
<b>422032</b>	5-HT <sub>2C</sub> , H <sub>3</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2cc(O)ccc2)cc(Cl)c1</chem>
<b>422034</b>	5-HT <sub>6</sub> , M <sub>5</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2ccc(O)cc2)cc(Cl)c1</chem>
<b>Hyperoside</b>	5-HT <sub>6</sub> , M <sub>3</sub>	<chem>O1C(CO)C(O)C(O)C(O)C1OC1=C(OC=2C(=C1O)C(O)=CC(=O)C=2)c1cc(O)c(O)cc1</chem>
<b>Quercitrin</b>	5-HT <sub>6</sub> , H <sub>4</sub>	<chem>O1C(C)C(O)C(O)C(O)C1OC1=C(OC=2C(=C1O)C(O)=CC(=O)C=2)c1cc(O)c(O)cc1</chem>
<b>498963</b>	5-HT <sub>2C</sub> , H <sub>1</sub>	<chem>FCCOC1CCCC(C(O)C2CCN(CC2)CCc2ccc([NH+](O)[O-])cc2)c1OC</chem>
<b>499029</b>	5-HT <sub>2C</sub> , H <sub>1</sub>	<chem>FCCOC1CCCC(C(O)C2CCN(CC2)CCc2ccc(OC)cc2)c1OC</chem>
<b>499030</b>	H <sub>2</sub> , H <sub>3</sub>	<chem>FCCOC1CCCC(C(=O)C2CCN(CC2)CCc2ccc([NH+](O)[O-])cc2)c1OC</chem>
<b>METHAMPHETAMINE</b>	$\beta_2$ , H <sub>2</sub>	<chem>N(C(Cc1ccccc1)C)C</chem>
<b>Desvenlafaxine</b>	H <sub>2</sub> , H <sub>3</sub>	<chem>OC1(CCCCC1)C(CN(C)C)c1ccc(O)cc1</chem>
<b>SULPIRIDE</b>	$\beta_2$ , SERT	<chem>S(=O)(=O)(N)c1cc(C(=O)NCC2N(CCC2)CC)c(OC)cc1</chem>
<b>258615</b>	D <sub>2</sub> , D <sub>3</sub>	<chem>OC(C(NC(C)(C)C)C)c1ccccc1</chem>
<b>273862</b>	$\alpha_{2C}$ , D <sub>1</sub>	<chem>N(CCC)(CCC)C1CCC(=CC1)C#N</chem>
<b>392093</b>	5-HT <sub>2A</sub> , M <sub>4</sub>	<chem>Clc1cc(ccc1Cl)C=1CCNCC=1C(OCC)=O</chem>
<b>392117</b>	D <sub>2</sub> , D <sub>3</sub>	<chem>Clc1cc(ccc1)C=1CCCC=1C(OC)=O</chem>
<b>392126</b>	D <sub>5</sub> , M <sub>1</sub>	<chem>O(C(=O)C=1CCCCC=1c1ccccc1)C</chem>
<b>392379</b>	5-HT <sub>1D</sub> , D <sub>5</sub>	<chem>FC(F)(F)Oc1ccc(cc1)C1CCCCC1C(OC)=O</chem>
<b>392385</b>	M <sub>1</sub> , M <sub>2</sub>	<chem>FC(F)(F)c1ccc(cc1)C1CCCCC1C(OC)=O</chem>
<b>416696</b>	M <sub>1</sub> , M <sub>2</sub>	<chem>Clc1cc(ccc1Cl)C=1CCCC=1C(=O)N(CC)CC</chem>
<b>416708</b>	M <sub>1</sub> , M <sub>2</sub>	<chem>Clc1cc(ccc1Cl)C=1CCCC=1C(=O)NCCF</chem>
<b>416717</b>	M <sub>1</sub> , M <sub>2</sub>	<chem>Clc1cc(ccc1Cl)C=1CCCC=1C(OCC(F)(F)F)=O</chem>
<b>582972</b>	D <sub>3</sub> , M <sub>2</sub>	<chem>Clc1cc(ccc1)C(=O)C(N(C)C)C</chem>
<b>582990</b>	M <sub>1</sub> , M <sub>2</sub>	<chem>Clc1cc(ccc1)C(=O)C(N1CCCCC1)C</chem>
<b>583011</b>	D <sub>3</sub> , D <sub>4</sub>	<chem>Fc1cc(ccc1)C(=O)C(NC(C)(C)C)C</chem>
<b>583012</b>	5-HT <sub>7</sub> , $\alpha_{1A}$	<chem>Br1cc(ccc1)C(=O)C(NC(C)(C)C)C</chem>
<b>583021</b>	$\alpha_{1A}$ , $\alpha_{2C}$	<chem>Clc1cc(ccc1)C(=O)C(NC1CC1)C</chem>
<b>583030</b>	5-HT <sub>7</sub> , D <sub>1</sub>	<chem>Clc1cc(ccc1)C(=O)C(NC1CCCC1)C</chem>



<b>583041</b>	$\alpha_{1A}$ , $\alpha_{2C}$	<chem>Clc1ccc(cc1)C(=O)C(NC(C)(C)C)C</chem>
<b>Iproniazid</b>	$\alpha_{1A}$ , $M_5$	<chem>O=C(NNC(C)C)c1ccncc1</chem>
<b>NPA</b>	5-HT <sub>7</sub> , $\alpha_{2C}$	<chem>OC(=O)c1cccc1C(=O)Nc1c2c(ccc1)cccc2</chem>
<b>229225</b>	5-HT <sub>2A</sub> , $M_4$	<chem>Fc1c2CCC(NC(C)C)Cc2ccc1</chem>
<b>520213</b>	$D_4$ , $M_3$	<chem>O1C2C(N(CC1)CCC)CCc1c2cc(cc1)C(O)=O</chem>
<b>DMA</b>	5-HT <sub>5</sub> , $D_4$	<chem>O(C)c1ccc(OC)cc1CC(N)C</chem>
<b>MEM</b>	$\alpha_{2C}$ , DAT	<chem>C1COCCOC</chem>
<b>336434</b>	5-HT <sub>1E</sub> , 5-HT <sub>6</sub>	<chem>BrC(CCCC(OCC)=O)c1cccc1</chem>
<b>336442</b>	5-HT <sub>2A</sub> , $M_5$	<chem>O1C(CCCC1=O)c1cccc1</chem>
<b>10383</b>	$D_4$ , $M_1$	<chem>[nH]1cc(nc1)CC(N)C</chem>
<b>10519</b>	5-HT <sub>2C</sub> , $M_1$	<chem>S(CCCN(C)C)C(N)=N</chem>
<b>549495</b>	$D_4$ , $M_1$	<chem>S(CCNC(N)=N)C(N)=N</chem>
<b>alpha-methylhistamine</b>	$D_4$ , $M_3$	<chem>[nH]1cncc1CC(N)C</chem>
<b>35034</b>	5-HT <sub>3</sub> , $M_4$	<chem>O(C)c1cccc1N1CCN(CC1)CCNC(=O)c1cc(OC)ccc1</chem>
<b>87552</b>	5-HT <sub>3</sub> , $D_2$	<chem>o1c2cc(NC(=O)Nc3c4ncccc4nc3)ccc2nc1C</chem>
<b>120076</b>	5-HT <sub>1D</sub> , $H_4$	<chem>Brc1cc(C(=O)NCC2N(CCC2)CC)c(OC)c(OC)c1</chem>
<b>156767</b>	5-HT <sub>1D</sub> , $\alpha_{1A}$	<chem>O(C)c1cc(ccc1OC)C(=O)CCC(=O)N1CCN(CC1)C(CC)CC</chem>
<b>231436</b>	5-HT <sub>7</sub> , DAT	<chem>N(CCN1CCc2c1cccc2)(C)C</chem>
<b>296993</b>	5-HT <sub>1A</sub> , $D_4$	<chem>Brc1cc(S(=O)(=O)N2CCC(N3c4c(COC3=O)cccc4C)CC2)c(OC)cc1</chem>
<b>299512</b>	$\alpha_{1A}$ , $H_2$	<chem>O1c2c(cccc2)C(=O)C=C1CN1CCN(CC1)Cc1cccc1</chem>
<b>307784</b>	5-HT <sub>6</sub> , $M_1$	<chem>O=C(Nc1cc(ccc1)C)CN1CCN(CC1)c1ccncc1</chem>
<b>317542</b>	$D_2$ , $H_2$	<chem>S(=O)(=O)(Nc1cc(N)cc(OC)c1)c1ccc(N)cc1</chem>
<b>325220</b>	5-HT <sub>1E</sub> , $H_4$	<chem>FCCc1cc(C(=O)NCC2N(CCC2)CC)c(OC)c(OC)c1</chem>
<b>332586</b>	5-HT <sub>1E</sub> , $\alpha_{2C}$	<chem>S(=O)(=O)(n1c2c(cc1)c([NH+](O)[O-])cc([NH+](O)[O-])c2)c1ccc(N)cc1</chem>
<b>350922</b>	5-HT <sub>6</sub> , $M_3$	<chem>O(C)c1cccc1N1CCN(CC1)CCCN1C(=O)C(NC(=O)C)CC1=O</chem>
<b>355470</b>	$H_4$ , $M_3$	<chem>S(=O)(=O)(Nc1cc(ccc1)CCN)c1cccc1</chem>
<b>355501</b>	$\beta_2$ , $H_2$	<chem>S(=O)(=O)(Nc1cc(ccc1)CCCN)c1cccc1</chem>
<b>383726</b>	5-HT <sub>1D</sub> , $H_2$	<chem>Clc1cccc1COc1ccc(OCCCN2CCOCC2)cc1C(=O)N(C)C</chem>
<b>383728</b>	$M_1$ , $M_3$	<chem>O1CCN(CC1)CCCOc1cc(C(=O)N(C)C)c(OCc2cccc2OC)cc1</chem>
<b>392106</b>	$M_1$ , $M_2$	<chem>S(=O)(=O)(C)c1ccc(cc1)C=1CCCC=1C(OC)=O</chem>
<b>400595</b>	5-HT <sub>1D</sub> , 5-HT <sub>7</sub>	<chem>Clc1cc(Cl)ccc1C=1C2CC(CC2)C=1C(OC)=O</chem>
<b>408619</b>	5-HT <sub>3</sub> , $H_3$	<chem>n1(c2c(cccc2)cc1)Cc1ccc(N)cc1</chem>

<b>421941</b>	$\beta_2$ , H <sub>4</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2cccc2)cc(OC)c1</chem>
<b>422033</b>	H <sub>2</sub> , H <sub>4</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2ccc(OC)cc2)cc(Cl)c1</chem>
<b>422043</b>	H <sub>2</sub> , H <sub>4</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2cc([NH+](O)[O-])c(OC)cc2)cc(Cl)c1</chem>
<b>422044</b>	5-HT <sub>6</sub> , H <sub>4</sub>	<chem>Clc1c2c(NC(=O)C(C)(C2=O)c2cc([NH+](O)[O-])c(O)cc2)cc(Cl)c1</chem>
<b>422048</b>	M <sub>5</sub> , DAT	<chem>Br1cc(ccc1O)C1(C)C(=O)c2c(NC1=O)cc(Cl)cc2Cl</chem>
<b>438478</b>	$\alpha_{2A}$ , DAT	<chem>n1cccc1CCN1CCCC1</chem>
<b>438651</b>	5-HT <sub>7</sub> , H <sub>4</sub>	<chem>O(C)c1cc(ccc1)CCN1CCCC1</chem>
<b>438652</b>	5-HT <sub>7</sub> , H <sub>4</sub>	<chem>O(C)c1cc(ccc1)CCN1CCCC1C</chem>
<b>438656</b>	5-HT <sub>1E</sub> , 5-HT <sub>5</sub>	<chem>O(C)c1cccc1CCN1CCCC1C</chem>
<b>438701</b>	$\alpha_{1A}$ , H <sub>4</sub>	<chem>O(C)c1cccc1CCN1CCCC1</chem>
<b>438703</b>	D <sub>1</sub> , H <sub>4</sub>	<chem>N1(CCCC1C)CCc1cc(ccc1)C</chem>
<b>438990</b>	D <sub>1</sub> , D <sub>4</sub>	<chem>O[NH+](O)c1ccc(cc1)CCN1CCCC1C</chem>
<b>438996</b>	H <sub>4</sub> , M <sub>3</sub>	<chem>O[NH+](O)c1ccc(cc1)CCN1CCCC1</chem>
<b>439048</b>	$\alpha_{1A}$ , H <sub>4</sub>	<chem>s1cccc1CCN1CCCC1C</chem>
<b>501019</b>	$\alpha_{2A}$ , H <sub>1</sub>	<chem>O(C(=O)c1[nH]c2c(cc(cc2)C(O)=O)c1)CC</chem>
<b>520792</b>	5-HT <sub>2c</sub> , H <sub>1</sub>	<chem>Fc1cc(OC(F)(F)F)ccc1-c1ccc(cc1)CN1CC(OC1=O)C</chem>
<b>520864</b>	5-HT <sub>3</sub> , D <sub>4</sub>	<chem>Fc1cc(OC(F)(F)F)ccc1-c1ncc(cc1)CN1CC(OC1=O)C</chem>
<b>626640</b>	5-HT <sub>3</sub> , 5-HT <sub>6</sub>	<chem>O(C1CC(N(C1)C(=O)C)C(=O)N1CCCN(CC1)C1CCC1)c1cccnc1</chem>
<b>647925</b>	5-HT <sub>1E</sub> , 5-HT <sub>5</sub>	<chem>O(c1cc(ccc1)C#N)c1ncc(cc1)C(=O)N1CCN(CC1)C(C)C</chem>
<b>648035</b>	5-HT <sub>2A</sub> , H <sub>2</sub>	<chem>Fc1ccc(Oc2ncc(cc2)C(=O)N2CCCN(CC2)C)cc1</chem>
<b>DMT</b>	$\beta_2$ , D <sub>3</sub>	<chem>[nH]1cc(c2c1cccc2)CCN(C)C</chem>

## References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL: How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* 2010, 9:203-214.
2. Bender A: Databases: Compound bioactivities go public. *Nat Chem Biol* 2010, 6:309.
3. Ki determinations were generously provided by the National Institute of Mental Health's Psychoactive Drug Screening Program, Contract # HHSN-271-2008-00025-C (NIMH PDSP). The NIMH PDSP is directed by Bryan L. Roth MD, PhD at the University of North Carolina at Chapel Hill and Project Officer Jamie Driscoll at NIMH, Bethesda MD, USA. (<http://pdsp.med.unc.edu/>). [date unknown], [no volume].
4. Oprea TI, Tropsha A: Target, chemical and bioactivity databases - integration is key. *Drug Discovery Today: Technologies* 2006, 3:357-365.
5. de Cerqueira Lima P, Golbraikh A, Oloff S, Xiao Y, Tropsha A: Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *Journal of Chemical Information and Modeling* 2006, 46:1245-1254.
6. Shen M, Beguin C, Golbraikh A, Stables JP, 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.
7. Tropsha A, Zheng W: Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. *Curr.Pharm.Des* 2001, 7:599-612.
8. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A: QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm.Res.* 2008, 25:1902-1914.
9. Tropsha A, Gramatica P, Gombar VK: The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR & Combinatorial Science* 2003, 22:69-77.
10. Kroeze WK, Sheffler DJ, Roth BL: G-protein-coupled receptors at a glance. *Journal of Cell Science* 2003, 116:4867 -4869.
11. Fredriksson R, Lagerström MC, Lundin L-G, Schiöth HB: The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.* 2003, 63:1256-1272.
12. Civelli O, Saito Y, Wang Z, Nothacker H-P, Reinscheid RK: Orphan GPCRs and their ligands. *Pharmacology & Therapeutics* 2006, 110:525-532.

13. Robas N, O'Reilly M, Katugampola S, Fidock M: Maximizing serendipity: strategies for identifying ligands for orphan G-protein-coupled receptors. *Curr Opin Pharmacol* 2003, 3:121-126.
14. Flower DR: Modelling G-protein-coupled receptors for drug design. *Biochim. Biophys. Acta* 1999, 1422:207-234.
15. Nonell-Canals A, Mestres J: In Silico Target Profiling of One Billion Molecules. *Molecular Informatics* 2011, 30:405-409.
16. Mark A. Johnson, Gerald M. Maggiora: *Concepts and applications of molecular similarity [Internet]*. John Wiley & Sons, New York; 1990.
17. Hansch C, Fujita T:  $\rho$ - $\sigma$ - $\pi$  Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society* 1964, 86:1616-1626.
18. Tropsha A: Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 2010, 29:476-488.
19. Zhu PJ, Zheng W, Auld DS, Jadhav A, Macarthur R, Olson KR, Peng K, Dotimas H, Austin CP, Inglese J: A miniaturized glucocorticoid receptor translocation assay using enzymatic fragment complementation evaluated with qHTS. *Comb. Chem. High Throughput Screen.* 2008, 11:545-559.
20. Xia M, Guo V, Huang R, Inglese J, Nirenberg M, Austin CP: A Cell-based beta-Lactamase Reporter Gene Assay for the CREB Signaling Pathway. *Curr Chem Genomics* 2009, 3:7-12.
21. Johnson RL, Huang R, Jadhav A, Southall N, Wichterman J, MacArthur R, Xia M, Bi K, Printen J, Austin CP, et al.: A quantitative high-throughput screen for modulators of IL-6 signaling: a model for interrogating biological networks using chemical libraries. *Mol Biosyst* 2009, 5:1039-1050.
22. Sun H, Veith H, Xia M, Austin CP, Huang R: Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data [Internet]. *Journal of Chemical Information and Modeling* 2011, doi:10.1021/ci200311w.
23. HANSCH C, MALONEY PP, FUJITA T, MUIR RM: Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* 1962, 194:178-180.
24. Dietrich SW, Dreyer ND, Hansch C, Bentley DL: Confidence interval estimators for parameters associated with quantitative structure-activity relationships. *J. Med. Chem.* 1980, 23:1201-1205.
25. Hadjipavlou-Litina D, Hansch C: Quantitative Structure-Activity Relationships of the Benzodiazepines. A Review and Reevaluation. *Chemical Reviews* 1994, 94:1483-1505.

26. Hansch C, Kurup A, Garg R, Gao H: Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms. *Chem. Rev.* 2001, 101:619-672.
27. Hansch C, Leo A, Mekapati SB, Kurup A: QSAR and ADME. *Bioorganic & Medicinal Chemistry* 2004, 12:3391-3400.
28. Kubinyi H: QSAR and 3D QSAR in drug design Part 1: methodology. *Drug Discovery Today* 1997, [no volume].
29. Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and regression trees*. Chapman & Hall; 1993.
30. Breiman L: Random Forests. *Machine Learning* 2001, 45:5-32.
31. Devillers J: A new strategy for using supervised artificial neural networks in QSAR. *SAR QSAR Environ Res* 2005, 16:433-442.
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. Vapnik V.: *The Nature of Statistical Learning Theory*. Springer; 2000.
34. Marron JS, Todd MJ, Ahn J: Distance Weighted Discrimination. *Journal of the American Statistical Association* 2007, 102:1267-1271.
35. Faulon J-L, Bender A: *Handbook of Chemoinformatics Algorithms*. Chapman and Hall/CRC; 2010.
36. Rücker C, Rücker G, Meringer M:  $\gamma$ -Randomization and its variants in QSPR/QSAR. *J Chem Inf Model* 2007, 47:2345-2357.
37. Topliss JG, Edwards RP: Chance factors in studies of quantitative structure-activity relationships. *Journal of Medicinal Chemistry* 1979, 22:1238-1244.
38. Golbraikh A, Tropsha A: Beware of  $q^2$ ! *J.Mol.Graph.Model.* 2002, 20:269-276.
39. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, 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.
40. Isaksson A, Wallman M, Göransson H, Gustafsson MG: Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters* 2008, 29:1960-1965.
41. 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.
42. Chawla N: DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW. In *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*. Springer; 2010:853-867.
  43. Hedlund PB, Huitron-Resendiz S, Henriksen SJ, Sutcliffe JG: 5-HT7 receptor inhibition and inactivation induce antidepressant like behavior and sleep pattern. *Biol Psychiatry* 2005, 58:831-837.
  44. Thomas DR, Hagan JJ: 5-HT7 receptors. *Curr Drug Targets CNS Neurol Disord* 2004, 3:81-90.
  45. Vanhoenacker P, Haegeman G, Leysen JE: 5-HT7 receptors: current knowledge and future prospects. *Trends Pharmacol Sci* 2000, 21:70-77.
  46. Ballaz SJ, Akil H, Watson SJ: The 5-HT7 receptor: role in novel object discrimination and relation to novelty-seeking behavior. *Neuroscience* 2007, 149:192-202.
  47. Shen Y, Monsma FJ, Metcalf MA, Jose PA, Hamblin MW, Sibley DR: Molecular cloning and expression of a 5-hydroxytryptamine<sub>7</sub> serotonin receptor subtype. *J Biol Chem* 1993, 268:18200-18204.
  48. Mengod G, Vilaro MT, Raurich A, Lopez-Gimenez JF, Cortes R, Palacios JM: 5-HT receptors in mammalian brain: receptor autoradiography and in situ hybridization studies of new ligands and newly identified receptors. *Histochem.J.* 1996, 28:747-758.
  49. Varnas K, Thomas DR, Tupala E, Tiihonen J, Hall H: Distribution of 5-HT7 receptors in the human brain: a preliminary autoradiographic study using [3H]SB-269970. *Neurosci.Lett.* 2004, 367:313-316.
  50. Prins NH, Briejer MR, Van Bergen PJ, Akkermans LM, Schuurkes JA: Evidence for 5-HT7 receptors mediating relaxation of human colonic circular smooth muscle. *Br J Pharmacol* 1999, 128:849-852.
  51. Mnie-Filali O, Lambas-Senas L, Zimmer L, Haddjeri N: 5-HT7 receptor antagonists as a new class of antidepressants. *Drug News Perspect* 2007, 20:613-618.
  52. Gray JA, Roth BL: Molecular targets for treating cognitive dysfunction in schizophrenia. *Schizophr Bull.* 2007, 33:1100-1119.
  53. Nandam LS, Jhaveri D, Bartlett P: 5-HT7, neurogenesis and antidepressants: a promising therapeutic axis for treating depression. *Clin Exp Pharmacol Physiol* 2007, 34:546-551.

54. Manuel-Apolinar L, Meneses A: 8-OH-DPAT facilitated memory consolidation and increased hippocampal and cortical cAMP production. *Behav Brain Res* 2004, 148:179-184.
55. Roberts AJ, Krucker T, Levy CL, Slanina KA, Sutcliffe JG, Hedlund PB: Mice lacking 5-HT receptors show specific impairments in contextual learning. *Eur J Neurosci* 2004, 19:1913-1922.
56. Tokarski K, Zahorodna A, Bobula B, Hess G: 5-HT7 receptors increase the excitability of rat hippocampal CA1 pyramidal neurons. *Brain Res* 2003, 993:230-234.
57. Agosti RM: 5HT1F- and 5HT7-receptor agonists for the treatment of migraines. *CNS Neurol Disord Drug Targets* 2007, 6:235-237.
58. Centurion D, Glusa E, Sanchez-Lopez A, Valdivia LF, Saxena PR, Villalon CM: 5-HT7, but not 5-HT2B, receptors mediate hypotension in vagosympathectomized rats. *Eur J Pharmacol* 2004, 502:239-242.
59. Zou BC, Dong L, Wang Y, Wang SH, Cao MB: Expression and role of 5-HT7 receptor in brain and intestine in rats with irritable bowel syndrome. *Chin Med J (Engl.)* 2007, 120:2069-2074.
60. Denhart DJ, Purandare AV, Catt JD, King HD, Gao A, Deskus JA, Poss MA, Stark AD, Torrente JR, Johnson G, et al.: Diaminopyrimidine and diaminopyridine 5-HT7 ligands. *Bioorg Med Chem Lett* 2004, 14:4249-4252.
61. Forbes IT, Cooper DG, Dodds EK, Douglas SE, Gribble AD, Ife RJ, Lightfoot AP, Meeson M, Campbell LP, Coleman T, et al.: Identification of a novel series of selective 5-HT7 receptor antagonists. *Bioorg Med Chem Lett* 2003, 13:1055-1058.
62. Mattson RJ, Denhart DJ, Catt JD, Dee MF, Deskus JA, Ditta JL, Epperson J, King H, Gao A, Poss MA, et al.: Aminotriazine 5-HT7 antagonists. *Bioorg Med Chem Lett* 2004, 14:4245-4248.
63. Medina RA, Sallander J, Benhamu B, Porras E, Campillo M, Pardo L, Lopez-Rodriguez ML: Synthesis of new serotonin 5-HT7 receptor ligands. Determinants of 5-HT7/5-HT1A receptor selectivity. *J Med Chem* 2009, 52:2384-2392.
64. Na YH, Hong SH, Lee JH, Park WK, Baek DJ, Koh HY, Cho YS, Choo H, Pae AN: Novel quinazolinone derivatives as 5-HT7 receptor ligands. *Bioorg Med Chem* 2008, 16:2570-2578.
65. Parikh V, Welch WM, Schmidt AW: Discovery of a series of (4,5-dihydroimidazol-2-yl)-biphenylamine 5-HT7 agonists. *Bioorg Med Chem Lett* 2003, 13:269-271.
66. Peters JU, Lubbers T, Alanine A, Kolczewski S, Blasco F, Steward L: Cyclic guanidines as dual 5-HT5A/5-HT7 receptor ligands: structure-activity relationship elucidation. *Bioorg Med Chem Lett* 2008, 18:256-261.

67. Pouzet B: SB-258741: a 5-HT7 receptor antagonist of potential clinical interest. *CNS Drug Rev* 2002, 8:90-100.
68. Volk B, Barkoczy J, Hegedus E, Udvari S, Gacsalyi I, Mezei T, Pallagi K, Kompagne H, Levay G, Egyed A, et al.: (Phenylpiperazinyl-butyl)oxindoles as selective 5-HT7 receptor antagonists. *J Med Chem* 2008, 51:2522-2532.
69. Yoon J, Yoo EA, Kim JY, Pae AN, Rhim H, Park WK, Kong JY, Park Choo HY: Preparation of piperazine derivatives as 5-HT7 receptor antagonists. *Bioorg Med Chem* 2008, 16:5405-5412.
70. Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A: Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J Comput Aided Mol Des* 2008, 22:593-609.
71. Peterson YK, Wang XS, Casey PJ, Tropsha A: Discovery of geranylgeranyltransferase-I inhibitors with novel scaffolds by the means of quantitative structure-activity relationship modeling, virtual screening, and experimental validation. *J Med Chem* 2009, 52:4210-4220.
72. Tang H, Wang XS, Huang XP, Roth BL, Butler KV, Kozikowski AP, Jung M, Tropsha A: Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation. *J Chem Inf Model.* 2009, 49:461-476.
73. Tropsha A, Golbraikh A: Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm.Des* 2007, 13:3494-3504.
74. Wang XS, Tang H, Golbraikh A, Tropsha A: Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J.Chem.Inf.Model.* 2008, 48:997-1013.
75. Vermeulen ES, van SM, Schmidt AW, Sprouse JS, Wikstrom HV, Grol CJ: Novel 5-HT7 receptor inverse agonists. Synthesis and molecular modeling of arylpiperazine- and 1,2,3,4-tetrahydroisoquinoline-based arylsulfonamides. *J Med Chem* 2004, 47:5451-5466.
76. Jalali-Heravi M, sadollahi-Baboli M: Quantitative structure-activity relationship study of serotonin (5-HT7) receptor inhibitors using modified ant colony algorithm and adaptive neuro-fuzzy interference system (ANFIS). *Eur J Med Chem* 2009, 44:1463-1470.
77. Fourches D, Muratov E, Tropsha A: Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* 2010, 50:1189-1204.



78. Roth BL, Craigo SC, Choudhary MS, Uluer A, Monsma FJ, Shen Y, Meltzer HY, Sibley DR: Binding of typical and atypical antipsychotic agents to 5-hydroxytryptamine-6 and 5-hydroxytryptamine-7 receptors. *J.Pharmacol.Exp.Ther.* 1994, 268:1403-1410.
79. Thomson Scientific: World Drug Index (WDI). <http://www.daylight.com/products/wdi.html> 2007, [no volume].
80. Hall L, Kellog G, Haney D: *Molconn-Z version 4.00 user guide [Internet]*. 2002.
81. Kier LB, Hall LH: An electrotopological-state index for atoms in molecules. *Pharm.Res.* 1990, 7:801-807.
82. Kier LB, Hall LH: *Molecular connectivity in chemistry and drug research*. Academic Press: New York; 1976.
83. Kier LB, Hall LH: *Molecular connectivity in structure-activity analysis*. Wiley: New York; 1986.
84. Kier LB, Hall LH: *Molecular Structure Description: The Electrotopological State*. Academic Press; 1999.
85. Kier LB, Hall LH: An Index of Electrotopological State of Atoms in Molecules. *J.Med.Chem.* 1991, 7:229.
86. Petitjean M: Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J.Chem.Inf.Comput.Sci.* 1992, 32:331-337.
87. Randi M: On Characterization on Molecular Branching. *J.Am.Chem.Soc.* 1975, 97:6609-6615.
88. Wiener H: Structural determination of paraffin boiling points. *J.Am.Chem.Soc.* 1947, 69:17-20.
89. Oloff S, Mailman RB, Tropsha A: Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J.Med.Chem.* 2005, 48:7322-7332.
90. Christopher J.C.Burges: A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and KNowledge Discovery* 1998, 2:955-974.
91. Farid Alizadeh, Ronald Goldfarb: Second-order cone programming. *Math.Program.* 2003, 95:3-51.
92. Golbraikh A, Tropsha A: Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J.Comput.Aided Mol.Des* 2002, 16:357-369.

93. Sharaf M, Illman D, Kowalski B: *Chemometrics*. John Wiley & Sons: New York; 1986.
94. Abbas AI, Hedlund PB, Huang XP, Tran TB, Meltzer HY, Roth BL: Amisulpride is a potent 5-HT<sub>7</sub> antagonist: relevance for antidepressant actions in vivo. *Psychopharmacology (Berl)* 2009, 205:119-128.
95. Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijter MB, Matos RC, Tran TB, et al.: Predicting new molecular targets for known drugs. *Nature* 2009, 462:175-181.
96. Gilbert N: *Statistics*. W.B. Saunders, Co.: Philadelphia, PA; 1976.
97. Baker N: Extracting Drug Activity Terms from Medline Annotations. In *Summit on Translational Bioinformatics*. American Medical Informatics Association; 2008.
98. Zhang S, Wei L, Bastow K, Zheng W, Brossi A, Lee KH, Tropsha A: Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J.Comput.Aided Mol.Des* 2007, 21:97-112.
99. Medina-Franco JL, Golbraikh A, Oloff S, Castillo R, Tropsha A: Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J.Comput.Aided Mol.Des* 2005, 19:229-242.
100. Aubry U, Carignan G, Charette D, Keeri-Szanto M, Lavallee JP: Neuroleptanalgesia with fentanyl-droperidol: an appreciation based on more than 1000 anaesthetics for major surgery. *Can.Anaesth.Soc.J.* 1966, 13:263-271.
101. Apfel CC, Cakmakaya OS, Frings G, Kranke P, Malhotra A, Stader A, Turan A, Biedler A, Kolodzie K: Droperidol has comparable clinical efficacy against both nausea and vomiting. *Br.J.Anaesth.* 2009, 103:359-363.
102. Resnick M, Burton BT: Droperidol vs. haloperidol in the initial management of acutely agitated patients. *J.Clin.Psychiatry* 1984, 45:298-299.
103. Flood P, Coates KM: Droperidol inhibits GABA(A) and neuronal nicotinic receptor activation. *Anesthesiology* 2002, 96:987-993.
104. Hyatt M, Muldoon SM, Rorie DK: Droperidol, a selective antagonist of postsynaptic alpha-adrenoceptors in the canine saphenous vein. *Anesthesiology* 1980, 53:281-286.
105. Toda N, Hatano Y: Antagonism by droperidol of dopamine-induced relaxation in isolated dog arteries. *Eur.J.Pharmacol.* 1979, 57:231-238.
106. Onrust SV, McClellan K: Perospirone. *CNS.Drugs* 2001, 15:329-337.

107. Hirose A, Kato T, Ohno Y, Shimizu H, Tanaka H, Nakamura M, Katsube J: Pharmacological actions of SM-9018, a new neuroleptic drug with both potent 5-hydroxytryptamine<sub>2</sub> and dopamine<sub>2</sub> antagonistic actions. *Jpn.J.Pharmacol.* 1990, 53:321-329.
108. Kato T, Hirose A, Ohno Y, Shimizu H, Tanaka H, Nakamura M: Binding profile of SM-9018, a novel antipsychotic candidate. *Jpn.J.Pharmacol.* 1990, 54:478-481.
109. DiMasi JA, Hansen RW, Grabowski HG: The price of innovation: new estimates of drug development costs. *J Health Econ.* 2003, 22:151-185.
110. DiMasi JA, Hansen RW, Grabowski HG, Lasagna L: Cost of innovation in the pharmaceutical industry. *J Health Econ.* 1991, 10:107-142.
111. Sprous DG, Palmer RK, Swanson JT, Lawless M: QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. *Curr Top.Med Chem* 2010, 10:619-637.
112. Obrezanova O, Segall MD: Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity. *Journal of Chemical Information and Modeling* 2010, doi:doi: 10.1021/ci900406x.
113. Bruce CL, Melville JL, Pickett SD, Hirst JD: Contemporary QSAR Classifiers Compared. *Journal of Chemical Information and Modeling* 2007, 47:219-227.
114. Truchon JF, Bayly CI: Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *Journal of Chemical Information and Modeling* 2007, 47:488-508.
115. Nilakantan R, Nunn DS, Greenblatt L, Walker G, Haraki K, Mobilio D: A family of ring system-based structural fragments for use in structure-activity studies: database mining and recursive partitioning. *J Chem Inf Model.* 2006, 46:1069-1077.
116. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 2006, 34:D668-D672.
117. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucl.Acids Res.* 2008, 36:D901-D906.
118. Cordon-Cardo C, O'Brien JP, Boccia J, Casals D, Bertino JR, Melamed MR: Expression of the multidrug resistance gene product (P-glycoprotein) in human normal and tumor tissues. *J Histochem.Cytochem.* 1990, 38:1277-1287.
119. Huls M, Russel FG, Masereeuw R: The role of ATP binding cassette transporters in tissue defense and organ regeneration. *J Pharmacol Exp Ther* 2009, 328:3-9.

120. Ozawa N, Shimizu T, Morita R, Yokono Y, Ochiai T, Munesada K, Ohashi A, Aida Y, Hama Y, Taki K, et al.: Transporter Database, TP-Search: A Web-Accessible Comprehensive Database for Research in Pharmacokinetics of Drugs. *Pharmaceutical Research* 2004, 21:2133-2134.
121. Cabrera MA, Gonzalez I, Fernandez C, Navarro C, Bermejo M: A topological substructural approach for the prediction of P-glycoprotein substrates. *J Pharm.Sci* 2006, 95:589-606.
122. Weisman JL, Liou AP, Shelat AA, Cohen FE, Guy RK, DeRisi JL: Searching for new antimalarial therapeutics amongst known drugs. *Chem Biol Drug Des* 2006, 67:409-416.
123. Schultz TW, Netzeva TI: Development and evaluation of QSARs for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity. In *\_Modeling En\_V\_ironmental Fate and Toxicity\_*. CRC Press: Boca Raton; 2004:265-284.
124. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Oberg T, Dao P, Cherkasov A, Tetko IV: Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J Chem Inf Model* 2008, 48:766-784.
125. Netzeva TI, Schultz TW: QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data. *Chemosphere* 2005, 61:1632-1643.
126. Aptula AO, Roberts DW, Cronin MTD, Schultz TW: Chemistry-Toxicity Relationships for the Effects of Di- and Trihydroxybenzenes to Tetrahymena pyriformis. *Chemical Research in Toxicology* 2005, 18:844-854.
127. Schultz TW, Sinks GD, Miller LA: Population growth impairment of sulfur-containing compounds to Tetrahymena pyriformis. *Environ Toxicol.* 2001, 16:543-549.
128. Schultz TW, Yarbrough JW, Woldemeskel M: Toxicity to Tetrahymena and abiotic thiol reactivity of aromatic isothiocyanates. *Cell Biol Toxicol.* 2005, 21:181-189.
129. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD: Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action. *QSAR Comb.Sci.* 2007, 26:238-254.
130. Huang XP, Setola V, Yadav PN, Allen JA, Rogan SC, Hanson BJ, Revankar C, Robers M, Doucette C, Roth BL: Parallel Functional Activity Profiling Reveals Valvulopathogens Are Potent 5-Hydroxytryptamine2B Receptor Agonists: Implications for Drug Safety Assessment. *Molecular Pharmacology* 2009, 76:710-722.
131. Zhu H, Martin TM, Ye L, Sedykh A, Young DM, Tropsha A: Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol* 2009, 22:1913-1921.

132. The globally harmonized system of classification and labelling of chemicals. 2010, [no volume].
133. Jolly WL, Perry WB: Estimation of atomic charges by an electronegativity equalization procedure calibrated with core binding energies. *Journal of the American Chemical Society* 1973, 95:5442-5450.
134. Wang R, Fu Y, Lai L: A New Atom-Additive Method for Calculating Partition Coefficients. *Journal of Chemical Information and Computer Sciences* 1997, 37:615-621.
135. Ioffe BV: *Chemistry Refractometric Methods*. Himiya:Leningrad; 1983.
136. Kuz'min VE, Artemenko AG, Muratov EN, Volineckaya IL, Makarov VA, Riabova OB, Wutzler P, Schmidtke M: Quantitative structure-activity relationship studies of [(biphenyloxy)propyl]isoxazole derivatives. Inhibitors of human rhinovirus 2 replication. *J. Med. Chem.* 2007, 50:4205-4213.
137. Polishchuk PG, Muratov EN, Artemenko AG, Kolumbin OG, Muratov NN, Kuz'min VE: Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *Journal of Chemical Information and Modeling* 2009, 49:2481-2488.
138. Kuz'min, Artemenko A, Muratov E: Hierarchical QSAR technology based on the Simplex representation of molecular structure. *Journal of Computer-Aided Molecular Design* 2008, 22:403-421.
139. Muratov EN, Artemenko AG, Varlamova EV, Polischuk PG, Lozitsky VP, Fedchuk AS, Lozitska RL, Gridina TL, Koroleva LS, Sil'nikov VN, et al.: Per aspera ad astra: application of Simplex QSAR approach in antiviral research. *Future Med Chem* 2010, 2:1205-1226.
140. Sedykh AY, Klopman G: A Structural Analogue Approach to the Prediction of the Octanol Water Partition Coefficient. *Journal of Chemical Information and Modeling* 2006, 46:1598-1603.
141. Sedykh A, Klopman G: Data analysis and alternative modelling of MITI-I aerobic biodegradation. *SAR and QSAR in Environmental Research* 2007, 18:693-709.
142. Altman DG, Bland JM: Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ* 1994, 308:1552.
143. Hess A, Iyer H: Fisher's combined p-value for detecting differentially expressed genes using Affymetrix expression arrays. *BMC Genomics* 2007, 8:96.
144. FIELDING AH, BELL JF: A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation* 1997, 24:38-49.

145. Deeks JJ, Altman DG: Diagnostic tests 4: likelihood ratios. *BMJ* 2004, 329:168-169.
146. Schepetkin IA, Kirpotina LN, Khlebnikov AI, Quinn MT: High-Throughput Screening for Small-Molecule Activators of Neutrophils: Identification of Novel N-Formyl Peptide Receptor Agonists. *Molecular Pharmacology* 2007, 71:1061-1074.
147. Jenkins JL, Kao RYT, Shapiro R: Virtual screening to enrich hit lists from high-throughput screening: A case study on small-molecule inhibitors of angiogenin. *Proteins* 2003, 50:81-93.
148. Flower DR: On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences* 1998, 38:379-386.
149. Willett P, Barnard JM, Downs GM: Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* 1998, 38:983-996.
150. Dearden JC, Cronin MT, Kaiser KL: How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ.Res* 2009, 20:241-266.
151. Swamidass SJ, Bittker JA, Bodycombe NE, Ryder SP, Clemons PA: An economic framework to prioritize confirmatory tests after a high-throughput screen. *J Biomol Screen* 2010, 15:680-686.
152. Austin CP, Brady LS, Insel TR, Collins FS: NIH Molecular Libraries Initiative. *Science* 2004, 306:1138-1139.
153. Young D, Martin T, Venkatapathy R, Harten P: Are the Chemical Structures in Your QSAR Correct? *QSAR & Combinatorial Science* 2008, 27:1337-1345.
154. Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, Oberg T, Todeschini R, Fourches D, Varnek A: Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* 2008, 48:1733-1746.
155. Ito H, Halldin C, Farde L: Localization of 5-HT<sub>1A</sub> receptors in the living human brain using [carbonyl-<sup>11</sup>C]WAY-100635: PET with anatomic standardization technique. *J. Nucl. Med.* 1999, 40:102-109.
156. Parks CL, Robinson PS, Sibille E, Shenk T, Toth M: Increased anxiety of mice lacking the serotonin<sub>1A</sub> receptor. *Proc. Natl. Acad. Sci. U.S.A.* 1998, 95:10734-10739.
157. Kennett GA, Dourish CT, Curzon G: Antidepressant-like action of 5-HT<sub>1A</sub> agonists and conventional antidepressants in an animal model of depression. *Eur. J. Pharmacol.* 1987, 134:265-274.

158. Wheeler Vega JA, Mortimer AM, Tyson PJ: Conventional antipsychotic prescription in unipolar depression, I: an audit and recommendations for practice. *J Clin Psychiatry* 2003, 64:568-574.
159. Li Z, Ichikawa J, Dai J, Meltzer HY: Aripiprazole, a novel antipsychotic drug, preferentially increases dopamine release in the prefrontal cortex and hippocampus in rat brain. *Eur. J. Pharmacol.* 2004, 493:75-83.
160. Bantick RA, De Vries MH, Grasby PM: The effect of a 5-HT1A receptor agonist on striatal dopamine release. *Synapse* 2005, 57:67-75.
161. Todeschini R. C: *Handbook of Molecular Descriptors*. Wiley; 2000.
162. Mauri A, Consonni V, Pavan M, Todeschini R: DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. *MATCH* 2006, 56:237-248.
163. Wiener H: Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *J. Am. Chem. Soc.* 1947, 69:2636-2638.
164. Alexandru T. B: Highly discriminating distance-based topological index. *Chemical Physics Letters* 1982, 89:399-404.
165. Balaban A: Five New Topological Indices for the Branching of Tree-Like Graphs. *Theoretica Chimica Acta* 1979, 53:355-375.
166. Gasteiger J, Marsili M: Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* 1980, 36:3219-3228.
167. Fourches D, Barnes JC, Day NC, Bradley P, Reed JZ, Tropsha A: Cheminformatics Analysis of Assertions Mined from Literature that Describe Drug-Induced Liver Injury in Different Species. *Chem Res Toxicol* 2010, 23:171-183.
168. Vidal D, Mestres J: In Silico Receptorome Screening of Antipsychotic Drugs. *Mol. Inf.* 2010, 29:543-551.
169. Vassilatis DK, Hohmann JG, Zeng H, Li F, Ranchalis JE, Mortrud MT, Brown A, Rodriguez SS, Weller JR, Wright AC, et al.: The G protein-coupled receptor repertoires of human and mouse. *Proc. Natl. Acad. Sci. U.S.A.* 2003, 100:4903-4908.
170. Rothman RB, Baumann MH, Savage JE, Rauser L, McBride A, Hufeisen SJ, Roth BL: Evidence for Possible Involvement of 5-HT2B Receptors in the Cardiac Valvulopathy Associated With Fenfluramine and Other Serotonergic Medications. *Circulation* 2000, 102:2836 -2841.
171. Kroeze WK, Hufeisen SJ, Popadak BA, Renock SM, Steinberg S, Ernsberger P, Jayathilake K, Meltzer HY, Roth BL: H1-histamine receptor affinity predicts short-

- term weight gain for typical and atypical antipsychotic drugs. *Neuropsychopharmacology* 2003, 28:519-526.
172. Roth BL, Sheffler DJ, Kroeze WK: Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov* 2004, 3:353-359.
  173. Roth BL, Kroeze WK: Screening the receptorome yields validated molecular targets for drug discovery. *Curr. Pharm. Des.* 2006, 12:1785-1795.
  174. Prinz F, Schlange T, Asadullah K: Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov* 2011, 10:712.
  175. Mullard A: Reliability of “new drug target” claims called into question. *Nat Rev Drug Discov* 2011, 10:643-644.
  176. Williams AJ, Ekins S: A quality alert and call for improved curation of public chemistry databases. *Drug Discov. Today* 2011, 16:747-750.
  177. Wold S, Eriksson L, Clementi S: Statistical Validation of QSAR Results [Internet]. In *Chemometrics Methods in Molecular Design*. 1995:309-338.
  178. Uehling DE, Shearer BG, Donaldson KH, Chao EY, Deaton DN, Adkison KK, Brown KK, Cariello NF, Faison WL, Lancaster ME, et al.: Biarylaniline phenethanolamines as potent and selective beta3 adrenergic receptor agonists. *J. Med. Chem.* 2006, 49:2758-2771.
  179. Blum LC, Reymond J-L: 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* 2009, 131:8732-8733.
  180. Walker T, Grulke CM, Pozefsky D, Tropsha A: Chembench: a cheminformatics workbench. *Bioinformatics* 2010, 26:3000 -3001.
  181. Ekins S, Williams AJ, Krasowski MD, Freundlich JS: In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov. Today* 2011, 16:298-310.
  182. Jones DR, Ekins S, Li L, Hall SD: Computational approaches that predict metabolic intermediate complex formation with CYP3A4 (+b5). *Drug Metab. Dispos.* 2007, 35:1466-1475.
  183. 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, 12:2111-2120.