

# CHEMINFORMATICS MODELING OF DIVERSE AND DISPARATE BIOLOGICAL DATA AND THE USE OF MODELS TO DISCOVER NOVEL BIOACTIVE MOLECULES

Man Luo

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 in Philosophy  
in the UNC Eshelman School of Pharmacy  
(Division of Medicinal Chemistry and Natural Products).

Chapel Hill

2011

Approved by:

Dr. Alexander Tropsha

Dr. Simon X. Wang

Dr. Michael Jarstfer

Dr. Stephen Frye

Dr. Mark Lovern

©2011  
Man Luo  
ALL RIGHTS RESERVED

## ABSTRACT

**MAN LUO:** Cheminformatics Modeling of Diverse and Disparate Biological Data and the Use of Models to Discover Novel Bioactive Molecules  
(Under the direction of Dr. Alexander Tropsha)

Ligand-based drug design is a popular and efficient computational approach to facilitate the drug discovery process. Current approaches mainly focus on optimizing the computational algorithms to improve the efficiency or accuracy of virtual screening; however, the success of ligand-based drug design relies not only on the effectiveness and robustness of the underlying algorithms, but much more importantly, on the quality of the data for model building. Although numerous chemical probe databases have emerged recently<sup>1-4</sup>, few evaluation of data quality and reliability have been performed.

Building upon our lab's experience in Quantitative Structure-Activity Relationship (QSAR) method and methods developed in the field of cheminformatics, this dissertation focuses on: 1) Investigation and comparison of the predictive power of QSAR methods with other ligand-based drug discovery approaches, such as Similarity Ensemble Approach (SEA) and Prediction of Activity Spectra for Substances (PASS); 2) Using QSAR methods to validate the consistency and reliability of biomedical data in disparate data sources. 3) Developing a novel, rigorous and dataset-specific QSAR workflow for the application on multiple therapeutic targets in order to identify diverse hits with high potency in practical virtual screening projects. These works succeed in thoroughly investigating the current approaches for ligand-based drug discovery, exploring the consistency and quality of major

annotated cheminformatics databases, and identifying many pharmaceutically important ligands. The success of our studies harshly challenges some popular multi-target profile prediction methods and contributes to the development of cheminformatics by emphasizing the importance of determining trustworthy data sources.

*To my beloved parents, all my family members, and  
Hongjie Zhu, whose support, encouragement and  
personal sacrifice have made this research possible.*

*To my mentors, and all senior professionals, who inspired,  
guided, helped, and touched  
a naïve mind, thus will change my life forever.*

## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance and the help of many individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

I am deeply indebted to Dr. Alexander Tropsha for his scientific guidance, his faith and generous support at the critical moments of my life, his allowance of my exploring different scientific projects and related fields, and his efforts to keep me on the right track. I am very thankful to Dr. Xiang Wang, not only for his patient guidance of detailed modeling techniques, but also for those invaluable research ideas he generously offered.

I appreciate all the kind help and strong support from Drs. Mark Lovern, Michael Jarstfer, and Stephen Frye. Their time, efforts and suggestions are so precious to me.

I am very grateful for Drs. Alexander Golbraikh, Denis Fourches, Alexander Sedykh, Ashutosh Tripathi, and all my labmates in Molecular Modeling Lab, for their warm help, hearty scientific discussions, and all the happiness we shared over the years.

I cannot acknowledge enough my family and beloved boyfriend Hongjie Zhu, for their continuous support and great encouragement throughout my life, without which I could not have achieved it.

## TABLE OF CONTENTS

LIST OF TABLES .....	xiii
LIST OF FIGURURES .....	xiv
CHAPTER .....	1
1. INTRODUCTION .....	1
1.1 Background.....	1
1.2 Introduction to Ligand-based Drug Discovery .....	2
1.3 Current Limitations for Existing Methodologies .....	6
1.4 Overview of Chapter 2 .....	7
1.5 Overview of Chapter 3 .....	8
1.6 Overview of Chapter 4 .....	8
1.7 Overview of Chapter 5 .....	9
1.8 Overview of Chapter 6 .....	10
2. METHODS .....	11
2.1 Introduction .....	11
2.2 Background Information of QSAR .....	12
2.3 Descriptors Used .....	12

2.3.1 Dragon descriptors .....	12
2.3.2 MACCS key fragment-based descriptors .....	13
2.4 QSAR Methodology .....	15
2.4.1 K-Nearest Neighbors (kNN) .....	19
2.4.3 Random forest (RF) .....	23
2.4.4 Support Vector Machines (SVM) .....	26
2.4.5 Robustness of QSAR models.....	29
2.5 Background Information of Generic Multi-target Techniques.....	30
2.5.1 Similarity Ensemble Search (SEA).....	30
2.5.2 Prediction of Activity Spectra for Substances' (PASS).....	33
2.6 Comparison of the generic multi-target technique versus single target model building.....	35
3. VALIDATION OF THE CONGRUENCE OF DATA IN DISPARATE SOURCES AND IDENTIFICATION OF LOW QUALITY DATA.....	44
3.1 Introduction .....	44
3.2 The Major Annotated Chemogenomics Databases .....	45
3.2.1 PubChem database and the NIH Molecular Libraries Roadmap Initiative .....	45
3.2.2 NIMH Psychoactive Drug Screening Program (PDSP).....	46



3.2.3 The World of Molecular Bioactivity (WOMBAT).....	47
3.2.4 Chemical European Molecular Biology Library (ChEMBL) .....	47
3.3 Different Data Sources of 5-Hydroxy Tryptamine receptor subtype 1A Ligands .....	47
3.3.1 5-HT1A Agonists and Antagonists from PubChem .....	47
3.3.2 5-HT1A Binders and Non-binders from PDSP .....	50
3.3.3 5-HT1A Binders from WOMBAT .....	51
3.3 Methods .....	51
3.3.1 Datasets Curation and Descriptors Generation .....	51
3.3.2 Training, Test, and External Validation Set Selection.....	52
3.3.3 QSAR Models Generated from Disparate Data Sources .....	54
3.3.2 QSAR Model-Based Cross Validation between Disparate Data Sources.....	54
3.4 Results and Discussions .....	54
3.4.1 QSAR Classification Models .....	54
3.4.2 Validations of QSAR Classification Models .....	56
3.4.3 Inter-dataset Cross Validation .....	59
3.4.4 Experimental Validation .....	63
3.5 Conclusions .....	68

4. DEVELOPMENT OF THE “DIVIDE-AND-CONQUER” QSAR MODELING SCHEME FOR RECA INHIBITORS AND VIRTUAL SCREENING FOR IDENTIFYING NOVEL CHEMICAL SCAFFOLDS .....	69
4.1 Introduction .....	69
4.1.1 Introduction for RecA Protein Inhibitors .....	70
4.1.2 RecA Dataset .....	72
4.1.3 Libraries for Virtual Screening .....	73
4.2 Methods .....	73
4.2.1 Generation of Descriptors and Dataset Split .....	73
4.2.2 Model-based Data Curation and Activity Cliffs Identification .....	74
4.2.3 “Divide-and-Conquer” QSAR Modeling Scheme for Curated RecA Dataset .....	76
4.2.4 QSAR-based Virtual Screening .....	79
4.2.5 Experimental Validation of Virtual Screening Hits .....	79
4.3 Results and Discussions .....	80
4.3.1 QSAR Classification Models On First Version Dataset .....	80
4.3.2 Identifications of Compound Pairs with Activity Cliffs and Data Curation .....	81
4.3.3 QSAR Classification Models On Dataset After Curation .....	86
4.3.4 Virtual Screening To Identify Putative RecA Inhibitors .....	91

4.3.5 Experimental Validation .....	92
4.4 Conclusions .....	97
5. DEVELOPMENT OF COMBINATORIAL QSAR MODELS FOR 5-HYDROXYTRYPTAMINE 1A RECEPTOR AND VIRTUAL SCREENING OF LIBRARIES WITH DIFFERENT CHARACTERISTICS .....	99
5.1 Introduction .....	99
5.1.1 Introduction for the 5- Hydroxytryptamine receptor 1A .....	100
5.1.2 Introduction of the Dataset for QSAR Model Building.....	101
5.1.3 Introduction of the Libraries for Virtual Screening .....	101
5.2 Methods .....	103
5.2.1 Dataset Curation .....	103
5.2.2 QSAR Modeling and External Validation .....	104
5.2.3 Virtual Screening of Various Types of Libraries.....	105
5.2.4 Experimental Testing .....	106
5.3 Results .....	107
5.3.1 QSAR Classification Models.....	107
5.3.2 QSAR Model Validations .....	108
5.4 Discussions .....	129

5.5 Conclusions .....	131
6. SUMMARY AND FUTURE DIRECTIONS .....	133
6.1 Summary and Future Directions of Chapter 2.....	134
6.2 Summary and Future Directions of Chapter 3.....	135
6.3 Summary and Future Directions of Chapter 4.....	136
6.4 Summary and Future Directions of Chapter 5.....	137
REFERENCES .....	144

## LIST OF TABLES

### Table

3.1. External validation statistics for disparate 5-HT1A datasets by different QSAR methods.....	58
3.2. The cross-validation of 69 WOMBAT 5-HT1A inhibitors by consensus prediction of acceptable QSAR models that were generated for PDSP dataset.....	62
3.3. The cross validation of 46 PubChem confirmatory 5-HT1A agonists/antagonists by consensus prediction of acceptable QSAR models that were generated for PDSP dataset .....	62
3.4. The experimental validation of 5-HT1A binding affinity test for PubChem confirmatory agonists/antagonists. ....	65
4.1. Results for the five-fold external sets cross validation as well as the secondary external set (from WOMBAT) validation using three different machine learning methods.....	90
4.2. The experimental test for the five computational hits of 5-HT1A inhibitors by mining the TimTec GPCR targeted screening library.....	94
5.1. Results for the five-fold external sets cross validation as well as the secondary external set validation using three different machine learning methods. ....	111
5.2. The experimental test for the ten virtual hits of 5-HT1A binders identified by virtual screening. ....	124

## LIST OF FIGURES

### Figure

1.1. Popular data mining approaches in cheminformatics .....	5
2.1. Depiction of MACCS key fragment-based descriptors .....	14
2.2. The workflow of QSAR model building, validation and virtual screening .....	18
2.3. Prediction based on <i>k</i> -nearest neighbors.....	21
2.4. Prediction based on random forest algorithms.....	25
2.5. Support Vection Machine (SVM) with maximum seperation. ....	28
2.6. The algorithm for Similarity Ensemble Approach (SEA) and its web-based platform. ....	32
2.7. The software platform for the Prediction of Activity Spectra for Substances (PASS). ....	34
2.8. Data disposition of 7 different GPCR targets in PDSP, WOMBAT and ChEMBL. ....	38
2.9. Comparison of SEA and <i>k</i> NN-QSAR on internal prediction. ....	39
2.10. Comparison of SEA and <i>k</i> NN-QSAR on external prediction.....	41
2.11. Comparison of <i>k</i> NN-QSAR external prediction for different data sources. ....	42
2.12. Comparison of PASS and <i>k</i> NN-QSAR on external prediction.....	43
3.1. The similarity search results of 5-HT1A binders for each dataset using binders from the other dataset as probes. ....	60
4.1. The workflow of the “Divide-and-Conquer” QSAR modeling approaches as applied to the RecA dataset.....	78
4.2. Heatmap of pair-wise Tc analysis for the first version RecA dataset.....	83

4.3. Activity cliffs of RecA inhibitors and non-inhibitors. ....	84
4.4. Hierarchical Clustering of 145 RecA inhibitors. ....	87
4.5. Five-fold external set prediction results by <i>k</i> NN-QSAR for RecA dataset after cluster .....	89
5.1. The statistics of five-fold external validations of 5-HT1A compounds from PDSP for three QSAR methods and Y-Randomization test.....	110
5.2. The principal component analysis of three virtual screening libraries and modeling set compounds .....	115
5.3. Hit rate of 5-HT1A binders on diverse screening libraries using different $Z_{\text{cutoff}}$ values .....	116
5.4. The structural similarity analysis of virtual hits screened from different libraries .....	119
5.5. The full dose response curves for hit compounds and the positive control. ....	122

## ABBREVIATIONS

5-HT1A	5-Hydroxy Tryptamine receptor subtype 1A
AD	Applicability Domain
CARs	Class Association Rules
CBA	Classification Based on Association
CCR	Correct Classification Rate
CCR <sub>train</sub>	Correct Classification Rate for training set
CCR <sub>test</sub>	Correct Classification Rate for test set
CCR <sub>evs</sub>	Correct Classification Rate for external validation set
CCR <sub>ex</sub>	Correct Classification Rate for external set
CV	Cross Validation
FN	False Negative
FP	False Positive
GPCR	G-Protein-Coupled Receptor
<i>k</i> NN	<i>k</i> Nearest Neighbor
LOO-CV	Leave-One-Out Cross Validation
MOE	Molecular Operating Environment



PDSP	NIMH Psychoactive Drug Screening Program
QSAR	Quantitative Structure-Activity Relationship
RF	Random Forest
SA	Simulated Annealing
SE	Sensitivity
SP	Specificity
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
VS	Virtual Screening
WDI	World Drug Index
WOMBAT	World of Molecular Bioactivity

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Background**

Drug design is the inventive process of finding new medications based on the knowledge of the biological target or a series of pharmacological agents that modulate it. Other than the classical medicinal chemistry of Structure-Activity Relationship (SAR) analysis, the Computer-Assisted Drug Design (CADD) techniques have become more and more popular recently. They have been proven to greatly improve the efficiency of the drug discovery process, and save a huge amount of money at the same time. For example, the introduction of rational drug design with the aid of computational works led to the discovery of Gleevec to inhibit bcr-abl kinase for the treatment of Chronic Myelogenous Leukemia (CML)<sup>5</sup>. The combination of computational chemistry concepts, robust software, and high-end computer hardware are used to assist the medicinal chemists identifying or designing ligands that are more likely to interact with the biological target of interest. CADD methods can be categorized as ligand-based and structure-based drug design based on the availability of crystal structures for the target of interest. If the three-dimensional (3D) structure of the target protein is available, structure-based drug design approaches could be used, which basically use active and/or binding site identification methods, and docking methods along with different scoring functions to rank screening compounds' structures as well as their

poses. Those compounds with the highest ranks would be proposed to be computational hits, and then rendered for experimental tests, if applicable. If the structure of the target protein is not known, which is a more common case; ligand-based drug design methods are used. The methodologies are based solely on the structure and activity data for series of ligands to a biological target of interest; the protein structure is not used. In this case, all chemical structures are represented by numeric characteristics called molecular descriptors. The generated descriptor matrix consisting of  $m$  rows (each row represents a compound) and  $n$  columns (each column represents a descriptor) will then be analyzed by diverse data analysis approaches, with the aim of better understanding the underlying nature of current data, and making predictions for new molecules. It can also be said that each molecule is represented by a point in the multidimensional descriptor space. In this chapter, a brief introduction of ligand-based drug discovery and various popular data analysis methods in cheminformatics research will be introduced. Then, the limitations for existing concepts and methodologies will be covered, leading to the topics that this dissertation will attempt to address.

## **1.2 Introduction to Ligand-based Drug Discovery**

Ligand-based drug design relies on experimental binding data to a target of interest for a set of small molecules. By analyzing these molecules using various 3D data analysis approaches, a pharmacophore model can be derived, which defines the minimum necessary structural characteristics a molecule must possess in order to bind to the receptor. One of the keys to the success of finding novel ligands is the underlying data analysis approaches in cheminformatics.

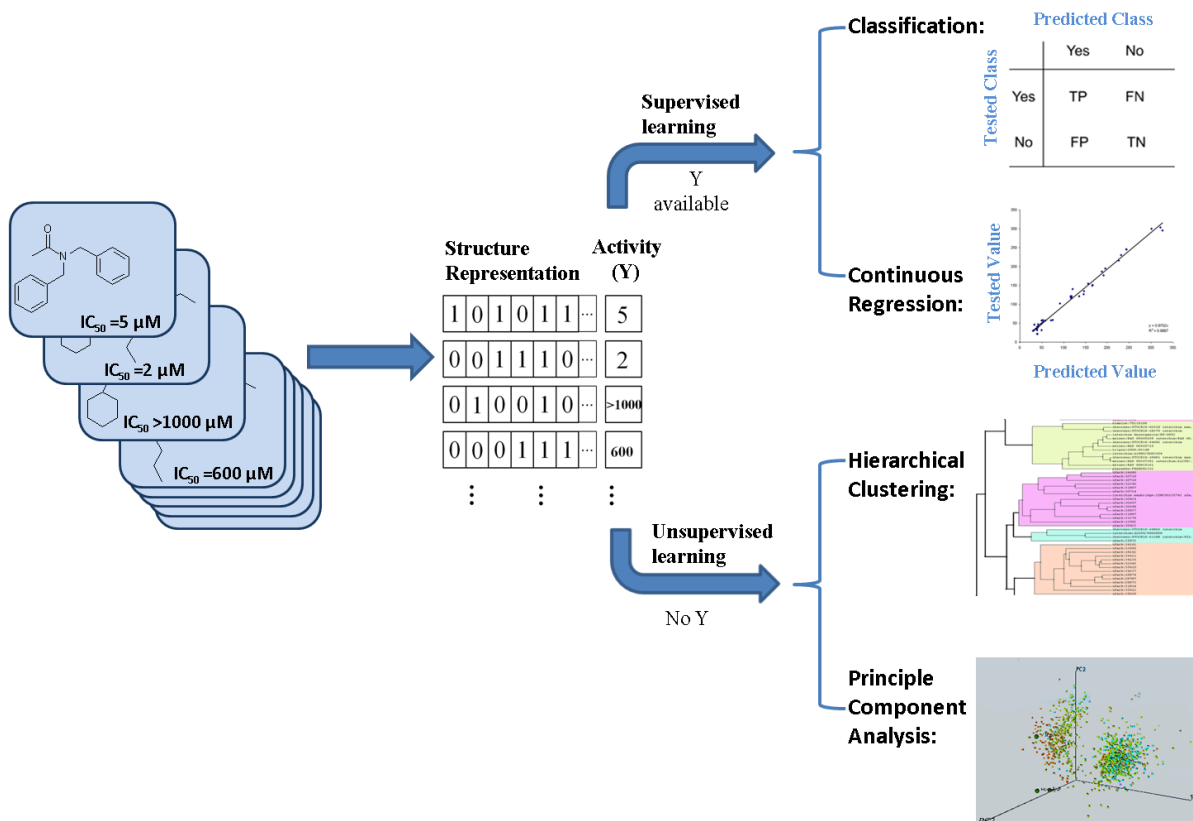
Intelligent data analysis with subsequent database mining is a common and efficient approach for the discovery of lead compounds. Usually, the input data for data analysis is a

descriptor matrix (see above). Various descriptors can be calculated using available software, such as Dragon<sup>6</sup>, Molconn-Z<sup>7</sup>, MOE<sup>8</sup>, MACCS keys<sup>9</sup>, eTc. Different physicochemical properties (calculated or measured), invariants of molecular graphs, indicators of presence/absence of specific chemical groups in a molecule and counts of different fragments, eTc. can serve as descriptors. Data analysis approaches can be divided into supervised learning and unsupervised learning. In supervised learning, biological activities or properties of molecules (Y) are used, and the main goal is building models capable of accurate prediction of activities or properties of compounds not included in the modeling dataset. In unsupervised learning, activity data (Y) are not used. The main goals of unsupervised learning are establishing hidden data structure and finding general properties of the dataset, like how many well-distinguished clusters of compounds exist, obtaining principal components accounting for most (for example, 95% of the variance), factor analysis, eTc. (**Figure 1.1**). The Quantitative Structure-Activity/Property Relationship (QSAR/QSPR) analysis includes various supervised (or semi-supervised) learning approaches as used in cheminformatics<sup>10</sup>. The main goal of QSAR/QSPR analysis is building models capable of accurate prediction of biological activities or properties of compounds. Descriptor matrix (see above) and activities or properties of a series of compounds are used as input variables. The QSAR puzzle can be mathematically described as deriving the equation:

$$\text{Predicted Activity} = f(\text{descriptors}) \quad (1)$$

in which the error of prediction is minimized in some way. Different functions in (1) correspond to different multivariate statistical modeling techniques used to generate these predictive models. Each method tries to tune its function parameters to minimize the error of prediction for the training set. Then the models are validated using test set compounds, not

used in building models<sup>11, 12</sup>. If the biological endpoints belong to only a small number (2-4) of classes, for example, protein binders or non-binders, the models' predictive accuracy is the correct classification rate (CCR), which is the (weighted) average of correct classification rates of each class or category (in case of two classes they are called sensitivity and specificity). CCR is optimized for the training set during model building. For an acceptable model, the CCR should not be lower than some predefined value (for example, for a binary dataset, it should not be lower than 0.7, eTc.) Usually, sensitivity and specificity values are also reported to evaluate the performance of models for each class. If the endpoints are continuous, cross-validation  $q^2$  is usually maximized during model building, and the following statistics for the test set are also used to estimate the predictive power of the model:  $R^2$ ,  $R_0^2$ , MSE, eTc<sup>13</sup>. QSAR methods can be divided into linear (e.g. Partial Linear Squares and Multiple Linear Regression) and non-linear methods (e.g.  $k$ -Nearest Neighbor, Random Forest). On the other hand, unsupervised learning approaches only use the compounds' structural information, and no model training procedure is involved. **Figure 1.1** shows two popular unsupervised learning techniques, hierarchical clustering and principal component analysis, to analyze and better understand the characteristics of current data. Predictions can also be made by unsupervised learning, which predict the biological profile (binder or non-binder) of a new chemical entity to a given target based on the structural similarity as compared to known ligands.



**Figure 1.1. Popular data analysis approaches in cheminformatics.** Various types of descriptors are available for the representation of compounds' chemical structures. The numeric matrix after the descriptor calculation will then be analyzed by different machine learning methods, which are categorized by either supervised learning or unsupervised learning. Supervised learning methods try to find the relationship between the descriptors of chemical compounds with their biological activities (Ys). If the endpoints of Y are categorical, classification modeling methods will be applied, with the prediction results represented by the confusion matrix. If Ys are continuous, regression models will be used. Unsupervised learning methods try to find the relationships between those chemical compounds themselves, without considering the Ys.

### 1.3 Current Limitations for Existing Methodologies

The Cheminformatics research has been developing rapidly in recent decades, and numerous data analysis methods have emerged in the forms of both academic freeware and commercial software packages. It is ideal that models are available for all targets of interest, pharmaceutically relevant receptors and biologically meaningful proteins, so that systems-level investigations can become realistic with the efficient multi-profile predictions. However, calculation speed is not all that we should care about; prediction accuracy should always be the most important feature we care about. Some web-based cheminformatics prediction servers, such as Similarity Ensemble Approach<sup>14</sup> (SEA), become very popular recently, which only apply the simple similarity search technique for the purpose of efficient multi-profile predictions. The fast feedback and friendly simple interface attract many users, especially scientists not in the cheminformatics field. In this dissertation, diverse data analysis techniques will be investigated and comprehensively compared.

Successful ligand-based drug discovery projects involve much more than simply applying different data analysis approaches in readily available web-based servers or software packages. Observations suggest that most efforts were made to optimize the computational algorithms in the purpose of improving the efficiency and/or accuracy of virtual screening<sup>15</sup>; however, the success of the methods relied not only on the effectiveness and robustness of the underlying algorithms, but much more importantly, on the quality of the data for model building<sup>16</sup>. Although numerous chemical probe databases have emerged recently, such as PubChem<sup>1</sup>, ChEMBL<sup>2</sup>, WOMBAT<sup>3</sup>, , seldom evaluation of data quality and reliability was performed.

In addition, since many successful stories<sup>11, 12, 17, 18</sup> proved the robustness and predicting accuracy of Quantitative Structure-Activity Relationships (QSAR) approaches, many researchers simply apply the conventional QSAR (Multiple Linear Regression) and the following virtual screening workflow, without taking consideration of the dataset they currently have. In this dissertation, we propose that instead of carrying the same QSAR workflow universally, dataset-specific QSAR approach should be used, which adjusts the procedures based on the special features of existing data or the different objectives one wants to achieve. For example, we will try to show that combining supervised learning of QSAR with unsupervised learning techniques (hierarchical clustering) could successfully address the problem of building models on a dataset with highly diverse compound structures.

## **1.4 Overview of Chapter 2**

In this chapter, various state-of-the-art data analysis tools in cheminformatics for predicting compounds' biological properties will be briefly introduced. For supervised statistical learning, QSAR methods will be discussed, encompassing different algorithms such as *k*-Nearest Neighbors (*k*NN), Random Forest (RF) and Support Vector Machine (SVM). The concept of applicability domains and external validation tests will also be covered. For other popular cheminformatics techniques, the underlying algorithms for two multi-profile prediction approaches, the Similarity Ensemble Search (SEA) and Prediction of Activity Spectra for Substances' (PASS) will be introduced. These methods will be comprehensively compared by 7 cases of biological receptors. We will demonstrate that the SEA method shows the worst records for both the internal retrieval rate and external



prediction accuracy for all tested datasets; PASS have a much higher prediction accuracy; and QSAR models are the best among all.

### **1.5 Overview of Chapter 3**

In this chapter, we first present and investigate several popular cheminformatics databases, including PubChem, PDSP and WOMBAT, and then apply QSAR approach to validate the consistency of data deposition in these databases to identify low signal-to-noise ratio data source. QSAR models for the same biological targets are generated separately for data obtained from different sources, and used for inter-database cross-validations. Computationally suggested false positive compounds are further tested experimentally to support our predictions. Results show that in the investigated cases both PDSP and WOMBAT datasets are trustworthy data sources, but the PubChem dataset is not. The nine PubChem 5-HT1A binders are identified to be false positives by the inter-database CV, and are further validated in the experimental tests, confirming our model predictions to be 100% accurate. These studies will contribute to the development of cheminformatics by suggesting an ever-increasing role of determining trustworthy data sources before model building.

### **1.6 Overview of Chapter 4**

Antibiotic resistance is an escalating problem requiring the discovery of novel antibiotic classes acting on nonclassical cellular targets. Targeting the protein involved in DNA repair, RecA, offers possible attractive solution, because it directly combat the development and transmission of antibiotic resistance and thus makes antibiotics more effective. In this chapter, we firstly developed QSAR models for the first version dataset containing 53 RecA inhibitors and 3,435 non-inhibitors, and identified those RecA non-

inhibitors with false labeling. The dataset is then curated and more tested compounds are included, making the second version dataset contains 145 RecA inhibitors and 26,288 non-inhibitors. Due to the high structural diversity this new dataset presents, we developed combinatorial QSAR models using the novel “Divide-and-Conquer” approach, which involves hierarchical clustering prior to QSAR modeling. The variable selection  $k$ NN, RF and SVM will be employed for model building within each group, and the two groups that render the best model statistics (group 1 and 2) are used in virtual screening. Computational hits are tested experimentally, revealing novel and potent RecA inhibitors for further drug discovery.

## **1.7 Overview of Chapter 5**

The 5-Hydroxy Tryptamine receptor subtype 1A (5-HT1A) has been an attractive target to treat mood disorders such as anxiety and depression. In this chapter, we develop classification combinatorial QSAR models for 5-HT1A receptor using data retrieved from the PDSP Ki database, the trustworthy data source that has already been confirmed in Chapter 3. We employ a rigorous model development workflow, including extensive internal and external validation, and apply them for consensus prediction for the purpose of mining chemical libraries with different characteristics: drug-like libraries from the World Drug Index and Prestwick, GPCR-targeted libraries from TimTec and ASINEX, and diversity libraries from TimTec and ASINEX. Results shows that the computational hits from a drug-like library share the most similar structures with 5-HT1A binders already known, while hits from GPCR-targeted library are much more structurally diverse, and hits from a diversity library are the most diverse ones. Five hits from each library are further tested in radioligand binding assays, and in total of nine novel structures of 5-HT1A binders are discovered.

## **1.8 Overview of Chapter 6**

In this chapter, we present a summary of all above studies. It not only identifies pharmaceutically important hits for drug discovery, but also encompasses a thorough investigation of the current approaches for multi-target profile predictions, but will also shed light on the consistency and quality of major annotated chemogenomics databases. These studies will contribute to the development of cheminformatics and influence the process of drug discovery by suggesting an ever increasing role of QSAR modeling and the importance of determining trustworthy data sources.

## **CHAPTER 2**

### **METHODS**

#### **2.1 Introduction**

Various cheminformatics tools, such as QSAR modeling, are fundamentally based on the similarity principle implying that compounds with similar chemical structures have similar biological properties. Consequently one can predict the biological target property of a molecule from that of chemically similar compounds for which the property is already known. However, QSAR modeling builds quantitative predictive models based on a similarity matrix using various machine learning techniques, while SEA as well as other generic multi-target techniques simply applies similarity search comparisons; herein, the predictive performance of QSAR methods will be compared with that of the approach predicting generic multi-target profiles, namely ‘Similarity Ensemble Approach’ (SEA)<sup>14</sup> and ‘Prediction of Activity Spectra for Substances’ (PASS)<sup>19</sup>. The outcome of this chapter (a compendium of generated models) will serve as a reliable tool for the virtual biological profiling of small molecules in Chapter 3, 4 and 5. In this chapter, the background information about QSAR as well as other generic multi-target cheminformatic techniques will be briefly introduced first, and then a comparison of methods will be performed on various datasets, followed by assessment results and discussion.

## 2.2 Background Information of QSAR

QSAR, which stands for **Quantitative Structure-Activity Relationship**, is a statistic learning methodology of searching, optimizing and validating the best possible mathematic equations that quantitatively correlate a set of chemical structures with their experimentally defined biological or chemical activities. The chief hypothesis of the QSAR approach is that if an implicit structure-activity relationship exists for a given data set, it can be formally manifested via a variety of QSAR models obtained with different descriptors and optimization protocols.

QSAR's most general mathematical form is:

$$\text{Activity} = f(\text{physiochemical properties and/or structural properties})$$

Once established, QSAR models can be then used to predict the biological responses of other similar chemical structures.

## 2.3 Descriptors Used

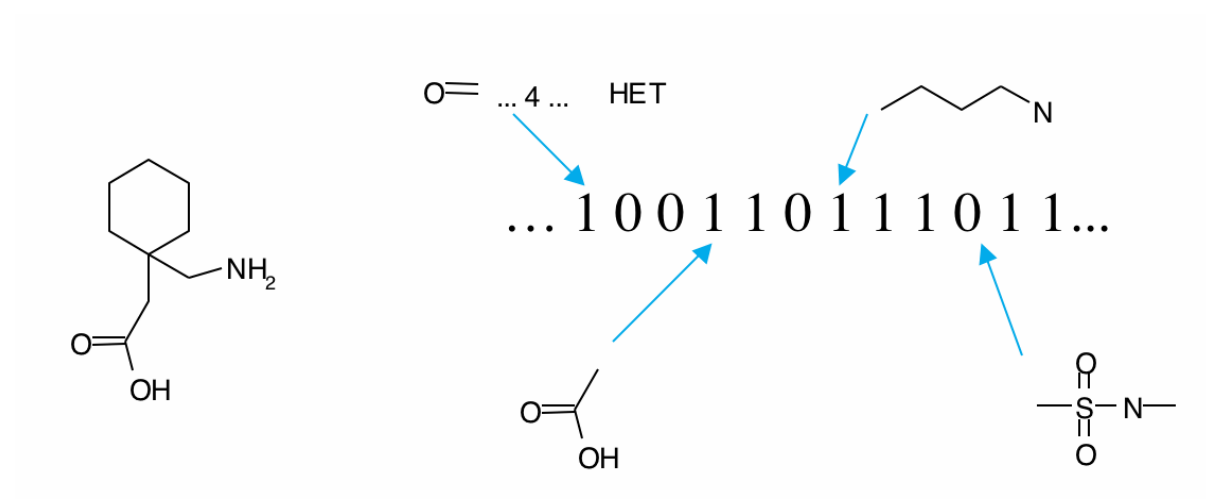
### 2.3.1 *Dragon descriptors*

A set of 843 theoretical molecular descriptors was computed using DRAGON software<sup>6</sup>. The descriptors were generated from the SMILES strings available for each compound. The descriptors include the following types: 0D constitutional (atom and group counts); 1D functional groups; 1D atom centered fragments; 2D topological descriptors; 2D walk and path counts; 2D autocorrelations; 2D connectivity indices; 2D information indices; 2D topological charge indices; 2D Eigenvalue-based indices; 2D edge adjacency indices; 2D Burden eigenvalues and molecular properties. In studies in this dissertation, no 3D

descriptors were used, and all descriptors were calculated with hydrogens, and were range-scaled. Descriptors which had the same value for all compounds or had less than 5% variance were deleted. If two descriptors were at least 98% correlated one of them was deleted. The definition of these descriptors and related literature references are reported elsewhere<sup>6</sup>.

### **2.3.2 MACCS key fragment-based descriptors**

For the feature list version of the MACCS Structural Keys used in our studies, each feature indicates the presence of one of the 166 public MDL MACCS structural keys (fragments) computed from the molecular graph. The fingerprint is represented as a sparse list of keys present in the molecule, as shown in Figure 2.1.



**Figure 2.1. Depiction of MACCS key fragment-based descriptors**

## 2.4 QSAR Methodology

Computer-aided drug design laboratories<sup>20-24</sup>, including ours, has concentrated on the development and application of fast, nonlinear, automated variable selection algorithms for QSAR modeling. These methods include but not limited to Genetic Algorithms-Partial Least Squares<sup>11</sup> (GA-PLS), *k*-Nearest Neighbor<sup>25</sup> (*k*NN), Random Forest (RF)<sup>26</sup>, and Support Vector Machine (SVM)<sup>27</sup>. It has been demonstrated that such methods could efficiently improve QSAR models, compared with those without variable selection<sup>25</sup>. An important aspect of these methods is, as we have shown earlier<sup>11, 25, 28</sup>, that they can be used for combinatorial library design and database mining. Most of the above techniques can deal with both binary and continuous endpoints, but for the continuous ones, a large variance from experiment tests contributes a heavy negative effect on the predictive model building process. Therefore, in this dissertation, binary response variables are exclusively considered, and all of the descriptions below are based on the results derived from methods used only in binary classification QSAR.

Another recent trend in QSAR modeling, which is especially emphasized in our studies, is model validation. It is known that the increase in the number of independent variables leads to a higher probability of chance correlation between predicted and observed activities<sup>29</sup>; Therefore, model validation is one of the most important aspects of the analysis. To validate a QSAR model, the majority of authors apply the leave-one-out (LOO) or leave-some-out (LSO) cross-validation procedure. The outcome from this procedure is a cross-validated Correct Correlation Rate (CCR). For a balanced dataset (dataset that has comparable numbers of binders and non-binders), the CCR is defined as<sup>30</sup>:



$$CCR = \frac{1}{2} \left( \frac{N_1^{corr}}{N_1^{total}} + \frac{N_2^{corr}}{N_2^{total}} \right) \quad (2)$$

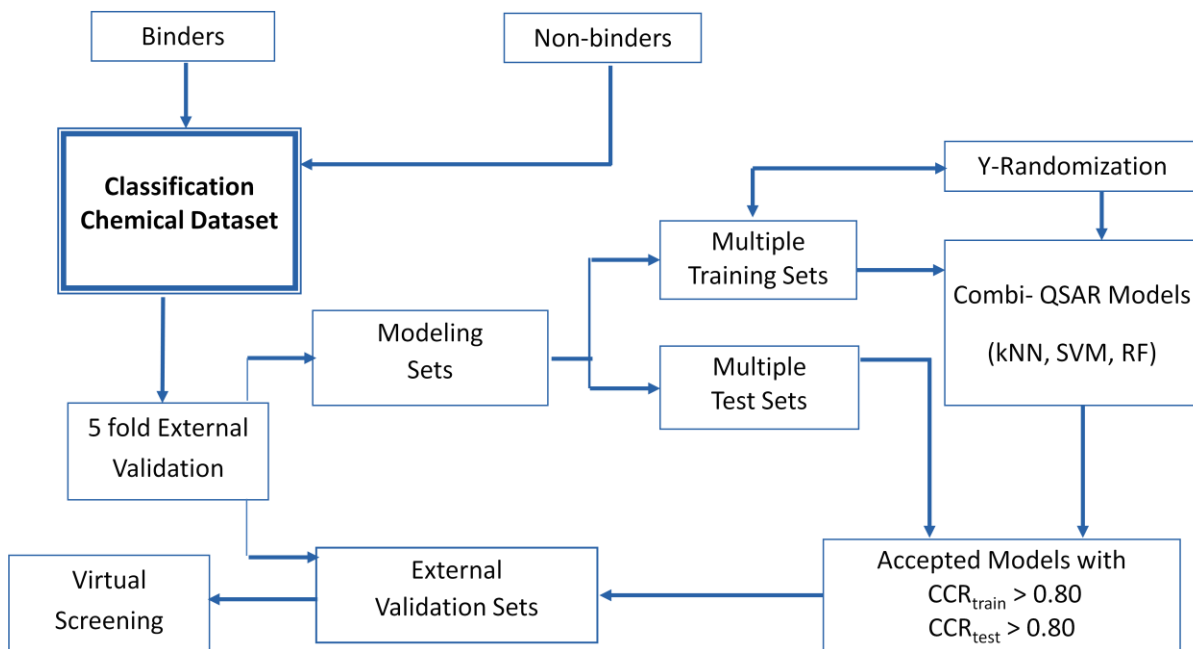
where  $N_1^{total}$  and  $N_2^{total}$  are the number of binders and non-binders in the dataset,  $N_1^{corr}$  and  $N_2^{corr}$  are the number of known binders predicted as binders (true positives) and the number of non-binders predicted as non-binders (true negatives). For imbalanced (biased) dataset, another alternative CCR to Equation 2 could include weights ( $W_1$  and  $W_2$ ) for each class, with smaller weights assigned for larger classes.

$$CCR = W_1 \cdot \frac{N_1^{corr}}{N_1^{total}} + W_2 \cdot \frac{N_2^{corr}}{N_2^{total}} \quad (3)$$

The statistical significance of the training and test set models is characterized by the LOO-CV  $CCR_{train}$  and predictive  $CCR_{test}$ , respectively. In summary, the variable selection  $kNN$  classification method generates stochastic models that usually finalized with values equal to (or close to) the global minimum, with the highest (or nearly highest) value of CCR characterized by the optimal  $k$  value, the number of nearest neighbors, and a subset of selected descriptors.

The summations in Equation 3 are performed over all compounds, which are used to build a model (training set). Many authors consider high  $CCR_{test}$  (for instance,  $CCR_{test} > 0.7$ ) as an indicator or even as the ultimate proof of the high predictive power of a QSAR model<sup>31</sup>. They do not test the models for their ability to predict the activity of compounds of an external test set (*i.e.*, compounds), which have not been used in the QSAR model development<sup>32-36</sup>. Some authors validate their models using only one or two compounds that were not used in model development<sup>37, 38</sup>.

Our laboratory has recently demonstrated<sup>17, 39</sup> that various commonly accepted statistical characteristics of QSAR models derived from a training set are insufficient to establish and estimate the predictive power of the models. As was suggested in several recent publications<sup>40-43</sup>, including our own work<sup>17, 39</sup>, the only way to ensure the high predictive power of a QSAR model is to demonstrate a significant correlation between predicted and observed activities of compounds for an external validation (test) set, which is not employed in model development. We argue that special approaches should be used to select a training set to ensure the highest significance and predictive power of QSAR models<sup>44, 45</sup>. Our approach to QSAR modeling does involve extensive validation as discussed below in Chapter 3, 4, and 5.



**Figure 2.2. The workflow of QSAR model building, validation and virtual screening.**

### 2.4.1 *K-Nearest Neighbors (kNN)*

#### 2.4.1.1 *kNN Classification Methodology*

The *kNN* classification QSAR method<sup>46, 47</sup> is based on the idea that the class that a compound belongs to can be defined by the class membership of its nearest neighbors (i.e., most similar compounds) taking into account weighted similarities between a compound and its nearest neighbors. Since our implementation of *kNN* approach includes simulated annealing (SA) based variable selection<sup>25</sup> as a stochastic optimization algorithm, the similarity is evaluated using only a subset of all descriptors, and is characterized by weighted Euclidean distance between compounds in multidimensional descriptor space. Thus, the class membership of compound *i* can be predicted from the following equation:

$$\hat{y}_i = \frac{\sum_{j=1}^k w_{ij} y_j}{\sum_{j=1}^k w_{ij}} \quad (4)$$

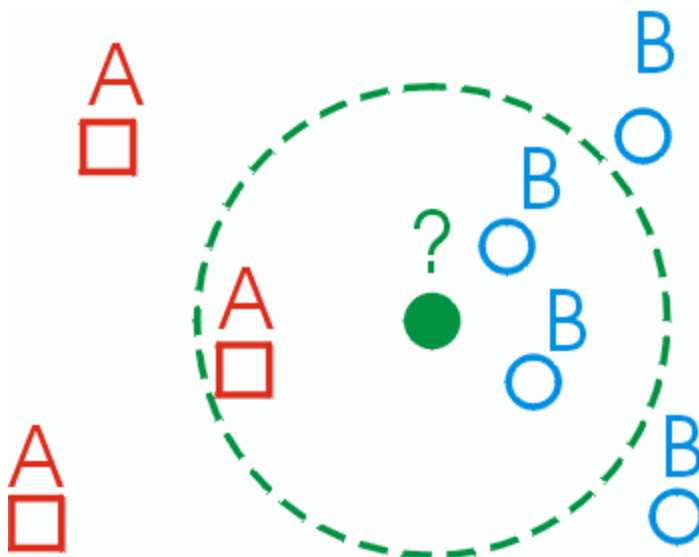
where weights  $w_{ij}$  are defined as

$$w_{ij} = \left[ 1 + \frac{d_{ij}^2}{\sum_{j=1}^k d_{ij}^2} \right]^{-1} \quad (4a)$$

and *k* is the number of the nearest neighbors (*k* = 1 to 9) of compound *i*,  $y_j$  is the class membership of compound *j* (1 or 2) and  $d_{ij}$  is the Euclidean distance between compound *i* and its  $j^{\text{th}}$  nearest neighbors. In practice, the value of  $\hat{y}_i$  is rounded to either 1 or 2 to determine the class membership of compound *i*:

$$\hat{y}'_i = \text{round}(\hat{y}_i) \quad (5)$$

The model is internally validated by leave-one-out cross-validation (LOO-CV) where each compound is eliminated from the training set and its class membership is predicted as the class the majority of its  $k$  nearest neighbors belong to. The descriptor set is optimized by simulated annealing approach with the Metropolis-like acceptance criterion to achieve the best  $\text{CCR}_{\text{train}}$  value.



**Figure 2.3. Prediction based on  $k$ -nearest neighbors.** When  $k=3$ , which is demonstrated here, the three nearest neighbor compounds of the query compound (green circle) are identified. Since two of them belong to class B, and they are closer (weight heavier) to the query compound than the class A compound, the external compound is predicted to be class B.

#### 2.4.1.2 Applicability Domain of $k$ NN Models

When developing QSAR models, each compound is represented as a point in  $M$ -dimensional descriptor space (where  $M$  is the total number of selected descriptors); thus, the molecular similarity between any two molecules can be characterized by the Euclidean distance between their representative points. The Euclidean distance  $d_{i,j}$  between two points  $i$  and  $j$  (which correspond to compounds  $i$  and  $j$ ) in  $M$ -dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (6)$$

Compounds with the smallest distance between one another are considered to have the highest similarity.

Theoretically, for any compound that can be represented by its chemical or physiological descriptors, one should be able to predict its class membership using the classification  $k$ NN approach. However, if the distance between the query compound and its  $k$  nearest neighbors in the training set is large, then the query compound is too dissimilar to the training set compounds, and the prediction of its activity using the  $k$ NN approach for this compound could be imprecise. Therefore, a similarity threshold (or model applicability domain) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules<sup>48</sup>. The similarity threshold is defined as follows:

$$D_T = \bar{y} + Z\sigma \quad (7)$$

Here,  $\bar{y}$  is the average Euclidean distance of the  $k$  nearest neighbors of each compound within the training set (where the value of  $k$  is the same as in predictive  $k$ NN QSAR models),  $\sigma$  is the standard deviation of these Euclidean distances, and  $Z$  is an arbitrary parameter to control the significance level. Typically, we set  $Z$  to 0.5, which places the boundary for deciding whether a compound is within or outside of the applicability domain at one half of the standard deviation. It is important to notice that increasing the value of  $Z$  would increase the number of compounds in the external set that are considered within the applicability domain but could decrease the accuracy of prediction due to inclusion of dissimilar nearest neighbors.

#### **2.4.3 Random forest (RF)**

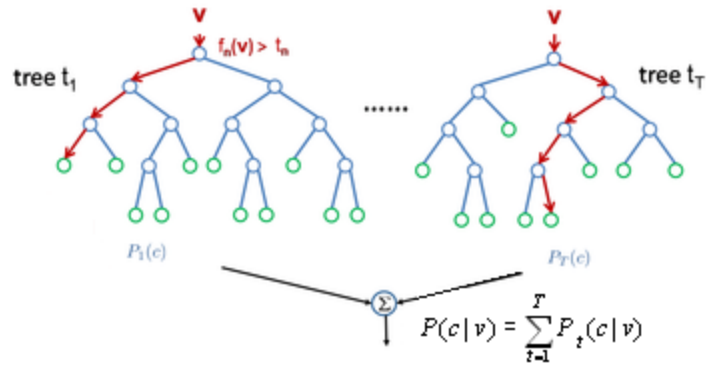
Random forest (RF) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler<sup>49</sup>. The term came from random decision forest that was first proposed by Tin Kam Ho of Bell Labs in 1995<sup>50</sup>. The method combines Breiman's "bagging" idea and the random selection of features, in order to construct a collection of decision trees with controlled variation. The selection of a random subset of features is an example of the random subspace method, which is a way to implement stochastic discrimination proposed by Eugene Kleinberg<sup>51</sup>.

Each tree is constructed as follows: 1) Let the number of training cases be  $N$ , and the number of variables in the classifier be  $M$ ; 2) The number  $m$  (much less than  $M$ ) represents the subset of  $M$ , and the number  $m$  of input variables are used to determine the decision at a node of the tree; 3) Choose a training set for this tree by choosing  $N$  times with replacement



from all  $N$  available training cases (i.e. take a bootstrap sample), and use the rest of the cases to estimate the error of the tree, by predicting their classes; 4) For each node of the tree, randomly choose  $m$  variables on which to base the decision at that node, and calculate the best split based on these  $m$  variables in the training set; 5) Each tree is fully grown and not pruned<sup>52</sup>.

In our studies, the R package of Random Forest (randomForest) was applied<sup>53</sup>, using default parameters.



**Figure 2.4. Prediction based on random forest algorithms.** The prediction output is the average results from all individual trees.

#### 2.4.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) was originally developed by Vapnik<sup>54</sup> as a general data modeling methodology where the training set error and the model complexity are incorporated into a special loss function and simultaneously minimized during model development. The importance of the prediction error versus the model complexity can be tuned during the optimization process, in order to generate models with reasonable complexity and avoid overfitting. SVM was later extended to afford the development of SVM regression models for datasets with non-integer variables.

We have implemented SVM for QSAR modeling as described earlier<sup>55</sup>. In brief, given a training set of pairs  $(x_i, y_i)$ ,  $i=1...m$ , where  $x_i$  is an array of descriptors of each compound and  $y_i$  is its biological activity (e.g., group label as binder or nonbinder), the sought correlation between structure and activity can be represented as  $y_i = f(x_i)$ . For simplicity, we define  $f(x_i)$  as a linear function:

$$f(x_i) = \langle w_i, x_i \rangle + b \quad (8)$$

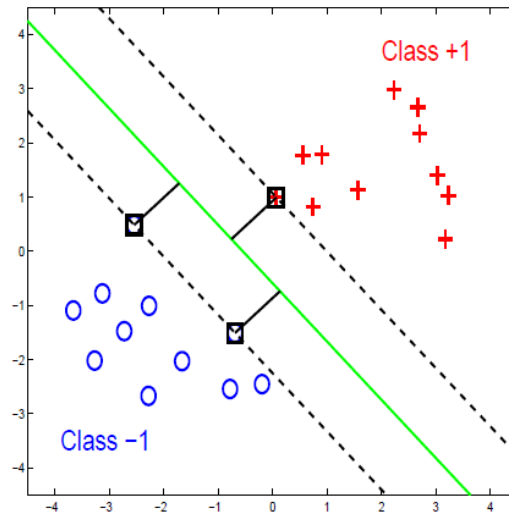
where  $w_i$  is the coefficient vector of the linear function and  $b$  is the bias. A major goal of the SVM regression algorithm is to minimize the loss function, which is a combination of prediction error defined by  $\xi_i$  and the magnitude of the coefficient  $C$  in the following equation:

$$loss_{\min} = \frac{\|w\|}{2} + C \sum_{i=1}^m \xi_i \quad (9)$$

with the constraint:

$$|y_i - (w\phi(x_i) + b)| = \xi_i \quad (10)$$

Here the training vectors  $x_i$  are mapped onto a high dimensional space by a kernel function  $\phi$ . In the end, SVM regression is expected to find a linear correlation between the actual activity and this high dimensional space  $\phi(x_i)$ . For this study, we have implemented a linear kernel.  $C$  is a penalty parameter of the error term that controls the weight between two terms in the SVM optimization process.



● □: support vectors

**Figure 2.5. Support Vector Machine (SVM) with maximum separation.**

#### 2.4.5 Robustness of QSAR models

The Y-randomization test is widely used to ensure model robustness<sup>56</sup>. It includes rebuilding the training set models using randomized activities (Y-vector) of the training set and comparing the resulting model statistics with that of the original test set. It is expected that models built with randomized activities should have significantly lower CCR values for both the training and test sets. In the model-building process, it is possible that sometimes, though infrequently, high CCR values may be obtained due to a chance correlation or structural redundancy of the training set. If QSAR models obtained in the Y-randomization test have relatively high LOO-CV  $CCR_{train}$  as well as predictive  $CCR_{test}$ , it implies that acceptable QSAR models cannot be obtained for the given dataset by the current modeling method. Herein, we applied the one-tail hypothesis to confirm the robustness of our QSAR models.

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  $CCR_{train}$  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, while the  $H_1$  hypothesis assumes the opposite, suggesting 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 three steps: (1) Determine the average value of  $CCR_{train}$  ( $\mu$ ) and its standard deviation ( $\sigma$ ) for random models; (2) Calculate the Z score that corresponds to the average value of  $CCR_{train}$  ( $h$ ) for the actual models using the following equation:

$$Z = (h - \mu) / \sigma \quad (12)$$

(3) Compare this Z score with the tabular critical values of  $Z_c$  at different levels of significance ( $\alpha$ )<sup>60</sup> to determine the level at which  $H_0$  should be rejected. If the Z score is higher than tabular values of  $Z_c$ , 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. In this study, the Y-Randomization test was applied to all data sets considered in this study, and the test was repeated twice in each case.

## **2.5 Background Information of Generic Multi-target Techniques**

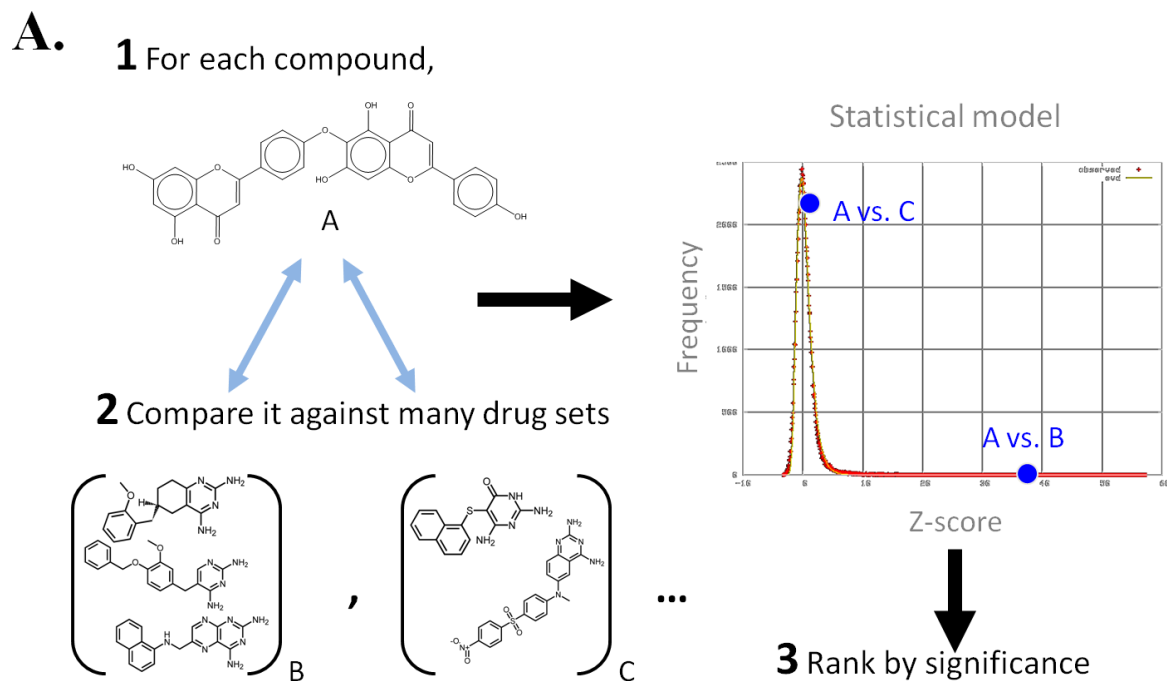
### **2.5.1 Similarity Ensemble Search (SEA)**

Similarity Ensemble Approach (SEA) is a recently developed cheminformatics tool used for predicting protein similarities as well as ligand's off-target binding<sup>57</sup>. It compares protein targets by the similarity of the ligands that bind to them, with prediction confidence expressed as expectation values, adapting the BLAST algorithms. For example, two subsets of the ligands in MDL Drug Data Report (MDDR) database are compared, which are annotated according to the receptor they modulate. Each ligand in each set was compared to each ligand in the other set. Tanimoto coefficients (Tc) of chemical similarity were calculated for each pair of ligands. Most ligand pairs with insubstantial similarity have the Tc range in 0.2 to 0.3. The compound pairs with Tc values over 0.57 are considered to be different. The raw similarity score of the set, which is the sum of ligand pair Tc over all pairs with  $Tc \geq 0.57$ , is calculated and the significance of it can be shown after correction for random expectation. Using a statistical model, any raw score for ligand sets of any size can be compared by Z-scores and expectation values (**Figure 2.6**). Therefore, protein comparison is realized by comparing the ligand sets that bind to them. Moreover, the prediction of

whether a ligand will bind to a specific target or not can be performed by comparing the ligand's structural similarity among the ligand set of the target. They claimed a confidence prediction is made when the expectation value  $\leq 10^{-10}$ .

In their paper published in Nature Biotechnology<sup>14</sup>, when screened internally against all MDDR molecular targets, SEA can recapitulate 19% of the off-targets binding reported in WOMBAT but missing from the MDDR. For new drug-target predictions, SEA successfully predicted the drug of N,N-dimethyltryptamine (DMT) strongly binds on serotonergic receptors, with experimental confirmation in a knockout mouse. The author stated that the chemical similarity approach is systematic and comprehensive, and may suggest side effects and new indications for many drugs.



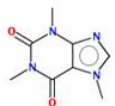


**B.**

sea viewer

**Query Info**

Code 152.19.136.137  
Name SEArch 152.19.136.137  
Size 1 (of 1)



[Back to Index](#)  
Page << 1 of 1 >>

	Code	#Ligands	Reference Name	E-value	MaxTC	Molecules
1	Ca(2+) channel+	92	voltage-dependent N-type calcium channel [+]	5.41e-71	1.00	<a href="#">view &gt;&gt;</a>
2	A2-	797	adenosine A2 receptor [-]	5.31e-39	1.00	<a href="#">view &gt;&gt;</a>
3	A1-	4193	adenosine A1 receptor [-]	3.40e-21	1.00	<a href="#">view &gt;&gt;</a>
4	GnoR-	31	guanosine binding site [-]	3.14e-15	1.00	<a href="#">view &gt;&gt;</a>
5	A2B-	927	adenosine A2B receptor [-]	3.16e-11	1.00	<a href="#">view &gt;&gt;</a>
6	A2A-	3127	adenosine A2A receptor [-]	5.87e-9	1.00	<a href="#">view &gt;&gt;</a>
7	A3-	2624	adenosine A3 receptor [-]	1.37e+0	1.00	<a href="#">view &gt;&gt;</a>
8	PDE4	43	cAMP-specific PDE	4.46e-54	0.55	<a href="#">view &gt;&gt;</a>
9	A2	157	adenosine A2 receptor	3.68e-14	0.71	<a href="#">view &gt;&gt;</a>
10	PDE4-	1375	cAMP-specific PDE [-]	1.93e-7	0.71	<a href="#">view &gt;&gt;</a>
11	PNP	13	purine nucleoside phosphorylase	8.49e-3	0.40	<a href="#">view &gt;&gt;</a>
12	TNFalpha-	743	tumor necrosis factor-alpha, cachectin [-]	2.36e-1	0.56	<a href="#">view &gt;&gt;</a>
13	H1-	809	histamine H1 receptor [-]	1.18e+0	0.46	<a href="#">view &gt;&gt;</a>
14	TPase-	50	thimidine phosphorylase [-]	1.74e+0	0.58	<a href="#">view &gt;&gt;</a>

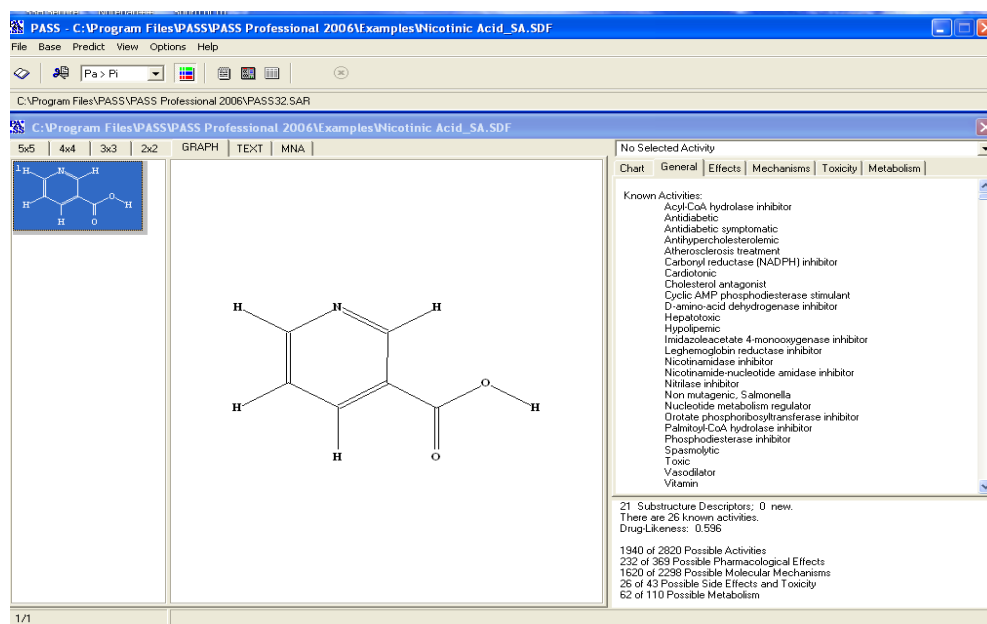
Note: Perfect matches always shown first.

**Figure 2.6. The algorithm for Similarity Ensemble Approach (SEA) (A) and its web-based platform (B).**

### ***2.5.2 Prediction of Activity Spectra for Substances' (PASS)***

The concept of Biological Activity Spectrum served as the basis for the 'Prediction of Activity Spectra for Substances' (PASS) software<sup>58</sup>. Different from other works, which focused on predicting a chemical compound's interaction with a specific biological entity, Dr. Poroikov's lab predicts each compound regarding a wide pharmaceutical and toxicological profile of activities, which is considered to be a 'biological activity spectrum of a substance'.

In PASS software version 2006, the prediction is based on a SAR analysis of the training set containing more than 60,000 compounds. It uses 'Multilevel Neighborhoods of Atoms' (MNA) as the chemical structure descriptors and the algorithm of the 'Activity Spectra Estimation'<sup>59</sup> as the training procedure. The result of a new compound is presented by the activity spectrum, which is the ranked list of the probabilities 'to be active'  $P_a$ , 'to be inactive'  $P_i$ , and the type of activity. A compound is considered active if the value  $(P_a - P_i)$  exceeds the cutoff value, e.g., by default  $(P_a - P_i) < 0.0$ <sup>59</sup>. Also, Poroikov's group states that if  $P_a > 0.7$ , the compound is very likely to reveal this activity in experiments, but in this case the chance of being the analogue of the known pharmaceutical agents for this compound is also high; If  $0.5 < P_a < 0.7$  the compound is likely to reveal this activity in experiments, but this probability is less, and the compound is not so similar to the known pharmaceutical agents; If  $P_a < 0.5$  the compound is unlikely to reveal this activity in experiments, but if the presence of this activity is confirmed in the experiment the compound might be a New Chemical Entity.



**Figure 2.7. The software platform for the Prediction of Activity Spectra for Substances (PASS).**

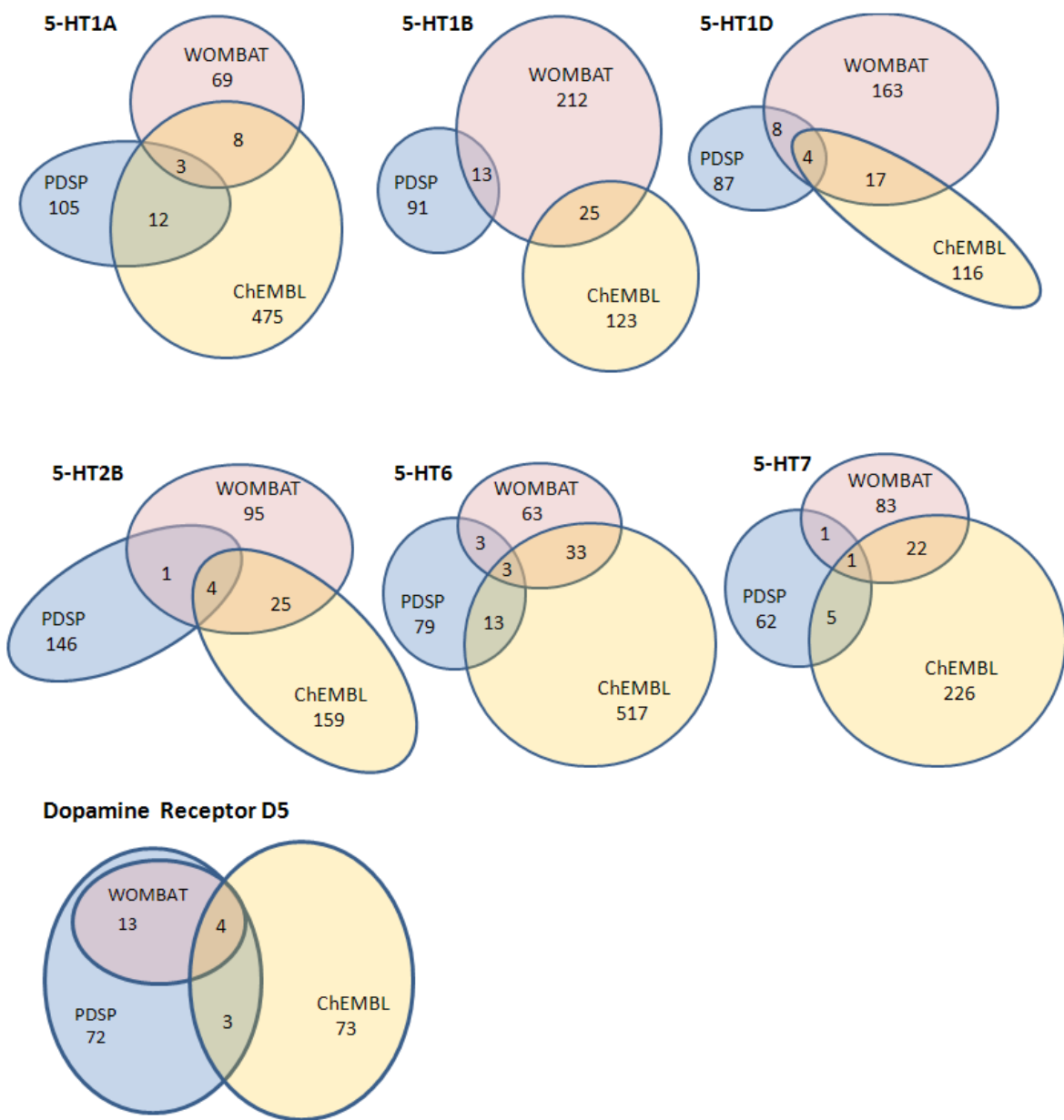
## 2.6 Comparison of the generic multi-target technique versus single target model building

We comprehensively compared and validated our *k*NN-QSAR technique with other popular multi-target prediction approaches, which include Similarity Ensemble Approach (SEA) and Prediction of Activity Spectra for Substances (PASS). Biological targets involved in the comparison included 7 GPCR receptors: 5-HT1A, 5-HT1B, 5-HT1D, 5-HT2B, 5-HT6, 5-HT7, and dopamine receptor D5.

Datasets used for model building and external prediction were extracted from PDSP, WOMBAT and ChEMBL. These are leading small molecule chemogenomics databases, which contain 2D/3D structures, calculated properties (e.g. logP, Molecular Weight, eTc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). Information were primarily collected from reliable publications (e.g. Journal of Medicinal Chemistry) so the labeling was considered accurate.

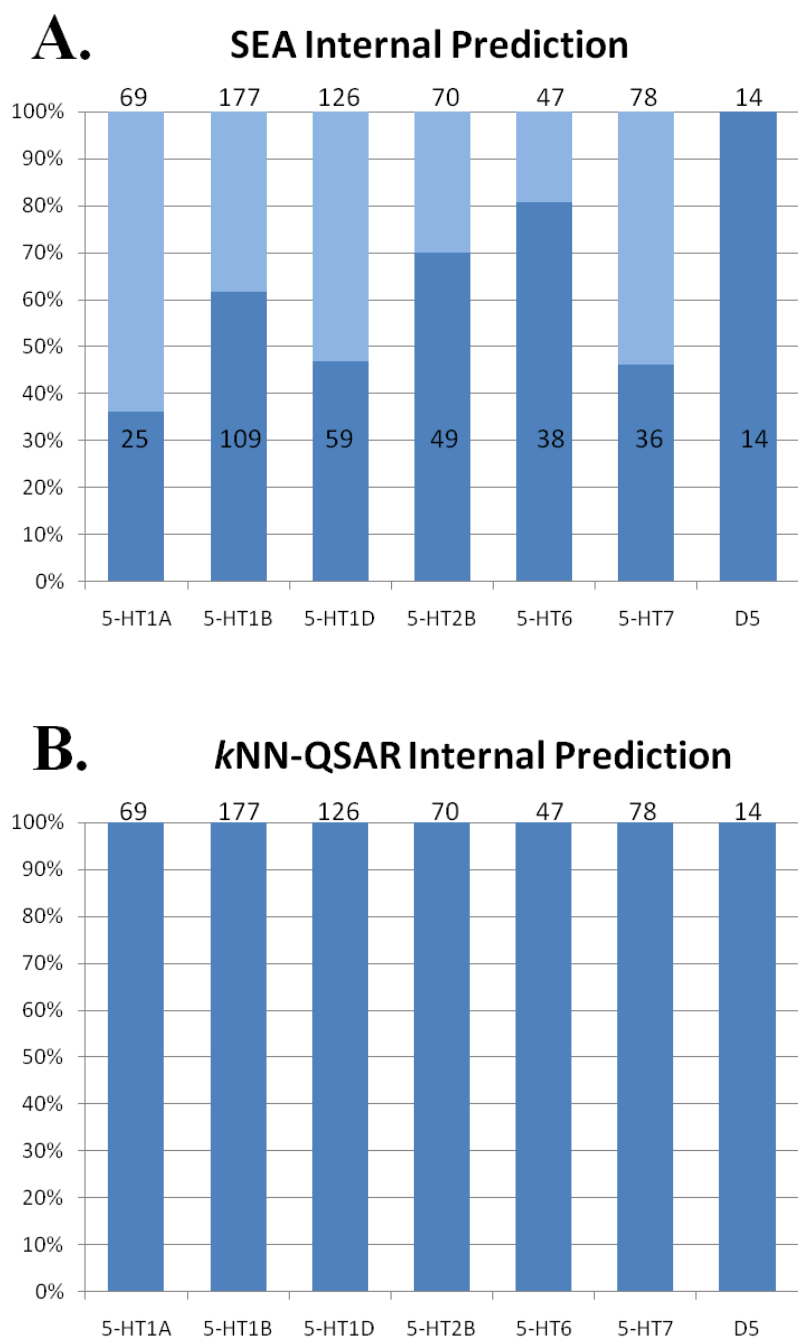
The prediction comparison was conducted for both the internal compounds' retrieval rate as well as the external compounds' prediction accuracy for seven cases. Since compounds with non-binder labeling were available only in the PDSP database, those non-binders have to be "shared" with binders from both PDSP and WOMBAT to generate classification *k*NN-QSAR models. For the case of 5-HT1A dataset, classification QSAR models were generated separately for the two combinations of datasets: 105 binders and 61 non-binders from PDSP, and 69 binders from WOMBAT and the same 61 non-binders from PDSP. Since SEA used WOMBAT data as their modeling set, the predictive power comparison between classification QSAR models generated for binders from WOMBAT (with non-binders from PDSP) and SEA approach was considered strictly fair. We compared

the internal recovery rate of modeling set compounds performed by *k*NN-QSAR versus SEA using the same training set compounds (both from WOMBAT). To validate their external prediction accuracy, we applied the individual *k*NN-QSAR and SEA models to predict unique ligands deposited in ChEMBL. Note that only ligands covered by neither WOMBAT nor PDSP were used. The data deposition of binders (including agonists and antagonists) for those 7 biological targets in PDSP, WOMBAT and ChEMBL were shown in **Figure 2.8**.



**Figure 2.8. Data disposition of 7 different GPCR targets in PDSP, WOMBAT and ChEMBL.**

The internal recovery rate and the external prediction accuracy are shown in **Figure 2.9-2.12**. The internal recovery rate of SEA was around or less than 70%, except for the case of dopamine receptor D5, which contained only 14 compounds; however, *k*NN-QSAR always achieved 100% for its internal recovery rate. Though only seven cases were tested here, we considered these results to be representative for the other cases, as well. The reason for the much lower recovery rate SEA had compared to *k*NN-QSAR may due to the fact that only a simple similarity comparison was involved when making a prediction. The prediction results would be significantly biased, especially when a set of known ligands for a receptor are unevenly distributed in the chemical space. In **Figure 2.9 (A)**, for example, if the 44 wrongly predicted 5-HT1A binders were diversely distributed in the chemical space, but far from the other 25 binders, if only the similarity comparison is considered in the prediction process, then those 44 compounds would much more likely to be falsely predicted. Since model building is involved in *k*NN-QSAR method, with rigorous model validation procedures, such false predictions would not commonly occur in QSAR.

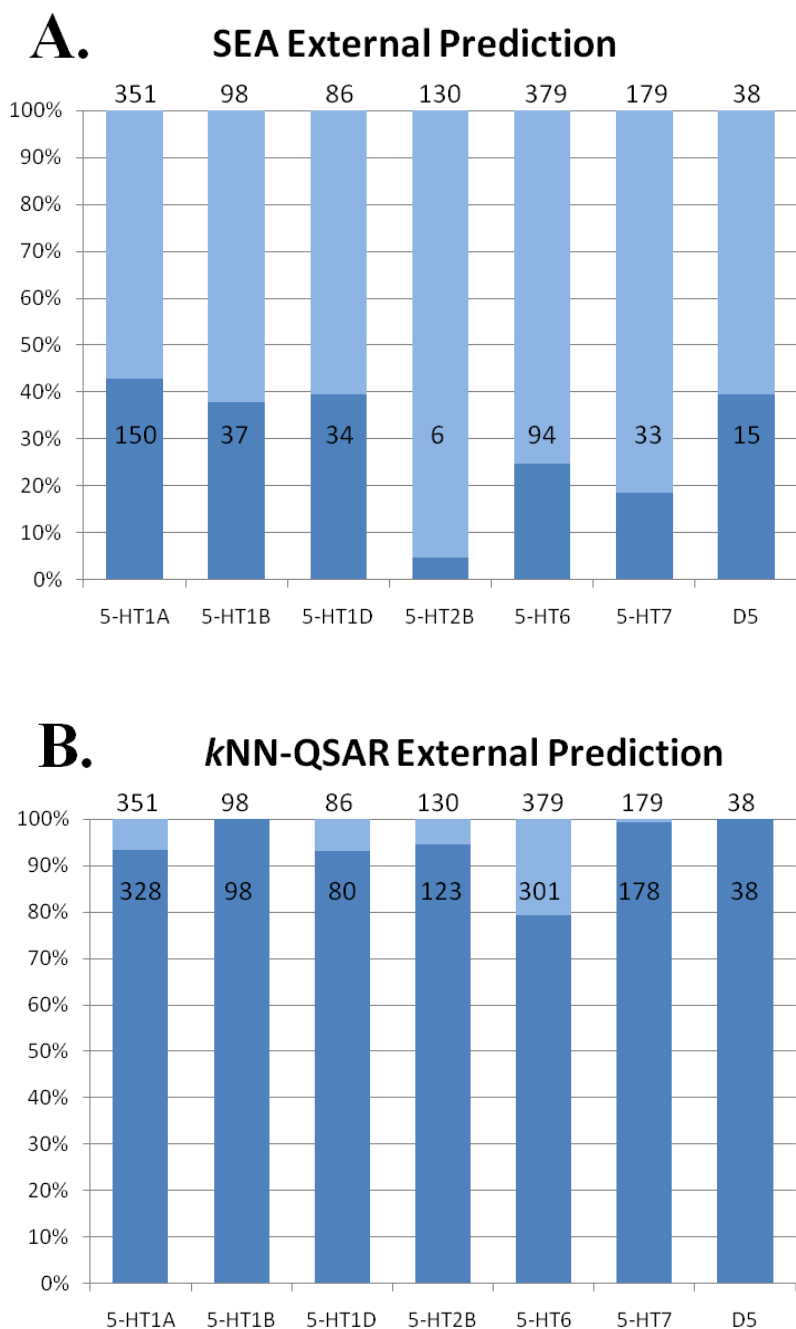


**Figure 2.9. Comparison of SEA and *k*NN-QSAR on internal prediction.** *k*NN-QSAR models were built for those 7 GPCR targets individually, and the numbers of binders in each modeling sets were shown on top of each bar. The numbers of binders which were correctly recovered by either SEA (A) or *k*NN-QSAR (B) were shown in dark blue.

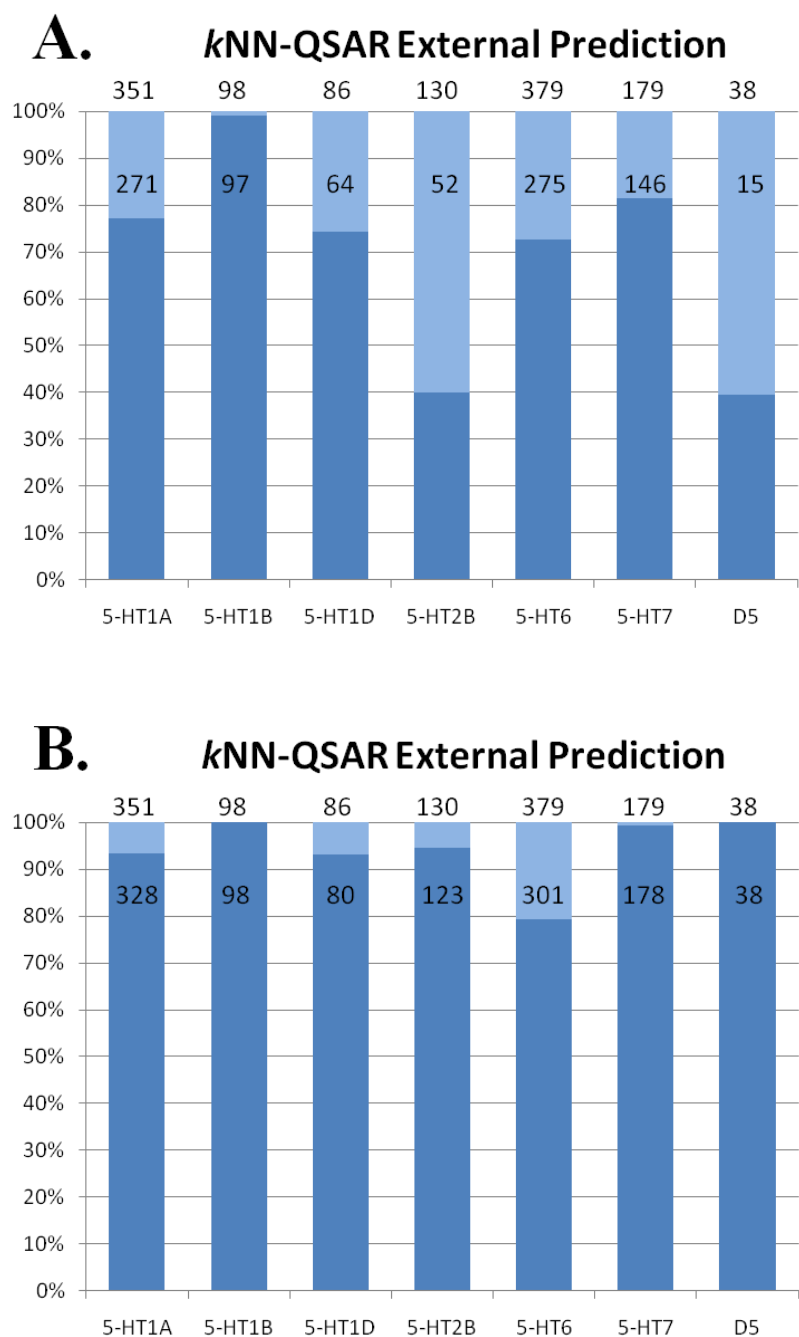


For the predictions of external compounds, the results further highlight the advantages of using QSAR over SEA and PASS. SEA, being the worst method, can only predict less than half of the external validation set compounds, while QSAR achieves around 90% for most of the seven cases. The prediction accuracy for PASS is also much higher than SEA, though not as good as QSAR. This is because a simple linear regression method is used in PASS for its internal model building process. It was noticed that QSAR models generated for binders and non-binders from consistent data source (PDSP) have higher external predictive power than models generated for binders and non-binders from mixed sources (binders from WOMBAT and non-binders from PDSP) (**Figure 2.11**). The reason for that is currently unknown, but we hypothesized it to be related to models' applicability domains (AD). This hypothesis was further confirmed by results shown in **Figure 2.11 (A)** and **Figure 2.12 (B)**. Both showed the external predictive power for models generated for binders from WOMBAT and non-binders from PDSP; however, the external prediction accuracy dramatically increased along with the application of AD. Compounds in the ligand sets from WOMBAT may be very similar with each other, thus the models built for them have only limited AD.

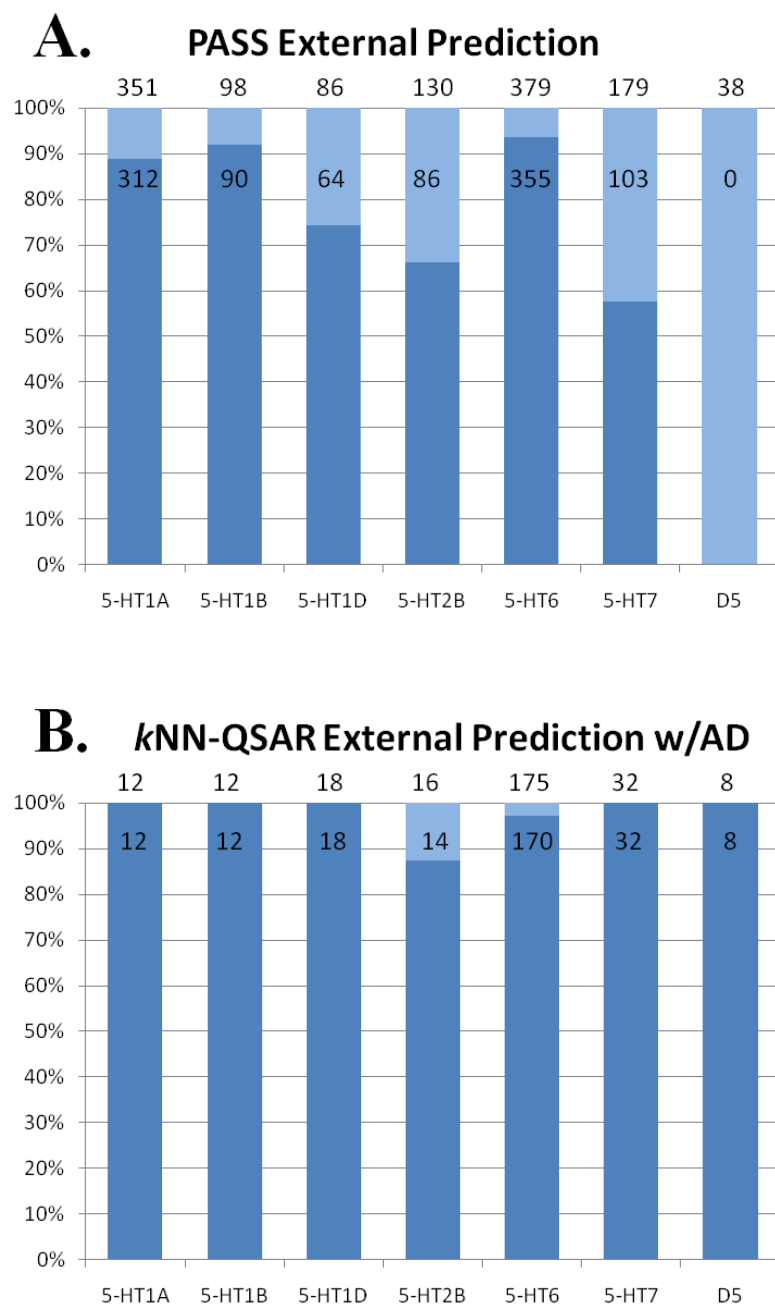
Because the modeling set compounds for PASS cannot be traced, not necessarily the compounds in WOMBAT database, so the model comparison between PASS vs. SEA and QSAR were not strict enough. But overall, the performances of PASS models are much better than SEA, though not as good as QSAR (**Figure 2.13**).



**Figure 2.10. Comparison of SEA and QSAR on external prediction.** *k*NN-QSAR Models were generated for binders and non-binders from PDSP. Unique compounds from ChEMBL were used as external tests to compare the external predictive power of SEA and QSAR. The numbers of compounds in each external test sets are shown on top of each bar and the numbers of compounds correctly predicted by either SEA (A) or QSAR (B) are shown in dark blue.



**Figure 2.11. Comparison of *k*NN-QSAR external prediction for different data sources.** *k*NN-QSAR models are generated separately for binders from WOMBAT with non-binders from PDSP (A), and binders/non-binders from PDSP (B). Unique compounds from ChEMBL were used as external tests to compare the external predictive power QSAR generated from different data sources. The numbers of compounds in each external test sets are shown on top of each bar, and the numbers correctly predicted by these two types of models are shown in dark blue.



**Figure 2.12. Comparison of PASS and *k*NN-QSAR on external prediction.** *k*NN-QSAR models were generated on binders from WOMBAT and non-binders from PDSP with applicability domains applied. Unique compounds from ChEMBL were used for external tests to compare the external predictive power of PASS and *k*NN-QSAR. The numbers of compounds in each external test sets are shown on top of each bar in (A), and the numbers of compounds within models' AD were shown on top of each bar in (B). The numbers of compounds correctly predicted by either PASS (A) or *k*NN-QSAR (B) are shown in dark blue.

## **CHAPTER 3**

### **VALIDATION OF THE CONGRUENCE OF DATA IN DISPARATE SOURCES AND IDENTIFICATION OF LOW QUALITY DATA**

#### **3.1 Introduction**

The rapidly growing number of chemical annotations deposited in various databases represents the arrival of cheminformatics era. However, a significant gap exists between the huge amount of data and our capability of data processing and analysis. Herein, I will apply cheminformatics approaches to validate the congruence of data deposition in several databases and to identify low quality (by means of low signal-to-noise ratio) data sources. The success of my study will not only shed light on the consistency and quality of major annotated cheminformatics databases, but also provide solid base for the high quality QSAR modeling.

In my work, QSAR models for the same biological targets are generated separately for data obtained from different sources, and are applied for inter-database cross-validation. Data that models suggest to be dubious are experimentally validated at UNC. These results contribute to the development of cheminformatics approaches and influence the process of drug discovery by stressing the importance of determining trustworthy data sources and highlighting the significance of predictive QSAR modeling.

## 3.2 The Major Annotated Chemogenomics Databases

### 3.2.1 PubChem database and the NIH Molecular Libraries Roadmap Initiative

Early stages in modern drug discovery often involve screening small molecules for their effects on a selected protein target or a model of a biological pathway. Innovative technologies developed in recent years enable rapid synthesis and high throughput screening of large libraries of compounds in industry. As a result, there is a huge increase in the number of compounds available on a routine basis to quickly screen for novel drug candidates against new targets/pathways. In contrast, such technologies have rarely become available to the academic research community, thus limiting the academia to conduct large scale chemical genomics research. Launched in 2004, the NIH Molecular Libraries Roadmap Initiative<sup>60</sup> aims to change this situation by integrating multiple chemical libraries and screening centers. A salient feature of the NIH Initiative is that the Centers are interested in any chemicals that may affect a biological pathway or function, regardless of their potential to become drugs. This feature has been elegantly demonstrated by Schreiber, *et al.* as the unique aspect of chemical genetics or chemical genomics research<sup>61</sup>. From a biological perspective, the NIH Initiative would like to address a broader diversity space as well, to be able to interrogate any biological pathways or networks with small molecule effectors<sup>62</sup>.

PubChem<sup>1</sup>, as an essential component of NIH's Molecular Libraries Roadmap Initiative, is the largest chemical database in public domains. As of recently, it includes information on more than 31 million chemical structures and bioactivity results from 1644 high-throughput screening programs. For each entry, it has the links to bioassay descriptions, literatures, references, and assay data. The BioAssay database in PubChem provides

searchable descriptions of over 1,000 bioassays, including descriptions of the conditions and readouts specific to a screening protocol.

However, many cheminformatics experts still are skeptical about PubChem data quality<sup>63</sup>. Many suggest that the major problem comes from the misguided perception that cheminformatics software can be directly used to address the needs of the MPL. Simply providing a place to deposit data is not sufficient to ensure its optimal use. The data as deposited in PubChem are not curated by screening centers. Identification of the PubChem data with rather low signal-to-noise ratio requires very thorough and laborious cheminformatics approaches.

### **3.2.2 NIMH Psychoactive Drug Screening Program (PDSP)**

NIMH Psychoactive Drug Screening Program (PDSP)<sup>64</sup> Ki Database (or Ki DB) is a public database of published binding affinities (specifically, Ki) of drugs and chemical compounds for receptors, neurotransmitter transporters, ion channels, and enzymes. This resource is maintained mainly by Dr. Brian Roth's lab in the University of North Carolina at Chapel Hill, and is funded by the NIMH Psychoactive Drug Screening Program (NIMH-PDSP) as well as by a gift from the Heffter Research Institute. It can be accessed at <http://pdsp.med.unc.edu/>. The Ki DB is the world's largest openly available database of ligand receptor affinity data and currently houses >47, 000 Ki values on >500 molecular targets. Ki-DB represents a curated, fully-searchable database of both published data and data internally derived from the NIMH-PDSP.

### **3.2.3 The World of Molecular Bioactivity (WOMBAT)**

The company supporting the publishers of World of Molecular Bioactivity (WOMBAT) is founded by Dr. Tudor I. Oprea M.D., Ph.D. in 2002. WOMBAT serves as a leading small molecule chemogenomics database, providing high quality information of small molecule bioactivity annotations<sup>3</sup>. The database is a collection of chemical annotations published in top medicinal chemistry journals such as Bioorganic & Medicinal Chemistry, Journal of Medicinal Chemistry, etc. Therefore, it is stated that although the compounds are tested in different laboratories, they should have robust and confident binding results.

### **3.2.4 Chemical European Molecular Biology Library (ChEMBL)**

ChEMBL is a database of bioactive drug-like small molecules, it contains 2D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and extracted bioactivities (e.g. binding constants, pharmacological and ADMET data). It attempts to normalize the bioactivities into a uniform set of end points and units. Also, when it is possible, the bioactivities are linked to published assays which are given varying confidence levels. The data is extracted and curated from primary scientific literature, and cover a significant fraction of the modern drug discovery space. As ongoing effort, additional data on clinical trial progress of compounds are being integrated into ChEMBL.

## **3.3 Different Data Sources of 5-Hydroxy Tryptamine receptor subtype 1A (5-HT1A)**

### **Ligands**

#### **3.3.1 5-HT1A Agonists and Antagonists from PubChem**

The Scripps Research Institute Molecular Screening Center deposited HTS data for 5-HT1A agonists and antagonists into PubChem (PubChem AID: 613, 718, and 755). AID718



and AID613 are confirmatory dose response assays for 5-HT1A agonists, while **AID755** is a confirmatory dose response assay for 5-HT1A antagonists. Compounds identified from a previously described set of experiments entitled "Primary HTS assay for 5-Hydroxytryptamine (Serotonin) Receptor Subtype 1a (5HT1a) agonists" and "Primary Cell Based High Throughput Screening Assay for Agonists of the 5-Hydroxytryptamine Receptor Subtype 1E (5HT1E)" were selected for testing in these assays. A cell line containing the human 5-HT1A receptor, the promiscuous G-alpha-15 protein (Ga15), and the beta-lactamase (BLA) reporter gene under control of the nuclear factor of activated T-cells (NFAT) promoter was used to measure 5-HT1A agonism. The amount of BLA activity is proportional to the concentration of agonist, which was measured with a fluorescent BLA substrate. All experimental details are available online

(<http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcassay&term=613>).

In cases where the highest concentration tested (~50 mM) did not result in more than 50% inhibition or where no curve fit could be achieved, the EC<sub>50</sub> was determined manually by the observed inhibition at the individual concentrations. Compounds with EC<sub>50</sub> values greater than 10 mM were considered inactive, and compounds with EC<sub>50</sub> equal to or less than 10 mM were considered active. A conservative estimate of the activity score for each compound was calculated for which no exact EC<sub>50</sub> value was given while maintaining a reasonable rank order of all compounds tested.

The experimental assays' protocols are listed below, which were extracted from PubChem website.

“**AID718** details the results of a confirmatory screening bioassay for agonists of 5-Hydroxytryptamine (Serotonin) Receptor Subtype 1A (5HT1A) conducted in the Scripps

Research Institute Molecular Screening Center. It is a relatively small dataset with 51 compounds and with a ratio of 1 active to 6 inactive compounds (15.9% minority class). The compounds were selected on the basis of primary HTS assay for agonists of the 5-Hydroxytryptamine Receptor Subtype 1E (5HT1E) of approximately 64,900 small molecules (PubChem AID = 574).

**AID613** is a confirmatory screening assay from the Scripps Research Institute Molecular Screening Center for 5-Hydroxytryptamine (Serotonin) Receptor Subtype 1a (5HT1A) agonists, which consists of 346 compounds with a ratio of one active compound to 19 inactive compounds (5.2% minority). The compounds have been selected for their activation results in the Primary HTS assay for 5-Hydroxytryptamine (Serotonin) Receptor Subtype 1a (5HT1A) agonists of approximately 64,900 compounds (PubChem AID = 567).

**AID755** is the result of a confirmatory screen for Antagonists of the 5-Hydroxytryptamine Receptor Subtype 1A (5HT1A) from the Scripps Research Institute Molecular Screening Center. It contains activity information of 44 compounds with a ratio of one active compound to 0.76 inactive compounds (131.6% minority). The screen is a reporter-gene assay and the tested 44 compounds were selected from the primary HTS assay for 5-Hydroxytryptamine (Serotonin) Receptor Subtype 1a (5HT1A) agonists and antagonists of approximately 64,900 compounds (PubChem AID = 567 and 574).”

We assume that all agonists and antagonists are 5-HT1A binders. There are seven agonists identified through AID718, and 17 from AID613, however, 3 of them were identical structures. So 21 agonists and 25 antagonists were confirmed by PubChem dose response

assay in total (there is no shared chemicals identified to be agonists and antagonists at the same time).

### **3.3.2 5-HT1A Binders and Non-binders from PDSP**

The NIMH Psychoactive Drug Screening Program (PDSP) database, which is maintained by Dr. Bryan Roth's lab, collects primary and secondary binding assays for many GPCR receptors, including 5-HT1A receptor. In their assays, each compound was added quadruplicate to 96-well format plate and radioligand was used to measure the fraction of the target compound that binds with the crude membrane fractions of cells expressing recombinant 5-HT1A receptors. Radioligands were purchased by PDSP from Perkin-Elmer or GE Healthcare. Competition binding assays were performed using transiently or stably expressing cell membrane preparations as previously described<sup>65, 66</sup>, and are available online (<http://pdsp.med.unc.edu>). The radioligand used for 5-HT1A binding assays was [<sup>3</sup>H]-8-OH-DPAT. All experimental details are available online (<http://pdsp.med.unc.edu/UNC-CH%20Protocol%20Book.pdf>).

All chemical structures and binding affinity information were deposited in PDSP Ki database, in which we retrieved the 5-HT1A binders and non-binders using their online search engine. In this study, we used 1mM as the cutoff value to define binders vs. non-binders, and we only retrieve ligands tested against cloned human cell lines using the hot ligand of [<sup>3</sup>H]-8-OH-DPAT. By submitting such queries, 105 unique compounds were extracted to be 5-HT1A binders, with binding affinity less than 1mM. There are 61 non-binders which were shown to have no binding to the 5-HT1A receptor at 1mM concentration, but share relatively high structural similarity with those 105 binders.

### **3.3.3 5-HT1A Binders from WOMBAT**

World of Molecular Bioactivity, WOMBAT, served as a leading small molecule chemogenomics database and the standard of quality in small molecule bioactivity annotations<sup>67</sup>. Compounds were tested in different labs but with robust and confident binding results. 5-HT1A binders were extracted from WOMBAT when they satisfied all the following thresholds: 1) Compounds were tested on cloned human species cell lines; 2) [<sup>3</sup>H]-8-OH-DPAT were used as the hot ligand in the experiments; 3) The compounds' binding affinities were lower than 1mM concentration. In such case, only 60 unique compounds were chosen.

## **3.3 Methods**

### **3.3.1 Datasets Curation and Descriptors Generation**

The SMILES<sup>68</sup> strings of each compound in 5-HT1A dataset were converted to 2D chemical structures using the Molecular Operating Environment (MOE) software package. The Dragon<sup>7</sup> software (version 5.5) was used to calculate a wide range of topological indices of molecular structure. These indices include (but are not limited to) the following descriptors: simple and valence path, cluster, path/cluster and chain molecular connectivity indices<sup>69-71</sup>, kappa molecular shape indices<sup>72, 73</sup>, topological and electrotopological state indices<sup>74-76</sup>, differential connectivity indices, graphs' radius and diameter<sup>77</sup>, Wiener and Platt indices, Shannon and Bonchev-Trinajstić information indices, counts of different vertices, counts of paths and edges between different kinds of vertices.

Overall, Dragon produced over 2,000 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, only 880 chemically relevant descriptors were initially calculated and 672 descriptors were eventually used for 5-HT1A binding datasets after deleting descriptors with zero value or zero variance. Dragon descriptors were range scaled prior to distance calculations since the absolute scales for Dragon descriptors can differ by orders of magnitude<sup>78</sup>. Accordingly, our use of range-scaling avoided giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional Dragon descriptor space.

### ***3.3.2 Training, Test, and External Validation Set Selection***

We have followed the rigorous QSAR workflow for model building, validation and database mining (**Figure 2.2**) established in our laboratory (see <sup>79</sup> for recent overview). For classification QSAR modeling, it would be ideal to have the balanced ratio between different compound classes in the modeling dataset. However, the 5-HT1A binding dataset from PDSP includes 105 inhibitors and 61 non-binders, i.e., thus it is imbalanced with the inhibitors to non-binders ratio of 5:3; while 69 compounds from WOMBAT are all binders, and the PubChem dataset contains 46 binders and 389 non-binders. In the absence of special statistical treatment, such imbalanced datasets will distort the QSAR model prediction; therefore, we conducted different strategies for model building on those multiple datasets. For PDSP [105/61] dataset, we do not want to lose any 5-HT1A binding data information, so different weights for 5-HT1A binders and non-binders were employed during the modeling process. As for 69 WOMBAT binders, we combine them with PDSP non-binders (61 in total) [69/61] to build classification QSAR models. Because PubChem dataset has the largest

inhibitors to non-binders ratio, being roughly 1:8, such a ratio would largely skew the prediction accuracy of the classification models in the absence of special statistical treatment. Thus, the distance matrix was calculated in the multidimensional descriptor space for all 435 compounds and a similarity search was carried out using 46 inhibitors as queries against the remaining 389 non-binders. 58 compounds were selected from the original 389 non-binders as most similar to the 46 inhibitors using  $Z_{\text{cutoff}}$  value of 0 (we note that this treatment makes the task of building the discriminatory binary QSAR models even more challenging). Consequently, these 58 non-binders combined with 46 PubChem inhibitors formed a new balanced dataset for QSAR model building. Furthermore, the five-fold external set cross validation protocol was conducted when building models for those three datasets. For each of the five fold, one-fifth of the compounds from the total 166 were randomly chosen as one external validation set, so that each compound will be in the external validation set once and only once. The remaining four-fifth of the compounds were considered a modeling dataset that was divided into multiple diverse and representative training and test sets using the Sphere Exclusion approach developed in our laboratory earlier<sup>80, 81</sup>.

### ***3.3.3 QSAR Models Generated from Disparate Data Sources***

We followed the conventional  $k$ NN classification model building workflow developed in our lab to generate high quality QSAR models, herein, the value of 0.8 was used for both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$ . RF was conducted in the R software using the package of RF (randomForest). SVM was conducted in the WinSVM software available in our lab, which implemented the conventional LIBSVM methodology. For all three methods, the same modeling set compounds were used to generate models, and the same external validation set compounds were applied, as well.

### ***3.3.2 QSAR Model-Based Cross Validation between Disparate Data Sources***

After building Classification QSAR models on binders/non-binders from PDSP, binders in WOMBAT with non-binders in PDSP, and binders/non-binders from PubChem, and these models were used to cross validate 5-HT1A compounds between different sources, i.e. the combinatorial QSAR models generated by  $k$ NN, RF and SVM on binders/non-binders from PDSP were used to validate 69 binders from WOMBAT and 46 agonists/antagonists from PubChem (PubChem Assay ID (AID): 613, 718 and 755). The application domain was also applied using  $Z_{\text{cutoff}} = 0.5$ . Some of the validation results for PubChem confirmatory compounds were set as non-binders by our models and were further experimentally tested in Dr. Bryan Roth's lab at UNC, Chapel Hill.

## **3.4 Results and Discussions**

### ***3.4.1 QSAR Classification Models***

The  $k$ NN QSAR method with variable selection afforded models with optimal accuracy characterized as CCR for both training and test sets. For  $k$ NN classification models

generated on 105 5-HT1A binders and 61 non-binders from PDSP, there were 838 models with both  $CCR_{train}$  and  $CCR_{test}$  equal or higher than 0.80. Most models with  $CCR_{test} \geq 0.80$  also had corresponding  $CCR_{train} \geq 0.80$ , but the opposite was not always true. The models with high values of both  $CCR_{train}$  and  $CCR_{test}$  ( $\geq 0.80$ ) were considered acceptable and were selected for consensus prediction. The  $CCR_{train}$  and  $CCR_{test}$  for the best kNN model are 0.91 and 0.99 respectively, implying that the model could correctly identify 51 binders out of 55 and 34 out of 38 non-binders ( $SE = 0.93$ ,  $SP = 0.89$ ,  $EN(1) = 1.80$ , and  $EN(2) = 1.85$ ) in the training set and almost all binders and non-binders in the test set. This remarkably high internal accuracy and the large number of acceptable models imply that the kNN classification method was generally successful in correctly distinguishing binders vs. non-binders using Dragon chemical descriptors.

For kNN classification models generated on 69 5-HT1A binders from WOMBAT and 61 non-binders from PDSP, both the internal and external model statistics are even higher. In total, there were over 30,000 models with both  $CCR_{train}$  and  $CCR_{test}$  equal or higher than 0.80, in which 6,234 models were over 0.90. Due to the extremely large number of qualified models, we chose to only use models with  $CCR_{train}$  and  $CCR_{test}$  equal or higher than 0.90 to perform further five-fold cross validation.

QSAR models generated for agonists/antagonists with non-binders from PubChem database had much worse statistics than the previous two. There were in total 7,852 models having both  $CCR_{train}$  and  $CCR_{test}$  equal or higher than 0.50, and only 123 models higher than 0.80. These statistics were very close to those obtained with Y-randomization test, which suggests rather low model quality and confidence.



### 3.4.2 Validations of QSAR Classification Models

In addition to the internal validation of  $k$ NN models using test sets, Y-randomization and external validation are the critical steps of the entire QSAR workflow (Figure 1). Only models that have been validated by these two steps can be utilized for external prediction and validation<sup>48</sup>.

#### 3.4.2.1 Y-randomization Test

In the Y-randomization test, the binary annotations of 5-HT1A as binders or non-binders were randomly shuffled and  $k$ NN, RF and SVM classification models were built with the same parameter setting. The test was performed three times for each training/test set split for datasets from PDSP, WOMBAT and PubChem, respectively. All runs of Y-randomization tests showed that there were relatively small number of models having both  $CCR_{train}$  and  $CCR_{test}$  higher than 0.70. However, there were only few (less than five) models with both CCR values higher than 0.80 (Figure 2). Notice that the CCR values generated in Y-randomization tests using either PDSP or WOMBAT datasets were much worse than the ones using real binding affinities. In contrast, the CCR values generated in PubChem model building based on either real or randomized activities were very close. It implied that the QSAR models obtained from PDSP and WOMBAT datasets with both CCRs greater than 0.80 were robust models, while models built for PubChem data were not reliable. As we will show, these results can also mean that data from PDSP and WOMBAT are more reliable than PubChem data depositions.

#### 3.4.2.2 External Set Validation

The five-fold cross-validation (see above) was employed to evaluate the predictive power of QSAR models. Consensus prediction by all  $k$ NN models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  greater than 0.8 was carried out, and the final prediction score was the average of predicted class labels over all models for which a compound was within the AD, which was defined by  $Z_{\text{cutoff}}$  value of 0.5. Also, a compound was considered out of the AD if it was predicted by less than 50% of the models. For validation of models generated by Random Forest and Support Vector Machine, exactly the same five-fold external cross-validation sets were used. The average prediction results for all five-fold predictions of dataset from PDSP, WOMBAT and PubChem were summarized in **Table 3.1** and the detailed prediction statistics for each five-fold were available in Table S1 of supporting materials. Under  $Z_{\text{cutoff}} = 0.5$ , most  $CCR_{\text{evs}}$  achieved a rather high prediction accuracy for datasets from PDSP and WOMBAT. Those 5-HT1A binders falsely predicted (average class number  $> 1.5$ ) were within an applicability domain of only a small portion of qualified  $k$ NN models, which were usually considered as unreliable predictions. In summary, QSAR models generated for 5-HT1A dataset from PDSP and WOMBAT with both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  equal to or greater than 0.80 gave high consensus prediction accuracy in the corresponding five-fold external validation.

On the other hand, the five-fold cross validation statistics for QSAR models generated on PubChem 5-HT1A data were much worse than PDSP and WOMBAT data, shown in **Table 3.1**. The average five-fold  $CCR_{\text{evs}}$  was only 0.59, statistically not different than the  $CCR_{\text{evs}}$  value of 0.48, which was obtained in the Y-randomization test. These results implied that the QSAR models generated on PubChem data were not robust and not predictive.

**Table 3.1.** External validation statistics for disparate 5-HT1A datasets by different QSAR methods

Datasets	Machine Learning Methods	Prediction CCR	Confusion Matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)
PDSP	k-Nearest Neighbor	0.76 (0.111)	104*	61	94	38	23	10	0.89 (0.084)	0.61 (0.177)	1.41 (0.212)	1.70 (0.229)
	Random Forest	0.80 (0.095)	105	61	93	41	20	12	0.88 (0.099)	0.68 (0.098)	1.46 (0.122)	1.72 (0.206)
	Support Vector Machine	0.80 (0.106)	105	61	98	41	20	7	0.93(0.043)	0.68(0.176)	1.50(0.229)	1.80(0.142)
WOMBAT	k-Nearest Neighbor	0.93(0.034)	69	61	67	54	7	2	0.97(0.034)	0.90(0.046)	1.79(0.086)	1.94(0.080)
	Random Forest	0.91(0.050)	69	61	65	54	7	4	0.94(0.064)	0.89(0.042)	1.78(0.082)	1.88 (0.129)
	Support Vector Machine	0.87(0.089)	69	61	59	54	7	10	0.85(0.095)	0.89(0.086)	1.74(0.132)	1.72(0.184)
PubChem	k-Nearest Neighbor	0.59(0.110)	45*	58	24	38	20	21	0.54 (0.188)	0.66 (0.102)	1.21 (0.258)	1.20 (0.226)
	Random Forest	0.59(0.100)	46	58	25	38	20	21	0.51(0.152)	0.67(0.210)	1.30(0.415)	1.16(0.191)
	Support Vector Machine	NA	46	58	NA	NA	NA	NA	NA	NA	NA	NA

<sup>a</sup> N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate.

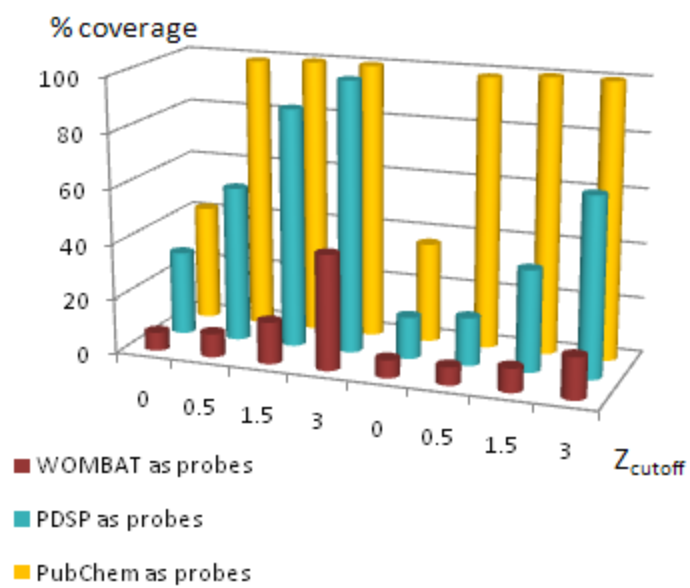
\* There is one compound in external set 1 that not within the applicability domain using  $Z_{\text{cutoff}} = 0.5$ .

### 3.4.3 Inter-dataset Cross Validation

#### 3.4.3.1 Applicability Domains of Models Generated for Various Datasets

To cross validate the 5HT1A models generated from three datasets, the applicability domains of the three datasets' were evaluated first. We applied the global similarity search approach, which involved all 672 Dragon descriptors, to search for structurally similar 5-HT1A binders in each of datasets using binders from one of the other datasets as probes. By applying different  $Z_{\text{cutoff}}$  values, the percentage of compounds that were identified within AD in each data source was shown in Figure 2. Regardless of dataset sources, the number of compounds within AD increased by raising the  $Z_{\text{cutoff}}$  values. When using 5-HT1A binders from WOMBAT for global similarity search, the least number of compounds were included within AD in both PDSP and PubChem dataset, indicating that the compounds from WOMBAT had very limited structural diversity, and thus the corresponding models had the smallest AD compared to other models. The chemicals from PDSP had greater structural diversity and the AD for the corresponding models was considerably wider. The compounds from PubChem had the most diverse structures, and nearly all compounds from either PDSP or WOMBAT were chosen using  $Z_{\text{cutoff}}$  of 0.5.

In order to perform the inter-dataset cross validation, both the model prediction accuracy and applicability domain should be taken into consideration. Models for PDSP and WOMBAT were both predictive, but models for PDSP had a much boarder AD than those for WOMBAT. Therefore, PDSP models were the most suitable to use for the validation of 5-HT1A binders from WOMBAT and PubChem.



**Figure 3.1.** The similarity search results of 5-HT1A binders for each dataset using binders from the other dataset as probes.

#### 3.4.3.2 WOMBAT Data Validation

We used models built for PDSP dataset (containing 105 5-HT1A binders and 61 non-binders) to verify the 69 5-HT1A binders from WOMBAT. The complete modeling set (i.e., including training and test sets) was used for the prediction as opposed to using only the corresponding training set. Among the 69 binders extracted from WOMBAT, all were within the applicability domain, and 65 were accurately annotated by *k*NN consensus prediction (CCR = 0.95, **Table 2**). The majority of ligands had been predicted correctly by *k*NN models, and for the only four falsely predicted 5-HT1A binders by *k*NN, they were within the applicability domain of only 70 models (i.e., approximately 30% of all models). As the model coverage was as low as 30% and the prediction scores were less than 1.70, these compounds' predictions were considered as low confidence. Similar consensus prediction results were achieved by both RF and SVM (**Table 2**), which confirmed that our models were both accurate and robust.

**Table 3.2.** The cross-validation of 69 WOMBAT 5-HT1A inhibitors by consensus prediction of acceptable QSAR models that were generated for PDSP dataset.

QSAR Methods	Prediction CCR	Confusion Matrix				Statistics
		N(1)	N(2)	TP	FN	SE
k-Nearest Neighbor	0.94	69	0	65	4	0.94
Random forest	0.94	69	0	65	4	0.94
Support Vector Machines	0.96	69	0	66	3	0.96

**Table 3.3.** The cross validation of 46 PubChem confirmatory 5-HT1A agonists/antagonists by consensus prediction of acceptable QSAR models that were generated for PDSP dataset.

QSAR Methods	Prediction CCR	Confusion Matrix				Statistics
		N(1)	N(2)	TP	FN	SE
k-Nearest Neighbor	0.48	46	0	22	24	0.48
Random forest	0.35	46	0	16	30	0.35
Support Vector Machines	0.56	46	0	26	20	0.56

### 3.4.3.3 PubChem Data Validation

All qualified models built for PDSP dataset were used to cross validate the 5-HT1A dataset from PubChem. In total of 21 5-HT1A agonists and 25 antagonists were found in PubChem dose response (confirmatory) assays (PubChem Assay ID: 613, 718 and 755). The consensus predictions were made by kNN, RF and SVM; however, 25 out of 46 were predicted non-binders by kNN, with 14 of those 25 having consensus prediction value higher than 1.70. These results strongly suggest that many of the agonists/antagonists reported in PubChem dose response (confirmatory) assays may be false positives. By applying random forest and support vector machines, the number of non-binders predicted was 30 and 20 respectively, out of the 46 total agonists/antagonists. 14 chemicals were predicted to be non-binders by all these three methods. 10 of them were commercially available, and were sent for further experimental validation at UNC.

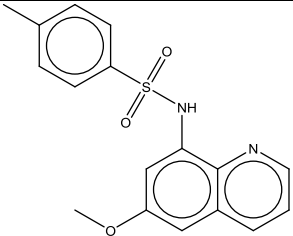
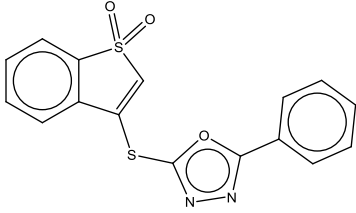
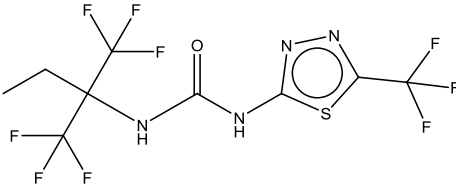
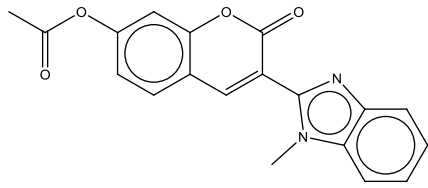
### 3.4.4 Experimental Validation

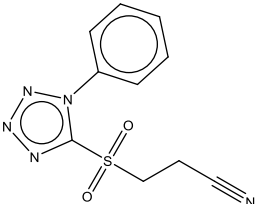
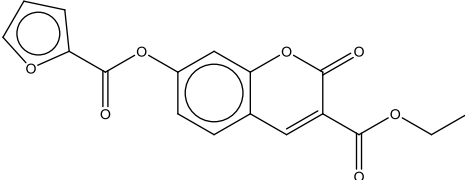
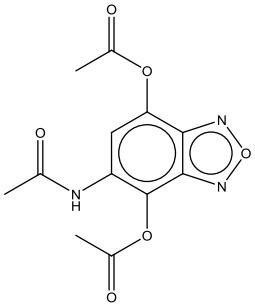
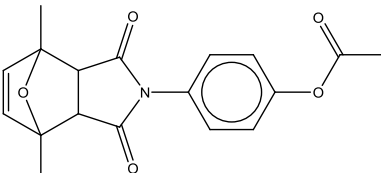
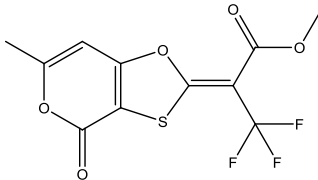
Of the 14 false positive PubChem chemicals predicted by all three methods, kNN, RF and SVM, there are 9 compounds commercially available, and were further tested experimentally in Dr. Bryan Roth's lab at UNC. All nine compounds (PubChem CID: 130606, 890649, 659822, 708260, 3242053, 733831, 597363, 2860584, 660939 Table 5) failed to achieve the 50% inhibitory activities at the single concentration assay. The most potent compound has the percentage of inhibition at 10  $\mu$ M being 43.0% while the weakest being only 10.6%. Although these experimental results confirm our QSAR models' prediction, the fact of such high false positive rate in PubChem data is still surprising to us because all the annotated agonists/antagonists are confirmed hits in dose-response cell-based assay conducted by the Scripps Research Institute Molecular Screening Center. In summary,



the above results did prove the high predictive power of our binary  $k$ NN, RF and SVM classification QSAR models built for the 5-HT1A dataset. These studies illustrate that the validated QSAR workflow, as employed in this paper, could be used as a general tool for identifying potential false positives and false negatives in assay results with low data quality.

**Table 3.4.** The experimental validation of 5-HT1A binding affinity test for PubChem confirmatory agonists/antagonists.

Structure	Serial No.	PubChem ID	PubChem label	kNN prediction score	RF prediction score	SVM prediction score	SEA prediction	Exp. Percent of inhibition <sup>a</sup>
	1	130606	Antagonist	1.97	1.76	2.00	Negative	12.10% <sup>b</sup>
	2	890649	Antagonist	1.86	1.76	2.00	Negative	10.60%
	3	659822	Antagonist	1.86	1.70	1.75	Negative	30.50%
	4	708260	Agonist	1.82	1.70	2.00	Negative	15.60%

	5	3242053	Antagonist	1.75	1.81	2.00	Negative	43.00%
	6	733831	Agonist	1.73	1.78	2.00	Negative	28.10%
	7	597363	Antagonist	1.73	1.77	2.00	Negative	20.30%
	8	2860584	Antagonist	1.73	1.66	2.00	Negative	20.10%
	9	660939	Antagonist	1.71	1.69	1.94	Negative	16.10%

- <sup>a</sup>: All nine commercially available PubChem compounds were sent to experimental validation for their 5-HT1A percent of inhibition test. Compounds with the percent of inhibition less than 50% are considered non-binders.
- <sup>b</sup>: The full IC<sub>50</sub> curves were generated in further experiment and were available in supplementary materials.

### 3.5 Conclusions

Our studies demonstrate that classification QSAR models can accurately differentiate true 5-HT1A binders from non-binders in reliable data sources, such as PDSP and WOMBAT. By contrast, we fail to generate qualified models using unreliable datasets, for example, data deposited in PubChem. The combinatorial QSAR modeling scheme is employed for all three 5-HT1A datasets and the models are rigorously validated using both internal (multiple training/test sets, Y-randomization test) as well as external validation (five-fold cross validation). We have demonstrated that this strategy affords QSAR models with high internal and external predictive power. As part of our QSAR modeling workflow, the predictors are further utilized for cross-validation of the 5-HT1A datasets from different sources. We find that the prediction results of our validated models highly agree with the experimental annotation of 69 5-HT1A binders as reported in WOMBAT database. Furthermore, the PubChem binders, which were identified by cross-validated QSAR models as false positive, were sent for experimental testing. The experimental results highly agreed with our models' consensus predictions and confirmed that those confirmatory assay hits deposited in PubChem are false positives.

## **CHAPTER 4**

### **DEVELOPMENT OF THE “DIVIDE-AND-CONQUER” QSAR MODELING SCHEME FOR RECA INHIBITORS AND VIRTUAL SCREENING FOR IDENTIFYING NOVEL CHEMICAL SCAFFOLDS**

#### **4.1 Introduction**

Good model building methodologies should take into consideration the different characteristics of modeling datasets, thus should vary case by case. Conventional QSAR approaches use various machine learning methods to build models and perform predictions, in regardless of caring about the high experimental variability of input data for model building. In this chapter, QSAR models would be generated to help identify those input data with low quality, and thus generate high quality models based on the curated dataset. Moreover, novel algorithms of the “divide-and-conquer” approach were developed based on existing QSAR classification methods to improve the performance of classifiers for largely imbalanced, sparsely structured datasets. These new algorithms outperformed traditional QSAR algorithms in mining the imbalanced RecA inhibitors dataset. Model-based virtual screening was performed thereafter, by mining 6.6 million molecules compiled from ZINC7.0 database, World Drug Index, and TimTec Diversity Set 10K. Proposed 11 computational hits were further experimentally tested against RecA ATPase activity assays,

and five were confirmed to be active against RecA. This new algorithm of QSAR approach demonstrated a unique edge in discovering chemical patterns for lead optimization.

#### **4.1.1 Introduction for RecA Protein Inhibitors**

The discovery of antibiotics, being one of the most important success stories in human medicine of the 20th century, has eased suffering and saved millions of lives in countless patients. However, antibiotic resistance in pathogenic bacteria is an escalating problem over time, making infections difficult if not impossible to treat. In the United States, infections encountered in the hospital or a health care facility affect more than 2 million patients, contributing to 88,000 deaths annually<sup>82</sup>. Notably, roughly 70% of those infections are resistant to at least one drug. Moreover, antibiotic drugs have accounted for \$23 billion in worldwide sales, making them the second largest therapeutic category in terms of sales. Therefore, the trend towards increasing numbers of infection, with the accelerating pace at which drug resistance increases, has provided huge research and business opportunities for novel antibiotic drug discovery.

New antibiotics discovered by traditional attempts, which mainly focus on finding new ways to combat bacteria, will be slow in coming to market due to the increasing rate that drugs develop resistance. Consequently, finding ways to combat antibiotic resistance directly is needed and might shed light on the discovery of novel antibiotic classes. Although the mechanisms by which antibiotic resistance evolves and spreads are not fully understood, the rapid rate at which bacteria develop drug resistance is largely due to mutations arising during mutagenic DNA synthesis and gene transfer between organisms. One protein central to both the development and transmission of antibiotic resistance is RecA, which is found in almost all bacteria and likely plays similar roles in all species. RecA is activated when the

bacterium is under stress and cannot divide in the usual way, thus acts as the foreman for a crew of repairman<sup>83</sup>. Therefore, it has been suggested that finding ways to inhibit RecA protein would open up a number of infectious therapeutic possibilities<sup>84</sup>. For example, using a RecA inhibitor in combination with mitomycin C, ciprofloxacin, or another DNA damaging antibiotic could provide a therapeutic strategy for treating NIAID class A and B pathogens. Furthermore, the discovery of RecA inhibitors might elucidate the molecular mechanisms related to the evolution of antibiotic drug resistance. Bacteria initiate the “SOS response”, which is initiated and controlled by RecA, when under attack by medicines and on the brink of destruction. Without RecA, the bacterial population’s ability to develop drug resistance would be suppressed.

Current discoveries of RecA inhibitors are encouraging but limited. The efforts of high-throughput screening encompass several classes of synthetic RecA inhibitors, including ATP-competitive small molecules, designed peptides, and select organometallic complexes<sup>85</sup>. However, these hits inhibit RecA protein with rather low potency. Thus there has been great need for potent RecA inhibitors of novel chemical scaffolds. The recent progress in model-based virtual screening has made the discovery of new chemicals more accurate and reliable; in addition, more diverse data in very large quantities have become available, allowing more opportunities for more focused hypothesis building. The paradigm is then beginning to change towards generating more focused structural hypotheses calling for much more limited testing of a smaller number of compounds with much higher probabilities of success; thus the experiment are being run to conform specific predictions rather than in a hope that some of the predicted hits may turn out useful. In this study, a unique modeling scheme that combines



unsupervised clustering with combinatorial QSAR methods was introduced for the discovery of novel scaffolds of RecA inhibitors.

#### **4.1.2 *RecA* Dataset**

##### **4.1.2.1 First Version of RecA Dataset**

The biological data for 3,488 compounds (53 inhibitors and 3,435 non-inhibitors) used in this study were generated in Dr. Singleton's laboratory. Compounds were tested by high-throughput screening against RecA's ATPase activity, and those with a percentage of inhibition at 17  $\mu$ M higher than 50% were subjected to the subsequent confirmatory binding assays<sup>85</sup>. Only limited number of primary hits were confirmed, therefore, additional RecA inhibitors were designed by Dr. Singleton according to those confirmed hits and were tested again, making the total number of RecA inhibitors to be 53. These confirmed RecA protein inhibitors have the inhibition  $IC_{50}$  range from 0.5 to 250. For the purpose of this work, we curated the data set to exclude duplicates within each group (inhibitors and non-inhibitors), and select the subset of organic compounds. Inorganic and organometallic compounds, as well as compound mixtures, were excluded because conventional chemical descriptors used in QSAR studies could not be computed in these cases. Also, 54 non-inhibitors were chosen based on a high structural similarity with the 53 inhibitors.

##### **4.1.2.2 Second Version of RecA Dataset**

The RecA dataset comprised of 26,433 compounds (145 inhibitors and 26,288 non-inhibitors) were generated after the curation of the first version of RecA dataset and the combination of additional up-to-date screening results from Boston University, IOC and Dr. Singleton's laboratory. Similar to the above curation procedure, duplicates were removed and

the main subsets of organic compounds were kept. Using global AD, 185 non-inhibitors were chosen based on high structural similarity with the 145 inhibitors.

#### **4.1.3 Libraries for Virtual Screening**

The virtual screening was performed on our in-house collection of ca. 6,600,000 molecules, including the ZINC database of ca. 6,500,000 compounds<sup>86</sup>, the World Drug Index (WDI) database of ca. 59,000 compounds<sup>87</sup>, and the TimTec Diversity library of 10,000 compounds. None of the compounds present in the modeling set were found in the screening libraries. Dragon 2D topological descriptors were calculated for each compound in the databases and linearly normalized based on the maximum and minimum values of each descriptor type in the modeling dataset of 145 RecA inhibitors and 185 non-inhibitors.

## **4.2 Methods**

### **4.2.1 Generation of Descriptors and Dataset Split**

#### **4.2.1.1 Generation of 2D Molecular Descriptors**

All chemical structures were cleaned using the wash function in MOE2007.09 to remove all but the largest fragment. Missing hydrogens were added during this procedure. The software of DRAGON2007-5.5 was used to calculate over 2,000 descriptors. Most of these descriptors can characterize chemical structures, including constitutional descriptors, walk and path counts, and functional group counts. In our study, 2D binary fingerprints and 2D frequency fingerprints were not included, and 698 chemically relevant descriptors remained after the removal of those descriptors with zero values or zero variance prior to model generation. The descriptor values were then linearly normalized to fall within a range

between zero and one based on the maximum and minimum values of each descriptor (i.e., range-scaled). Normalization was required to prevent unequal descriptor weighting during the QSAR model generation process as described below.

#### **4.2.1.2 Selection of External Validation Sets, Training and Test Sets**

In order to obtain external sets containing the same distribution of inhibitors and non-inhibitors as the original set, and also to increase the predictive statistical value, a five-fold Cross Validation (CV) analysis was performed for the whole dataset containing RecA inhibitors and non-inhibitors. The dataset was randomly split into five equal-size subsets of compounds and five independent sets of calculations were conducted, each of which involves only four-fifth of the compounds for model building and selection, and the remaining one-fifth of the compounds as an external test set.

The data set was subdivided into multiple training/test set pairs using the sphere exclusion method developed in our laboratory<sup>88, 89</sup>. By default, fifty different training/test set splits were initially tried using probe sphere radii defined by the minimum and maximum elements,  $D_{\min}$  and  $D_{\max}$ , of the distance matrix  $D$  between compound-vectors in the descriptor space and 42 splits were ultimately accepted. The number of compounds in the test set was varied to achieve the largest possible size of the test set, while ensuring that the training set models were still able to accurately predict the binding affinity of the test set compounds.

#### **4.2.2 Model-based Data Curation and Activity Cliffs Identification**

##### **4.2.2.1 Similarity Search and QSAR Models Generation**

For classification QSAR modeling, it would be ideal to have the balanced ratio between different compound classes in the modeling dataset. However, with the first version of RecA dataset containing 53 inhibitors and 3,435 non-inhibitors, the RecA dataset was imbalanced. In the absence of special statistical treatment, such a ratio would skew the prediction accuracy of the classification models. Therefore, a similarity search was performed against 3,435 RecA non-inhibitors using those 53 inhibitors as probes, and 55 compounds were chosen. All 698 Dragon descriptors were involved in this process, which was conducted by the software written in house.

We have followed the conventional *k*NN QSAR workflow for model building, and vigorously validated our models using both multiple internal test sets and five-fold external cross validation sets<sup>90</sup>.

#### 4.2.2.2 Identification of Compound Pairs with Activity Cliffs

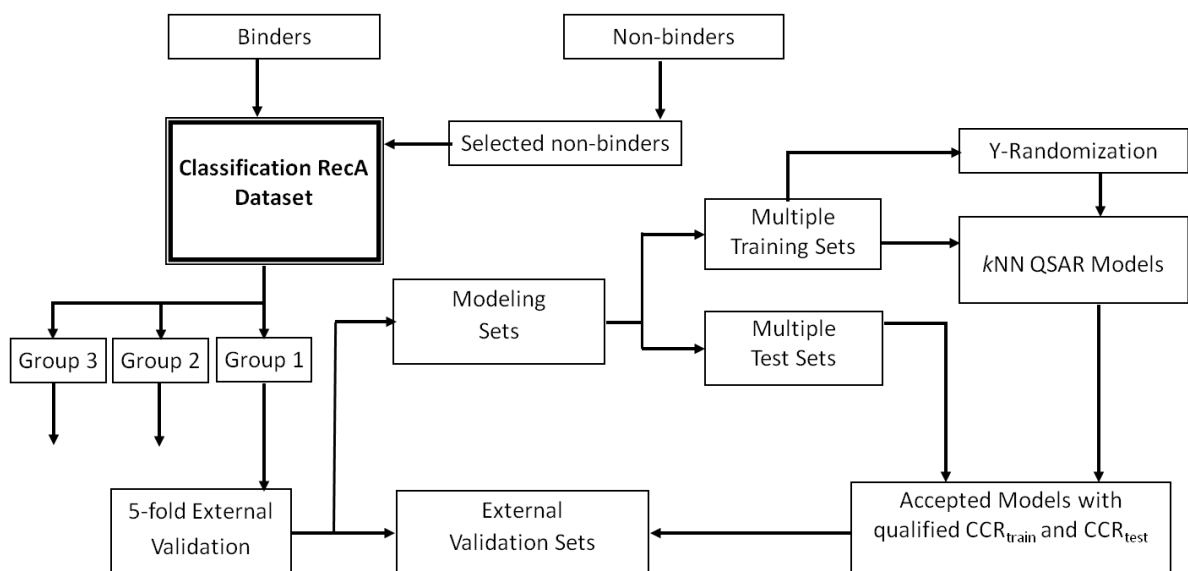
The first version of RecA dataset containing 53 inhibitors and 55 non-inhibitors were first clustered using hierarchical clustering method using R software. Then, we calculated the Tanimoto Coefficient (Tc) values for each pair of compound structures represented by 698 Dragon descriptors, and plotted the different values in heatmap. RecA compound pairs with activity cliffs were then identified. Several pairs with highly similar structures but opposite RecA inhibitory profiles were analyzed and further tested experimentally again.

#### 4.2.3 “Divide-and-Conquer” QSAR Modeling Scheme for Curated RecA Dataset

The second version of RecA dataset largely increased in size as well as structural diversity. Therefore, the dataset were first analyzed by the hierarchical cluster approach in the software of R. Three distinct groups of RecA inhibitors were identified according to their structural similarity between each other. A similarity search was then performed individually against RecA non-inhibitors using inhibitors in each group as probes, and three groups of RecA dataset were generated. The QSAR approach involving  $k$ NN, RF and SVM were conducted separately, followed by the rigously internal and external validation.

In addition, a Y-randomization test was carried out to establish model robustness. The test consists of rebuilding models using shuffled activities of the training set and evaluation of such models' predictive accuracy in comparison with the original model. It is expected that models obtained for the training set with randomized activities should have significantly lower values of statistical parameters such as  $CCR_{train}$  and, especially,  $CCR_{test}$ . Therefore, if most QSAR models generated in the Y-randomization test exhibit relatively high values of the statistical parameters for both training and test sets, it implies that a reliable QSAR model

cannot be obtained for the given dataset. This test was applied to all QSAR approaches in this study and was repeated twice for each division.



**Figure 4.1.** The workflow of the “Divide-and-Conquer” QSAR modeling approaches as applied to the RecA dataset.

#### **4.2.4 QSAR-based Virtual Screening**

As illustrated in the workflow shown above, the rigorously validated combinatorial QSAR models were employed for virtual screening. In total of 6.6 million compounds rendered the virtual screening process, including World Drug Index, ZINC7.0 database, and TimTec Diversity Set 10K. First, a global applicability domain was applied in the complete descriptor space in order to filter out compounds that differed in their structures from the modeling set compounds. All compounds filtered by the global AD were further put into the combinatorial QSAR approach for their RecA inhibitory predictions. Because robust QSAR classification models were successfully built for only compounds in group 1 and group 2, the global similarity search were applied separately using compounds in these two groups. We should point out that the chemical spaces represented by these two groups were different, so few compounds were chosen by these two similarity searches at the same time. During the consensus prediction, the results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction and the standard deviation of estimated means across all models was small. Furthermore, we restricted ourselves to the most conservative applicability domain for each model using  $Z_{\text{cutoff}} = 0.5$ .

#### **4.2.5 Experimental Validation of Virtual Screening Hits**

For all virtual hits chosen by consensus predictions of  $k$ NN, RF and SVM, 11 chemicals were further selected including one compound from World Drug Index, 1 from ZINC7.0 database, and nine from TimTec Diversity Set 10K. These structurally diverse and commercially available hits were purchased from different suppliers and experimentally



tested in Dr. Scott Singleton's laboratory for measuring their RecA ATPase inhibition activity.

### 4.3 Results and Discussions

#### 4.3.1 QSAR Classification Models On First Version Dataset

The first version of RecA dataset contains 53 RecA inhibitors and over 3,435 non-inhibitors. Descriptors were first generated for all compounds using DRAGON software 2007. The group of non-inhibitors, representing diverse structural classes, have only limited numbers of compounds sharing the same structural classes as active ones, while most others have structures highly dissimilar from those included in the group of RecA inhibitors. Therefore, in order to prevent the model building and validation from being biased toward correct prediction of the larger group, the RecA non-inhibitors, a similarity search was performed using those 53 RecA inhibitors as probes. There were 55 non-inhibitors chosen by their structural similarity with those inhibitors using Z cutoff of -0.4. One-fifth of the whole dataset were randomly split as the external validation set, and the rest four-fifth were used for model building and selection. The sphere exclusion method, which developed in our laboratory, was used to divide each modeling set into multiple training/test set pairs. Hundreds of  $k$ NN models were generated by using a simulated annealing variable selection procedure. All the models were evaluated by Correct Classification Rate (CCR) (equation 3) and were selected based on the criteria that both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  were equal to or higher than 0.80. In this case, there were 497 qualified models generated. However, when these qualified models were used to predict those external set compounds, the performance of  $CCR_{\text{external}}$  for all external compounds reached only 0.48. There were nine RecA inhibitors in

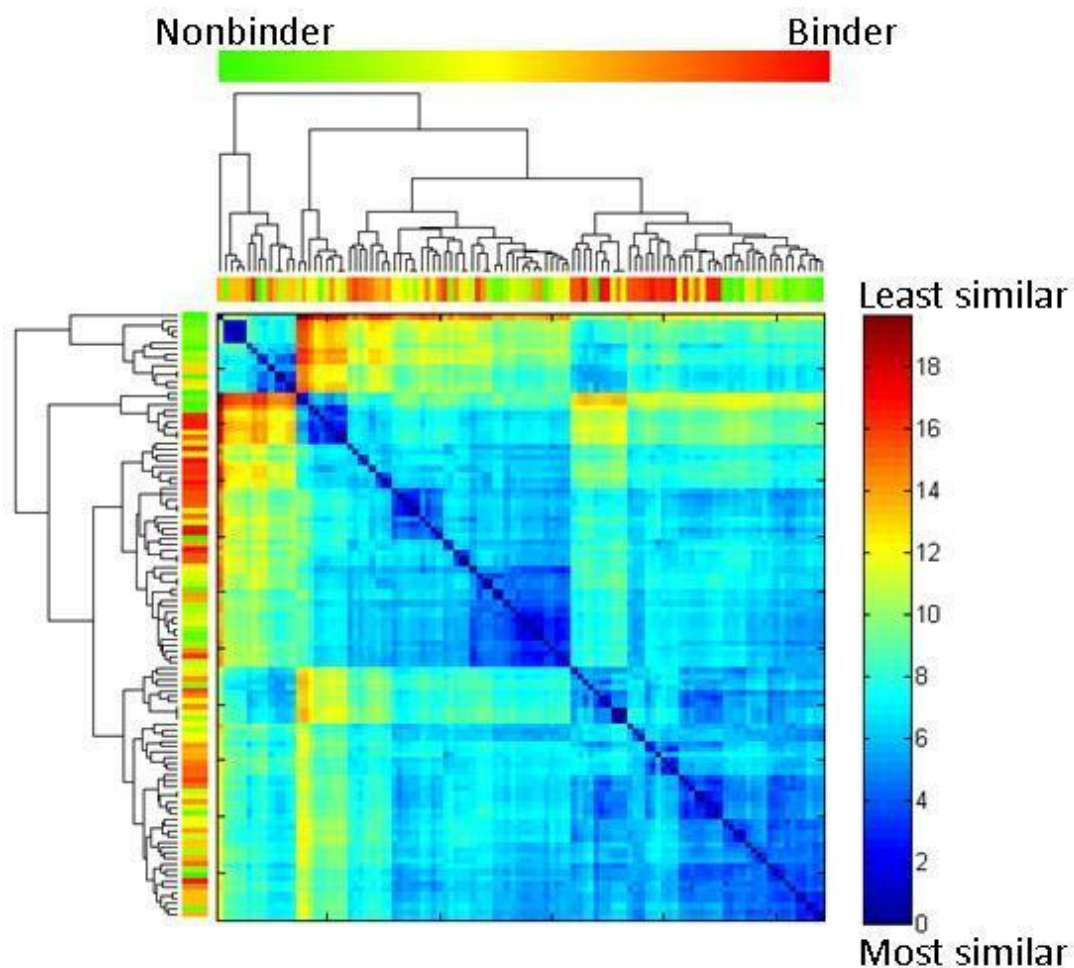
total in the external validation set, and six of which were wrongly predicted, representing the prediction statistics of the selectivity value being as low as 0.33.

To further explore the reason why model building achieved with only limited success, those 6 RecA inhibitors were put into the modeling set, with the external validation set substituted with another six RecA inhibitors randomly chosen from the original modeling set compounds. Then, *k*NN models were generated following the same above procedure including Sphere Exclusion protocol and simulated annealing variable selection; however, only 22 models qualified the threshold having both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  equal to or higher than 0.80. Moreover, the statistics of the external set prediction,  $CCR_{\text{external}}$ , by these consensus models were greatly improved, achieving 0.84 (compared to 0.48 previously). Herein, we observed that the model statistics declined dramatically for the dataset when those 6 RecA inhibitors presented: qualified model numbers drops considerably when those 6 compounds were in the modeling set, while  $CCR_{\text{external}}$  was very low when they were in the external validation set. Therefore, it was highly suspected that there were some mislabeled compounds or activity cliff pairs in this version of RecA dataset, which became the apparent obstacle for the building of successful *k*NN-QSAR models..

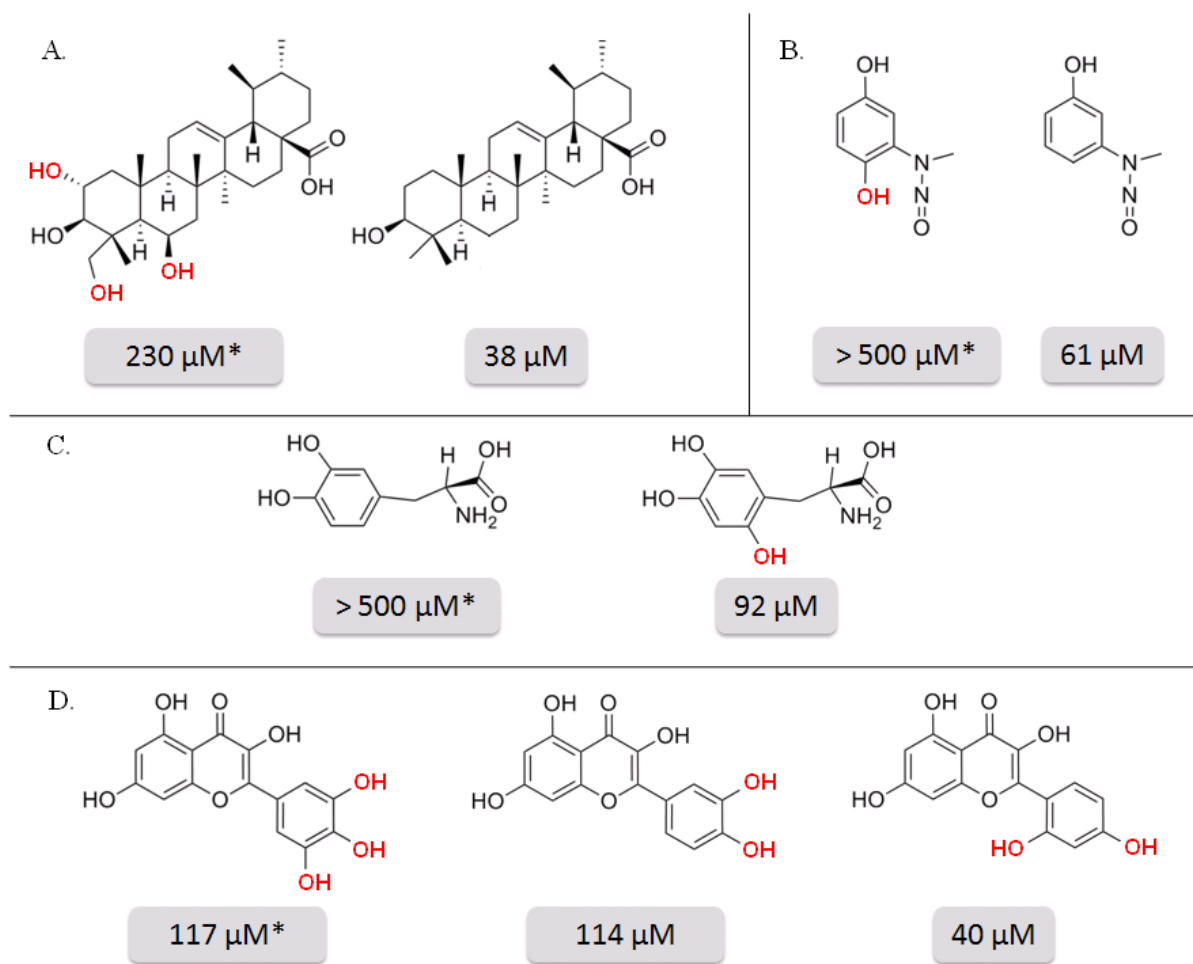
#### ***4.3.2 Identifications of Compound Pairs with Activity Cliffs and Data Curation***

Tanimoto Coefficient (Tc) values for each pair of compound structures in the first version of RecA dataset (53 inhibitors and 55 non-inhibitors) was calculated, followed by the conduction of pair-wise comparison. RecA inhibitors and non-inhibitors were first clustered using a hierarchical clustering method, and the tree structure was shown on the above and left side of the heatmap in **Figure 4.2**. The color bars below/beside the hierarchical cluster tree represented the compound class each structure belonged to: red as RecA inhibitor and green

as RecA non-inhibitor. Though most inhibitors were clustered together with inhibitors, and most non-inhibitors with were clustered with non-inhibitors, it could still be clearly seen that some inhibitors were positioned in close enough proximity with non-inhibitors in the tree, which suggested the existence of activity cliffs in the dataset. The center of the heatmap, **Figure 4.2**, showed the pair-wise comparison scores between each structures in the dataset, which further confirmed this hypothesis. The right side color bar, from red to yellow to green to blue, showed the increasing degrees of structural similarity between compound pairs: the more red of each small square, the less similar the compound pairs were, while the more blue of it, the more similar the compounds were. While it was commonly known that the diagonal of the heatmap should be dark blue: each compound is identical to itself and thus the bluest, several other dark blue matrices were noticed in the heatmap. The identification of those dark blue matrices unveiled the activity cliffs of several structurally highly similar compound pairs (**Figure 4.3**) as well as some undiscovered duplicates.



**Figure 4.2. Heatmap of pair-wise Tc analysis for the first version RecA dataset.** RecA inhibitors (53) and non-inhibitors (55) were first clustered using hierarchical clustering method, as shown on the above and left side of the heatmap. The color below the hierarchical clustering represents whether the individual compound belongs to the category of either inhibitor (red) or non-inhibitor (green).



\*previously labeled as non-inhibitor.

**Figure 4.3. Activity cliffs of RecA inhibitors and non-inhibitors.** These four pairs of activity cliff compounds were identified by binary classification QSAR models, and were further tested again experimentally. Two pairs (B and C) were verified to be true activity cliffs. The previously identified non-inhibitors in pairs A and D were found to be weak RecA protein inhibitors, with the  $\text{IC}_{50}$  being 230  $\mu\text{M}$  and 117  $\mu\text{M}$ .

As shown in **Figure 4.3**, the backbones of chemical structures within each activity cliff group were identical, and the only difference is either absence/presence of a phenolic hydroxy or different location of them. The first compound within each group (with asterisk) were first labeled as RecA non-inhibitors; however, such small differences usually would not render large difference for RecA binding activity. Therefore, we suggested these 4 compounds to be re-tested in our collaborator's lab using the same protocol RecA inhibitors were originally identified. The experimental results confirmed two compounds to be true activity cliffs with their pair ones (B and C in **Figure 4.3**), and the other two (A and D in **Figure 4.3**) to be mislabeled false negatives with low RecA binding affinity ( $> 100 \mu\text{M}$ ). These promising results showed that *k*NN-QSAR models were not only powerful enough to identify possible mislabeled compounds, but could also be used to help detect true activity cliff compound pairs.

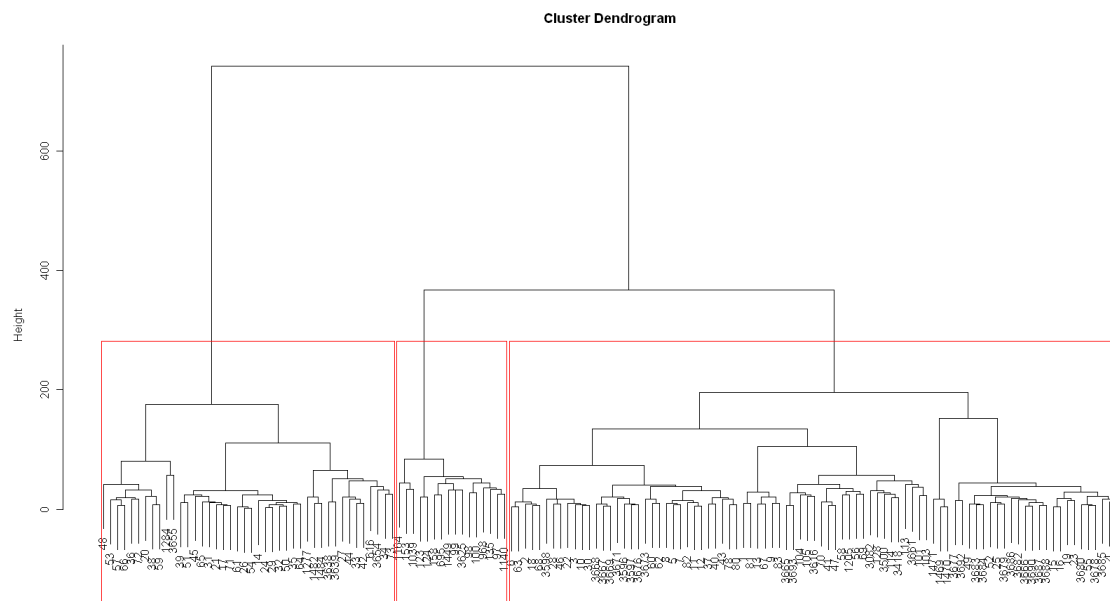
Our previous experience indicated that the presence of activity cliff compound pairs would dramatically decrease the statistics for *k*NN QSAR model generation as well as consensus prediction, therefore, for each pair compounds, we only keep the RecA inhibitors, and remove the other RecA non-inhibitors out of the whole dataset. Moreover, for the duplicate compound pairs present in the group of both RecA inhibitors and non-inhibitors, those labeled as non-inhibitors were removed. Therefore, there were nine compound in RecA non-inhibitors group were removed in total.

Furthermore, results of additional screened compounds from Boston University, IOC and our collaborator's lab were added in the first version dataset. After removing duplicate chemicals, the second version of RecA dataset were generated, which encompassed 145 confirmed RecA inhibitors and 26,288 non-inhibitors.

#### **4.3.3 QSAR Classification Models On Dataset After Curation**

The RecA dataset size increased a lot after incorporating newly identified RecA inhibitors and non-inhibitors from other multiple sources. The RecA inhibitors themselves were highly diversified in chemical space, thus presenting a great challenge for QSAR modeling. To address it, the hierarchical cluster approach was firstly used to analyse the RecA inhibitors dataset, with the results shown in Figure 4. It was clearly seen that the cluster dendrogram could be cut into three sub-trees easily, representing three distinct groups of RecA inhibitors according to their structural similarity among each other. The largest group (group 1) contained 87 RecA inhibitors; the mid-sized group (group 2) had 42; and the smallest group (group 3) had only 16 compounds.

Using the hierarchical clustering, a method of unsupervised learning, before the QSAR study of supervised learning shed light on building models on these highly diversified dataset. Then, similarity search was performed against RecA non-inhibitors using inhibitors in each cluster as probes. There were 113 non-inhibitors chosen by their structural similarity with those inhibitors in group 1 using  $Z_{\text{cutoff}}$  of 1, making the entire dataset of 200 RecA inhibitors and non-inhibitors for  $k$ NN-QSAR modeling. Similarly, 55 non-inhibitors were chosen for 42 RecA inhibitors in group 2, and 19 compounds were chosen for 16 RecA inhibitors in group 3. Among these non-inhibitors, only one compound was chosen by both group 1 and group 2 probes, representing the non-overlapping chemical structures each group of compounds represented.

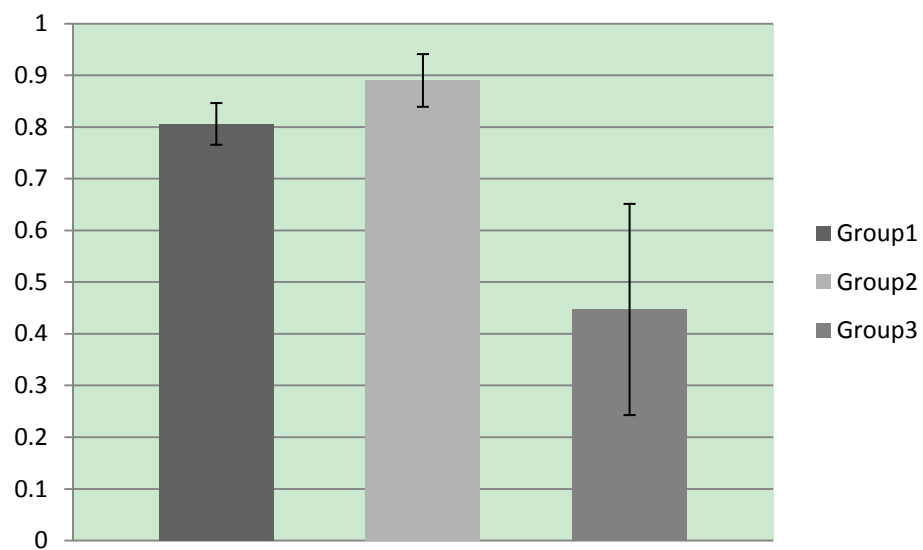


**Figure 4.4. Hierarchical clustering of 145 RecA inhibitors.** The descriptors for all 145 RecA inhibitors were calculated by Dragon software, and the hierarchical clustering was performed by R. According to the structural similarity of compounds between each other, three groups of compounds were clustered. The largest group (group 1) has 87 RecA inhibitors; the mid-sized group (group 2) has 42; and the smallest group (group 3) has only 16 compounds.



The five-fold CV approach was employed in each of the three RecA dataset split, so that only four-fifth of the dataset compounds were used for model building and selection. These modeling sets were then rendered the split of multiple training/test set pairs using the Sphere Exclusion technique, and multiple models were generated with simulated annealing variable selection approach. All model building statistics were available in Table 1. Predictive *k*NN-QSAR models were successfully built in both group 1 and 2. For group 1, there were 5,640 models passed the criteria that both  $CCR_{train}$  and  $CCR_{test}$  equal to or higher than 0.80. The number of models that qualified the same threshold for group 2 dataset exceeded 30,000, so we had to increase the cutoff value for both  $CCR_{train}$  and  $CCR_{test}$  to 0.9, and keep only manageable number of models for consensus prediction. Even for such a high criteria, there were still 20,863 models qualified. However, as for model statistics in group 3, less than 20 models were chosen given that both  $CCR_{train}$  and  $CCR_{test}$  were equal to or higher than 0.80. There were only 339 models qualified after the cutoff values were both decreased to 0.7, as shown in Table 1. These results from modeling sets suggested that the data quality in both group 1 and 2 were good enough to build statistical powerful models, while group 3 data were not, only yielding many poor models.

Moreover, models that passed the qualified threshold for both  $CCR_{train}$  and  $CCR_{test}$  were applied to predict RecA inhibitors and non-inhibitors in the external validation sets. Consistent with the above hypothesis, models built from group 1 and group 2 were able to classify correctly RecA inhibitors from non-inhibitors in the external sets, achieving at least 0.75 for  $CCR_{external}$  of group 1 and 0.80 for  $CCR_{external}$  for group 2; however, the prediction statistics for the external set compounds from group 3 were no better than random guesses, with the lowest  $CCR_{external}$  being 0.13.



**Figure 4.5.** Five-fold external set prediction results by *k*NN-QSAR for RecA dataset after cluster.

**Table 4.1.** Results for the five-fold external sets cross validation as well as the secondary external set (from WOMBAT) validation using three different machine learning methods.

Group of RecA inhibitors	External Sets	Number of Models	Prediction CCR	Confusion Matrix						Statistics			
				N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)
Group 1	1	920 <sup>b</sup>	0.84	17	23	13	21	2	4	0.76	0.91	1.80	1.59
	2	945	0.80	17	23	13	19	4	4	0.76	0.83	1.63	1.56
	3	622	0.85	17	23	15	19	4	2	0.88	0.83	1.67	1.75
	4	1629	0.79	18	22	13	19	3	5	0.72	0.86	1.68	1.51
	5	1524	0.75	18	22	13	17	3	5	0.72	0.77	1.68	1.47
Group 2	1	4377 <sup>c</sup>	0.89	7	14	6	13	1	1	0.86	0.93	1.85	1.73
	2	3630	0.85	10	9	7	9	3	0	0.70	1.0	1.35	2.0
	3	4999	0.83	8	11	6	10	2	1	0.75	0.91	1.61	1.76
	4	3537	0.95	10	9	9	9	1	0	0.90	1.0	1.78	2.0
	5	4320	0.93	7	12	6	12	1	0	0.76	1.0	1.82	2.0
Group 3	1	13 <sup>d</sup>	0.65	2	5	1	4	1	1	0.50	0.80	1.43	1.23
	2	82	0.13	3	4	0	1	3	3	0.0	0.25	0.0	0.40
	3	48	0.59	3	4	2	2	1	2	0.67	0.50	1.45	0.86
	4	152	0.41	4	3	2	1	2	2	0.50	0.33	0.86	0.80
	5	44	0.46	4	3	1	2	3	1	0.25	0.67	0.40	1.45

<sup>a</sup>: N(1) = number of inhibitors, N(2) = number of non-inhibitors, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-inhibitors predicted as inhibitors), FN = false negatives (inhibitors predicted as non-inhibitors), TN = true negative (non-inhibitors predicted as non-inhibitors), SE = sensitivity =  $TP/N(1)$ , SP = specificity =  $TN/N(2)$ , EN - the normalized enrichment,  $EN(1) = (2TP * N(2))/(TP * N(2) + FP * N(1))$ ,  $EN(2) = (2TN * N(1))/(TN * N(1) + FN * N(2))$ , and CCR = correct classification rate.

<sup>b</sup>: Models with both  $CCR_{train}$  and  $CCR_{test}$  over 0.8.

<sup>c</sup>: Models with both  $CCR_{train}$  and  $CCR_{test}$  over 0.9.

<sup>d</sup>: Models with both  $CCR_{train}$  and  $CCR_{test}$  over 0.7.

Furthermore, the Y-randomization test for each group further confirmed our hypothesis. For both group 1 and group 2, models obtained for the training sets with randomized activities had significantly lower values of statistical parameters, such as  $CCR_{train}$  and  $CCR_{test}$ , than the ones with original activity labels. However, for group 3, most QSAR models generated in the Y-randomization test exhibit relatively comparable statistics to the models generated without randomization, which implied that the QSAR models obtained in group 3 dataset were not robust enough.

#### **4.3.4 Virtual Screening To Identify Putative RecA Inhibitors**

Virtual screening was conducted using the consensus approach, which relies on averaging predictions from all qualified models, i.e. all models with both  $CCR_{train}$  and  $CCR_{test}$  equal to or greater than 0.80 instead of using only one single and best model. Since the number of total models qualified was still large, herein, for each cluster of RecA dataset, we only used the group of models yielding the highest  $CCR_{external}$  for the virtual screening. An important condition that ensures reliable predictions by the model is the use of AD. Therefore, two types of AD were employed in the virtual screening of compound databases, which includes a global AD and local AD. The global AD, which acts as a filter, ensures some level of global similarity between the predicted compounds and the compounds in the modeling set, while the local AD is defined for each of the individual classification models.

A large external database including around 6,500,000 compounds from the ZINC 7.0 database, around 59,000 from World Drug Index (WDI) dataset, and 10,000 from TimTec Diversity library were screened for putative RecA inhibitors. These original collections had many duplicates, such as many salt forms for the same chemical entity, therefore, all molecules were firstly “washed” using the Wash Molecules tool in MOE, keeping only the

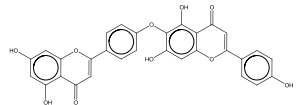
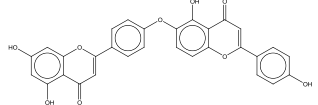
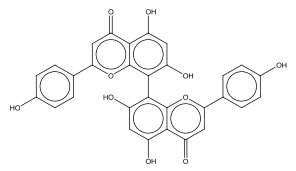
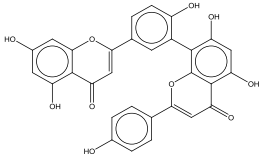
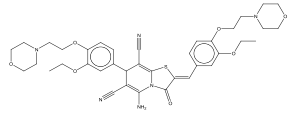
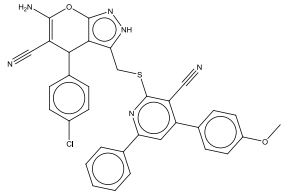
main organic chemical entities. Duplicates were then identified by XChem and were removed using MOE. We also removed all compounds included in our modeling and external validation sets. Dragon descriptors were generated for the remaining unique compounds in the database, which were then subjected to a global AD filter for the modeling set from RecA group 1 and group 2 compounds respectively. The  $Z_{\text{cutoff}}$  of 0.5 was applied because we considered the trade off for both the number of compounds chosen for consensus prediction and the confidence level we had for our model prediction accuracy. Next, all kNN-QSAR models with  $\text{CCR}_{\text{train}}$  and  $\text{CCR}_{\text{test}} \geq 0.80$  generated from group 1 RecA dataset were employed in consensus fashion to predict those compounds filtered by global AD from group 1 modeling set; while consensus models generated from group 2 dataset were used to predict compounds filtered by group 2 global AD. This resulted in a selection of 1,470 active hits for group 1 and 1,662 hits for group 2. kNN-QSAR consensus models were then used as final filters for the determination of putative RecA inhibitors. To obtain the higher confidence level for each prediction, we took both the consensus score (average class number) and model coverage into consideration. In particular, only the hits with average class number between 1.0 and 1.25 and the model coverage over 50% were selected. We found that there were 12 compounds from ZINC database, 11 from WDI library, and 10 from TimTec Diversity library that satisfied both criteria.

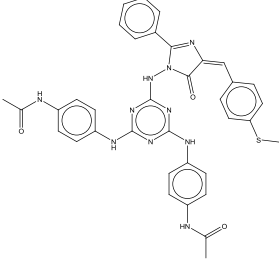
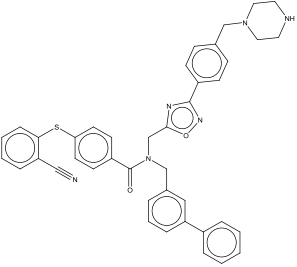
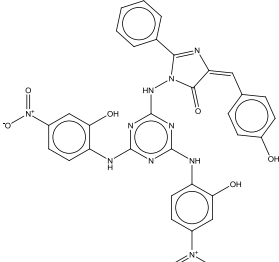
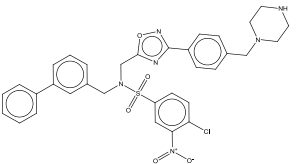
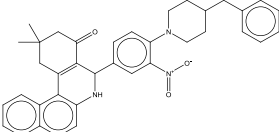
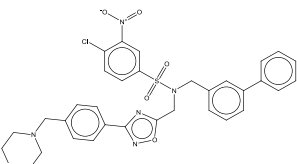
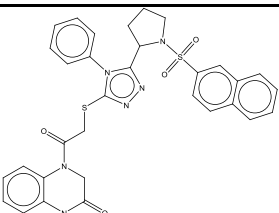
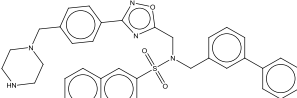
#### **4.3.5 Experimental Validation**

For all computational hits identified by consensus predictions, 11 compounds were chosen to be tested experimentally against RecA APTase inhibitory assays by considering both commercial availability and their prices. These test chemicals included Cupressuflavone from ZINC 7.0 database, Hinokiflavone from the World Drug Index, and nine chemicals (ID:

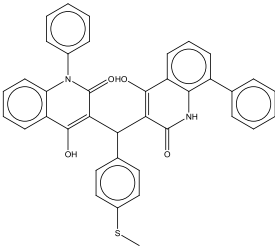
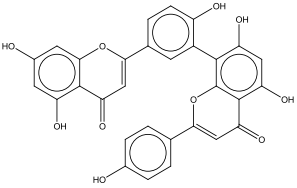
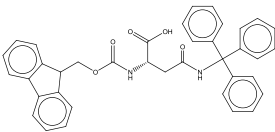
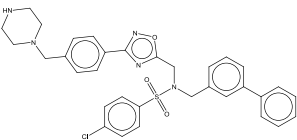
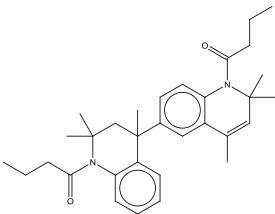
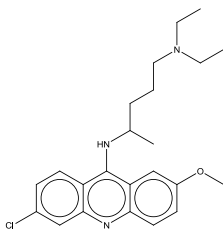
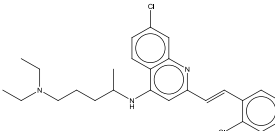
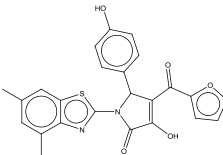
ST025018, ST026920, ST026922, ST044784, ST059022, ST069932, ST029446, ST002554 and ST072510) from the TimTec Diversity Set 10K. Five of them showed the ATPase inhibitory activities against RecA protein, in the range of 5~28  $\mu$ M. For each tested compounds, the full dose-response curve was obtained and the inhibition  $IC_{50}$  was calculated. We should emphasize that in our QSAR modeling approaches, only binary endpoints were used for generating models and virtual screening, so no estimation of exact binding affinities ( $K_i$  values) had been made. Consequently, these confirmed experimental results were considered very promising.

**Table 4.2.** The experimental test for the five computational hits of 5-HT<sub>1A</sub> inhibitors by mining the TimTec GPCR targeted screening library.

Structure	No.	Name	Source library	Most similar compound in modeling set	Tc	kNN-QSAR prediction score	Exp. IC <sub>50</sub> (μM)
	1	Hinokiflavone	World Drug Index		1.0	1.02	10
	2	Cupressuflavone	ZINC		1.0	1.14	Non-inhibitor
	3	ST025018	TimTec Diversity library		0.65	1.12	Non-inhibitor

	4	ST026920	TimTec Diversity library		0.42	1.07	8.3
	5	ST026922	TimTec Diversity library		0.38	1.07	12
	6	ST044784	TimTec Diversity library		0.53	1.23	Non-inhibitor
	7	ST059022	TimTec Diversity library		0.41	1.13	Non-inhibitor



	8	ST069932	TimTec Diversity library		0.33	1.22	Non-inhibitor
	9	ST029446	TimTec Diversity library		0.76	1.22	28
	10	ST002554	TimTec Diversity library		0.31	1.21	Non-inhibitor
	11	ST072510	TimTec Diversity library		0.31	1.25	5.3

## 4.4 Conclusions

Antibiotic resistance is an escalating problem requiring the discovery of novel antibiotic classes acting on nonclassical cellular targets. Targeting the nonessential genes, for example RecA, offers possible attractive solution. In this study, we have developed combinatorial Quantitative Structure-Activity Relationship (QSAR) model for hundreds of chemically diverse RecA inhibitors and their structurally similar inactive compounds resulting from high-throughput screening and the subsequent confirmatory binding assays. The initial attempts to classify 53 RecA inhibitors out of over 3,000 non-inhibitors met with only limited success, due to the fact that the activity cliffs exist in several highly similar compound pairs identified by pair-wise Tanimoto Coefficient (Tc) analysis. Then, a new dataset, containing 145 RecA inhibitors and 26,132 non-inhibitors, was created after both data curation of those activity cliff pairs and incorporation of more up-to-date experimental testing results. The new dataset was clustered into three groups according to structure similarity, then the variable selection *k*-Nearest Neighbor (*k*NN), Random Forest (RF) and Support Vector Machines (SVM), were employed for model building within each group using 2D topological Dragon chemical descriptors. Highly predictive QSAR models were generated with leave-one-out cross-validated (LOO-CV) Correct Correlation Rate (CCR) and the external CCR values were as high as 0.85, which is greatly improved compared to the CCR of 0.79 for model building without clustering. With two differently defined applicability domain thresholds, all validated QSAR models were employed concurrently for virtual screening (VS) of an in-house compound collection including 9.5 million molecules compiled from the ZINC7.0 database, the Word Drug Index (WDI) database, and the TimTec Diversity Set. VS resulted in 31 structurally unique consensus hits that were considered novel

putative RecA inhibitors. These computational hits had several novel structural features that were not present in the original data set. There were 11 computational hits, some of them possessing novel scaffolds, that were tested experimentally and 5 out of 11 were confirmed to be active against RecA, with  $IC_{50}$  values ranged 5 ~ 28  $\mu$ M. In summary, this study illustrates the power of the combinatorial QSAR-VS method as a general approach for the effective identification of structurally novel bioactive compounds.

## CHAPTER 5

### DEVELOPMENT OF COMBINATORIAL QSAR MODELS FOR 5-HYDROXYTRYPTAMINE 1A RECEPTOR AND VIRTUAL SCREENING OF LIBRARIES WITH DIFFERENT CHARACTERISTICS

#### 5.1 Introduction

The 5-Hydroxy Tryptamine receptor subtype 1A (5-HT1A) is highly expressed in the *raphe* nuclei region and limbic structures; for that reason 5-HT1A has been an attractive target to treat mood disorders such as anxiety and depression. We have developed binary combinatorial QSAR models for 5-HT1A binding using data retrieved from the PDSP Ki database by employing *k* Nearest Neighbor (*k*NN), Random Forest (RF) and Support Vector Machine (SVM) prediction methods. We have employed a rigorous model development workflow, including extensive internal and external validation. The classification accuracies of the models to discriminate 5-HT1A binders from the non-binders were as high as 86% for the external test set. These models were used to mine chemical libraries with different characteristics, including drug-like libraries from the World Drug Index and Prestwick, GPCR-targeted libraries from TimTec and ASINEX, and diversity libraries from TimTec and ASINEX. 15 computational hits were tested in radioligand binding assays with a success rate of 60%, and one compound was found to be very potent, having a binding affinity of 2.3 nM with 5-HT1A.

### **5.1.1 Introduction for the 5- Hydroxytryptamine (serotonin) receptor 1A**

The 5-HT<sub>1A</sub> receptor is a subtype of 5-HT receptor that binds the endogenous neurotransmitter serotonin (5-hydroxytryptamine, 5-HT). It is the most widespread of all the 5-HT receptors, an important family of G protein-coupled receptors (GPCRs). In the central nervous system, 5-HT<sub>1A</sub> receptors exist in the cerebral cortex, hippocampus, septum, amygdala, and raphe nucleus in high densities, while low amounts also exist in the basal ganglia and thalamus. It has been among the most important molecular targets that are actively being explored for potential drug discovery efforts in psychoactive treatment. Because of its dense concentration on cortical and hippocampal pyramidal neurons, 5-HT<sub>1A</sub> receptors have been actively studied in recent years for novel strategies for treating the cognitive deficits in schizophrenia. In fact, atypical antipsychotic drugs modestly enhance cognition, and several atypical antipsychotic drugs have 5-HT<sub>1A</sub> partial agonist activity (eg, aripiprazole, clozapine, olanzapine, ziprasidone, quetiapine). In addition, 5-HT<sub>1A</sub> receptor agonists, such as buspirone and flesinoxan, show efficacy in relieving anxiety and depression, and buspirone and tandospirone are currently approved for these indications in various parts of the world. Others, such as gepirone, flesinoxan, flibanserin, and PRX-00023, have also been investigated, though none has been fully developed and approved as of yet. Some of the atypical antipsychotics, like aripiprazole, are also partial agonists at the 5-HT<sub>1A</sub> receptor and are sometimes used in low doses as augmentations or standard antidepressants, for example, the selective serotonin reuptake inhibitors (SSRIs).

5-HT<sub>1A</sub> receptors have recently received considerable attention as treatments for neurodegenerative diseases. 5-HT<sub>1A</sub> receptor activation has been shown to increase dopamine release in the medial prefrontal cortex, striatum, and hippocampus, and may be

useful for improving the symptoms of Parkinson's disease. As mentioned above, some of the atypical antipsychotics are 5-HT<sub>1A</sub> receptor partial agonists, and this property has been shown to enhance their clinical efficacy. Enhancement of dopamine release in these areas may also play a major role in the antidepressant and anxiolytic effects seen upon postsynaptic activation of the 5-HT<sub>1A</sub> receptor. Moreover, 5-HT<sub>1A</sub> receptor antagonists such as lecozotan have been shown to facilitate certain types of learning and memory in rodents by stimulating the release of glutamate and acetylcholine in various areas of the brain. As a result, they are being developed as novel treatments for Alzheimer's disease. Taken together, there is a critical need in developing novel 5-HT<sub>1A</sub> receptor modulators to benefit the aforementioned diseases.

### ***5.1.2 Introduction of the Dataset for QSAR Model Building***

The 5-HT<sub>1A</sub> binders and non-binders were downloaded from the NIMH Psychoactive Drug Screening Program (PDSP). By querying in PDSP, 105 unique compounds were identified to be 5-HT<sub>1A</sub> binders. 78 nonbinders were also extracted, which were shown to have no binding to the 5-HT<sub>1A</sub> receptor at 1mM concentration. Most of these non-binders shared a relatively high structural similarity with those 105 binders.

### ***5.1.3 Introduction of the Libraries for Virtual Screening***

#### **5.1.3.1 Drug-like Screening Libraries.**

Drug-like databases are collections of currently marketed drugs or drug candidates in the approval process. For our study, we used the World Drug Index (WDI) database as well as the Prestwick Chemical Library (PCL).

WDI is maintained by Derwent Publications and contains 59,000 drugs and pharmacologically active compounds, including all marketed drugs. WDI993 also contains 175,000 synonyms, 73,000 trade names, 26,000 manufacturers, 6700 International Non-proprietary Names, 8000 US Adopted Names, 17,000 journal and conference references, and more, including extensive medical data, such as indications and usage, interactions, adverse effects, mechanism of action, and activity keywords.

PCL is a collection of the Prestwick chemical company. It contains 1,200 small molecules with 100% being marketed drugs, thus it represents the greatest possible degree of drug-likeness. The active compounds were selected for their high chemical and pharmacological diversity as well as for their known bioavailability and safety in humans.

#### **5.1.3.2 Targeted Screening Libraries.**

5-HT1A belongs to the big family of GPCR, therefore, GPCR-targeted libraries were virtually screened for the purpose of identifying new 5-HT1A ligands. In the study, the TimTec AntiTarg-G library and ASINEX Synergy GPCR CNS library were chosen. The TimTec AntiTarg-G library is a plated screening set of molecules that contain chemical lattices present in compounds reported in the technical or patent literature to possess GPCR-ligand properties. A pre-filtered diversity collection of 2,300 compounds is assembled which provides a high-value screening library of molecules for identifying the new GPCR ligands. Moreover, Structural constraints and novel pendants within these lattices provide the structural variability to identify new chemical directions for hit optimization.

Similarly, the ASINEX Synergy GPCR CNS library is a collection of the ASINEX Company and is composed of 3,233 compounds rich in GPCR drug-like pharmacophore fragments.

#### **5.1.3.3 Diversity Screening Libraries.**

The diversity libraries we decided to use are also from TimTec and ASINEX, named TimTec Diversity Set 10K and ASINEX Diverse Set-Platinum 5K. The diversity screening set from TimTec contains 10,000 samples selected from the company's stock of over 180,000 compounds as the most structurally diverse and competitively priced collection. The assorted set stands out as having a diverse selection of singletons identified in the TimTec stock pool of readily available compounds. In addition, it is also a compound collection that complies with Lipinski Rules of Five.

ASINEX Diversity Set-Platinum 5K, which contains 5,072 compounds, is an assortment of all other ASINEX libraries based on the compounds' structural multiplicity. The ASINEX Company claimed it to be a great starting point that requires a pure diversity of chemicals.

## **5.2 Methods**

### **5.2.1 Dataset Curation**

For the purposes of this work, the data was curated following the guidelines our laboratory suggested earlier<sup>91</sup>. First, all molecules were cleaned using the Wash Molecules module in MOE<sup>8</sup> (v.2009.10). This software processes chemical structures by



carrying out several standard operations including 2D depiction layout, hydrogen correction, salt and solvent removal, chirality and bond type normalization (all details can be found in the MOE manual<sup>8</sup>). Second, ChemAxon Standardizer<sup>92</sup> was used to harmonize the representation of aromatic rings. Finally, duplicates were detected by the analysis of the normalized molecular structures, which contained 75 duplicate compounds for 5-HT1A binders and 17 for non-binders (i.e., different salts or isomeric states). The functional data for duplicated compounds were verified to be identical, so in each case a single example was removed. The curated subset of the original 5-HT1A dataset used in this work included 166 unique organic compounds (105 actives and 61 inactives). All of the details about the dataset are available in the Supporting Information.

### ***5.2.2 QSAR Modeling and External Validation***

We have followed the rigorous QSAR workflow for model building, validation and virtual screening (Figure 1) established in our laboratory<sup>93</sup>. For classification QSAR modeling, it would be ideal to have the balanced ratio between different compound classes in the modeling dataset. However, with 105 binders and 61 non-binders, the 5-HT1A dataset was imbalanced. In the absence of special statistical treatment, such a ratio would skew the prediction accuracy of the classification models. While we do not want to lose any information, different weights for 5-HT1A binders and non-binders were employed during the modeling process. Furthermore, in order to perform a five-fold external set cross-validation protocol, the sample set of 166 compounds was divided into five subsets, with one subset used for external testing and the other four as model training and internal testing. This was repeated five times and a different one-fifth was used for external testing each time. The remaining compounds in the four-fifth section, which were considered modeling dataset,

were further partitioned into multiple pairs of chemically diverse and representative training and test sets of in different sizes, using the Sphere Exclusion approach developed in our laboratory earlier<sup>81, 94</sup>.

Moreover, the dataset of 69 additional 5-HT1A binders from WOMBAT served as independent external validation set. We used models built from 166 5-HT1A binder/non-binder dataset from PDSP to verify those 69 compounds. We should emphasis that these new binders are unique structures from existing PDSP binders. In the consensus prediction process, both model prediction values and model overages were taken into consideration. The success of this additional external validation would suggest that our QSAR models are predictive and robust enough to be applied for virtual screening.

### ***5.2.3 Virtual Screening of Various Types of Libraries***

As illustrated in the workflow of Figure 1, QSAR models that passed both internal and external validation were employed for virtual screening. A global applicability domain (calculated using all descriptors) was applied first in order to filter out compounds that structurally highly different from the compounds in the modeling set. All 105 known 5-HT1A binders extracted from PDSP were used as probes in the calculations. Then, the consensus prediction of various machine learning methods was only conducted on compounds chosen by the global AD. The results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction and the standard deviation of estimated means across all models was small. During the consensus prediction of *k*NN, we restricted ourselves to the most conservative applicability domain for each model using the (*cf.* Equation 4)  $Z_{\text{cutoff}} = 0.5$ .

The screening was performed on various chemical libraries with different characteristics: drug-like databases named the Prestwick Chemical Library <sup>95</sup> and the World Drug Index (WDI <sup>96</sup>), GPCR-targeted databases named the TimTec ActiTarg-G (GPCR) library <sup>97</sup> and the ASINEX Synergy GPCR CNS library <sup>98</sup>, and diversity databases from the TimTec Diversity Set 10K <sup>99</sup> and the ASINEX Diversity Set-Platinum 5K <sup>100</sup>.

All the modeling and virtual screening calculations were done at a 352-processor Beowulf Linux cluster of the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The compute nodes are Intel Xeon IBM BladeCenter of Dual Intel Xeon 2.8GHz, with 2.5GB RAM on each node. The cluster runs the Red Hat Enterprise Linux 4.0 (32-bit) and the nodes communicate via a Gigabit Ethernet network. The processing speed of QSAR-based screening is relatively high, *ca.* 100K compounds per minute. As could be expected, the processing speed was found to scale linearly with the size of the screening library.

#### **5.2.4 Experimental Testing**

For all virtual hits chosen by consensus predictions of *k*NN, RF and SVM, 15 chemicals were further selected including 5 compounds from Prestwick library, 5 from TimTec AntiTarg-G library, and 5 from TimTec Diversity Set 10K. These structurally diverse and commercially available hits were purchased from different suppliers and experimentally tested in PDSP in 5-HT<sub>1A</sub> radioligand binding assays.

## 5.3 Results

Overall, Dragon produced over 2,000 different descriptors. Most of these descriptors characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced solely for bookkeeping purposes. In our study, only 880 chemically relevant descriptors were initially calculated and 672 descriptors were eventually used for 5-HT1A binding dataset after deleting descriptors with zero value or zero variance. Dragon descriptors were range-scaled prior to distance calculations since the absolute scales for Dragon descriptors can differ by orders of magnitude<sup>101</sup>. Accordingly, our use of range-scaling avoided giving descriptors with significantly higher ranges a disproportional weight upon distance calculations in multidimensional Dragon descriptor space.

### 5.3.1 QSAR Classification Models

The *k*NN QSAR method with variable selection afforded multiple models with optimal accuracy characterized as CCR for both training and test sets. In total, there were 838 models with both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  equal to or higher than 0.80. Most models with  $CCR_{\text{test}} \geq 0.80$  also had corresponding  $CCR_{\text{train}} \geq 0.80$ , but the opposite was not always true. The models with high values of both  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  ( $\geq 0.80$ ) were considered acceptable and were selected for consensus prediction. The  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  were found to be as high as 0.91 and 0.99, respectively, which implies that the models could correctly identify 51 binders out of 55 and 34 out of 38 binders ( $SE = 0.93$ ,  $SP = 0.89$ ,  $EN(1) = 1.80$ , and  $EN(2) = 1.85$ ) in the training set and almost all binders and non-binders in the test set. This remarkably high internal accuracy and the large number of acceptable models imply that the

*k*NN classification method was generally successful in correctly distinguishing binders vs. non-binders using Dragon chemical descriptors.

### **5.3.2 QSAR Model Validations**

In addition to the internal validation of *k*NN, RF and SVM models using test sets, Y-randomization and external validation are the critical steps of the entire QSAR workflow (Figure 1). Only models that have been validated by these two steps can be utilized for external prediction and virtual screening<sup>48</sup>.

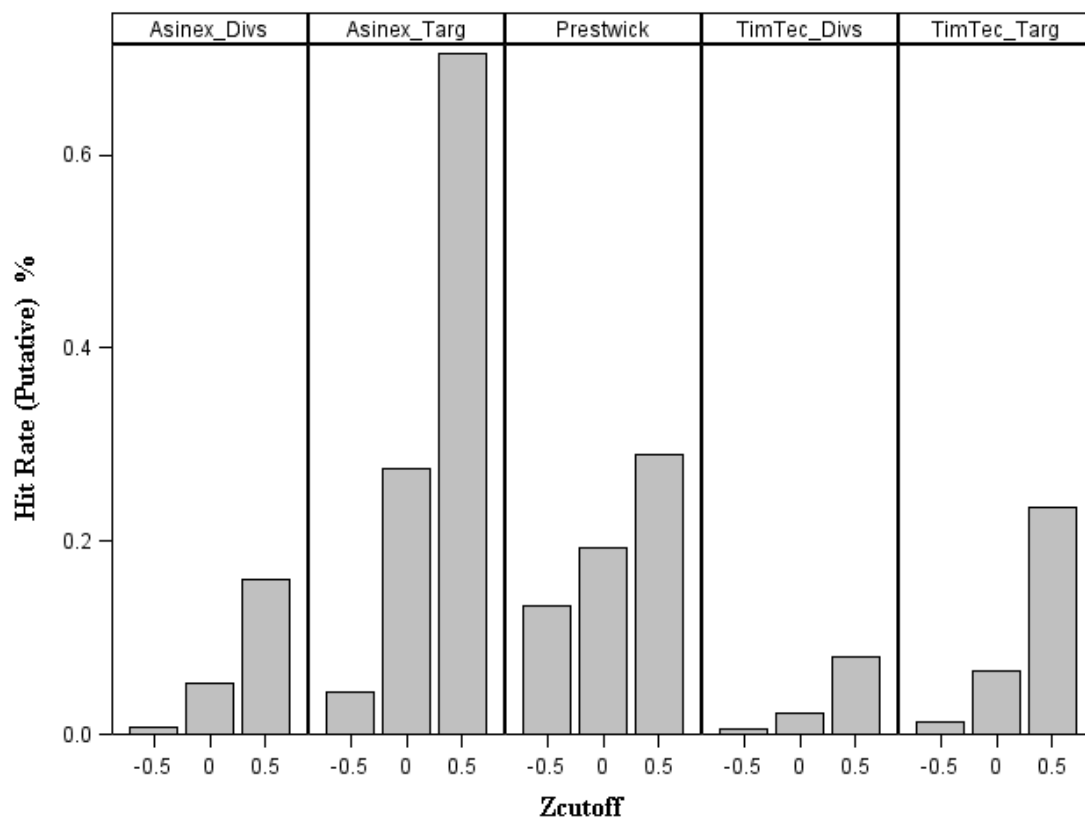
#### **5.3.2.1 Y-randomization Test**

In this Y-randomization test, the binary annotations of 5-HT1A as binders or non-binders were randomly shuffled, and *k*NN, RF and SVM classification models were built with the same parameter settings. The test was performed once for each training/test set split and all runs of Y-randomization tests showed that almost all models had both  $CCR_{train}$  and  $CCR_{test}$  less than 0.70. Moreover, the one-tail hypothesis was applied, and the Z score of 2.17 was calculated given the non hypothesis of QSAR models for the actual dataset being not significantly better than random models. After comparing this Z score with the tabular critical values of  $Z_c$  at different levels of significance ( $\alpha$ )<sup>60</sup>, we concluded that with 98.48% confidence the null hypothesis  $H_0$  should be rejected, and then confirmed that the difference of  $CCR_{train}$  before and after Y-randomization was significant.

#### **5.3.2.2 External Cross Validation**

The five-fold cross validation approach was employed for external prediction, i.e. the 33 compounds randomly excluded from modeling set for each fold. Consensus predictions were carried out using those predictive models with  $CCR_{train}$  and  $CCR_{test}$  greater than 0.8

under different  $Z_{\text{cutoff}}$  values ( $Z_{\text{cutoff}} = 0.5 \sim 3.0$ , Table 1). For Random forest and Support vector machine, exactly the same five-fold external sets were implied for validation, and the prediction results were compared and summarized in Table 2. Because of the applicability domain inherent to individual  $k$ NN QSAR models, the consensus prediction usually cannot cover the whole dataset, i.e., one binder in the first external set cannot be predicted by consensus models using  $Z_{\text{cutoff}} = 0.5$ . Table 1 shows the consensus scores for each of the five fold external sets. The consensus score, in terms of the average class number in classification QSAR, was calculated by the fraction of models that predicted a compound as non-binder over the total number of models used for prediction plus 1. Under  $Z_{\text{cutoff}} = 0.5$ , most of the external validation set achieved a rather high prediction accuracy. For the forth external set split, the prediction achieved 95% for binders and 77% for non-binders, leading to  $\text{CCR}_{\text{evs}} = 0.86$ . Those falsely predicted binders (average class number  $> 1.5$ ) were within an applicability domain of a small portion of all models, i.e., the model coverage was very low and the prediction value is no larger than 1.67. In general, the prediction with such a low coverage is viewed as a low confidence level. The higher  $Z_{\text{cutoff}}$  significantly raised the model coverage for binder and non-binder predictions because of the extended applicability domain for individual models. However, the prediction with extended applicability domain for consensus models also comes with lower confidence level. Generally speaking, in order to have reliable and accurate predictions, one has to have a broader model coverage and a smaller  $Z_{\text{cutoff}}$  value.



**Figure 5.1.** The statistics of five-fold external validations of 5-HT<sub>1A</sub> compounds from PDSP for three QSAR methods and Y-Randomization test.

**Table 5.1.** Results for the five-fold external sets cross validation as well as the secondary external set () validation using three different machine learning methods.

Machine Learning Methods	External Sets	Prediction CCR	Confusion Matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP	TN	FP	FN	SE	SP	EN(1)	EN(2)
<i>k</i> -Nearest Neighbor	1	0.86	19b	14	18	11	3	2	0.90	0.79	1.62	1.77
	2	0.61	20	13	15	6	7	5	0.75	0.46	1.16	1.30
	3	0.77	22	11	20	7	4	2	0.91	0.64	1.43	1.75
	4	0.86	20	13	19	10	3	1	0.95	0.77	1.61	1.88
	5	0.68	23	10	22	4	6	1	0.96	0.40	1.23	1.80
	WOMBAT	0.94	69	0	65	NA	NA	4	0.94	NA	NA	NA
Random Forest	1	0.80	20	14	16	11	4	3	0.80	0.79	1.47	1.70
	2	0.68	20	13	15	8	5	5	0.75	0.62	1.32	1.42
	3	0.84	22	11	21	8	1	3	0.95	0.73	1.83	1.68
	4	0.74	20	13	19	7	1	6	0.95	0.54	1.85	1.28
	5	0.83	23	10	22	7	1	3	0.96	0.70	1.81	1.69
	WOMBAT	0.94	69	0	65	NA	NA	4	0.94	NA	NA	NA
Support Vector Machine	1	0.87	20	14	19	11	1	3	0.95	0.79	1.86	1.68
	2	0.68	20	13	18	6	2	7	0.90	0.46	1.71	1.14
	3	0.95	22	11	22	10	0	1	1.00	0.91	2	1.90
	4	0.76	20	13	18	8	2	5	0.90	0.62	1.71	1.42
	5	0.76	23	10	21	6	2	4	0.91	0.60	1.64	1.55
	WOMBAT	0.96	69	0	66	NA	NA	4	0.96	NA	NA	NA

<sup>a</sup>: N(1) = number of binders, N(2) = number of non-binders, TP = true positive (binders predicted as binders), FP = false positives (non-binders predicted as binders), FN = false negatives (binders predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate.

<sup>b</sup>: Some N(1) binders of and N(2) non-binders were out of application domain of all consensus models, thus having no prediction. Only data for compounds found within the AD were used for statistical summaries.



In summary, 258 models with  $CCR_{\text{train}}$ ,  $CCR_{\text{test}}$  and  $CCR_{\text{evs}}$  equal to or greater than 0.80 could be applied for consensus prediction and virtual screening. The models chosen for the prediction had relatively small  $Z_{\text{cutoff}}$  ( $= 0.5$ ) and relatively broad coverage for compounds in external datasets ( $\geq 50\%$ ).

### 5.3.2.3 Independent External Prediction

We used models built from 166 5-HT1A binder/non-binder dataset from PDSP to verify the 69 5-HT1A binders from WOMBAT. We should emphasize that these new binders are unique structures from existing PDSP binders. Among the 69 binders (all were within the applicability domain), 65 were accurately annotated by  $k$ NN consensus prediction ( $CCR_{\text{ex}} = 0.94$ , **Table 1**). Thus, the majority of ligands were predicted correctly by our consensus models. Since the 4 falsely predicted 5-HT1A binders by  $k$ NN had the prediction values greater than 1.67, and were within the applicability domain of only 70 models (i.e., approximately 30% of all models), the  $k$ NN prediction is considered as of low confidence. When RF and SVM were applied, the prediction accuracy for the additional 69 binders from WOMBAT was also high, ranging from  $CCR_{\text{ex}} = 0.94$  to 0.96 (**Table 1**). The success of this additional external validation suggested that our QSAR models would be predictive and robust enough to be applied for virtual screening.

### 5.3.2.4 QSAR Models based Virtual Screening

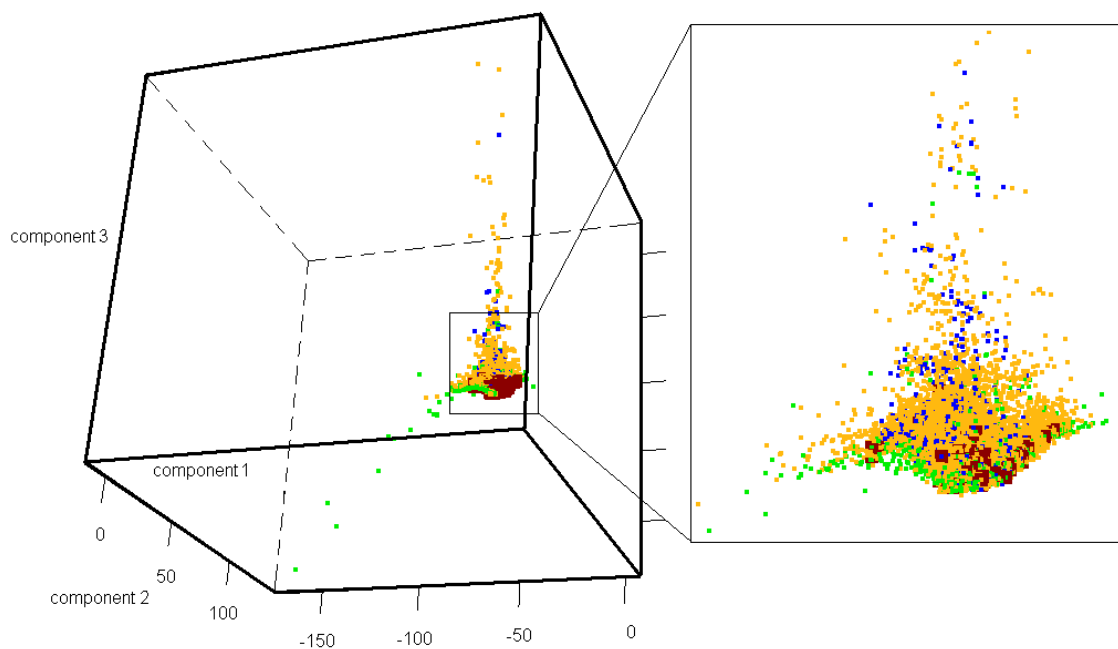
Instead of using only one single and best model for virtual screening, the consensus prediction approach was applied. To perform the consensus prediction, we averaged predictions from all qualified models, i.e. 258 models with both Internal and External  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  equal to or greater than 0.80 from  $k$ NN, 7 models (same criteria) from

Random Forest (RF) and 62 from Support vector machine (SVM). Models generated by the modeling set yielding the highest  $CCR_{\text{evs}}$  (for the third split of five-fold CV) were used for the virtual screening. Initially, 55,384 compounds from the Prestwick Chemical Library (PCL) and World Drug Index (WDI) dataset were screened for 5-HT1A binding, and the numbers of compounds chosen by AD within different  $Z_{\text{cutoff}}$  were shown in Figure 4. The compounds within  $Z_{\text{cutoff}} \leq 0.5$  were further predicted by  $k$ NN consensus models. 234 compounds from Prestwick were predicted as binders by at least one of the  $k$ NN consensus models. To narrow the hit list and obtain the higher confidence level for each prediction, we took both the consensus score (average class number) and model coverage into consideration. In particular, only the hits with average class number between 1.0 and 1.1 and the model coverage over 50% were selected. We found 125 compounds from Prestwick and 181 from WDI satisfied both criteria.

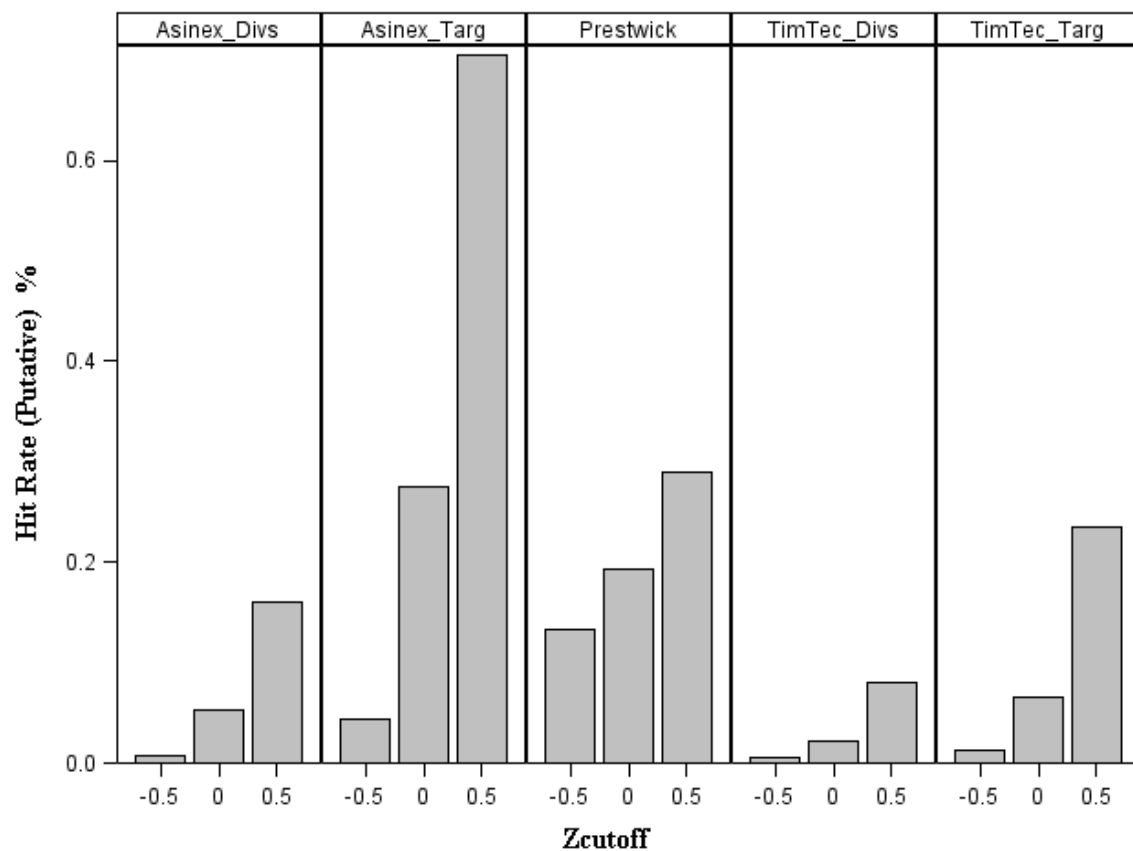
However, the majority of these virtual hits were highly similar to the compounds already known (compounds in the QSAR modeling set), so it would be least attractive to test these hits experimentally. To verify the diversity of those virtual hits, pairwise similarity calculations were performed. Each compound was represented by a fingerprint of 166 substructure keys (MACCS structural keys<sup>102</sup>), indicating the presence or absence of a particular chemical substructure. The pairwise similarity was measured by using the Tanimoto coefficients ( $T_c$ ) to compare the Prestwick virtual hits versus themselves, Prestwick virtual hits versus each hit's nearest neighbor from the binders in the modeling set (identified by Dragon descriptors and Euclidean distances), Prestwick virtual hits versus the binders used in model building, and the binders used in model building versus themselves. The majority of compound pairs between Prestwick virtual hits versus each hit's nearest

neighbor within the modeling set have Tc over 0.9, while other pair-wise similarity scores show a normal distribution, suggesting that the virtual hits are structurally highly similar with our already known 5-HT1A binders (**Figure 5**).

To explore more structurally diverse 5-HT1A compounds, we further screened GPCR-targeted libraries and diversity libraries from the commercial chemical sources of both TimTec and ASINEX. Therefore, the additional collection of 24,000 compounds were screened, which includes the TimTec ActiTarg-G (GPCR) library of about 2,300 compounds, the ASINEX Synergy GPCR CNS library of about 7,000 compounds, the TimTec Diversity Set 10K of 10,000 compounds and the ASINEX Diversity Set-Platinum 5K of about 5,100 compounds. By applying various AD, the putative hit rate for different screening libraries within various  $Z_{\text{cutoff}}$  values was shown in Figure 4, and the exact numbers of compounds chosen from them were also available in supplementary material (**Table S1**). It is obvious that many more chemicals were selected from the GPCR library than the diversity library by applying the same  $Z_{\text{cutoff}}$  value, verifying that the diversity library has much more structural-varied compounds compared with our modeling set than the GPCR-targeted library.



**Figure 5.2. The PCA plot of three virtual screening libraries and modeling set compounds.** Chemical compounds in modeling set are labeled red; Chemical compounds in Prestwick library are labeled green; Chemical compounds in TimTec AntiTarg-G library are labeled blue; Chemical compounds in TimTec Diversity Set 10K are labeled orange.

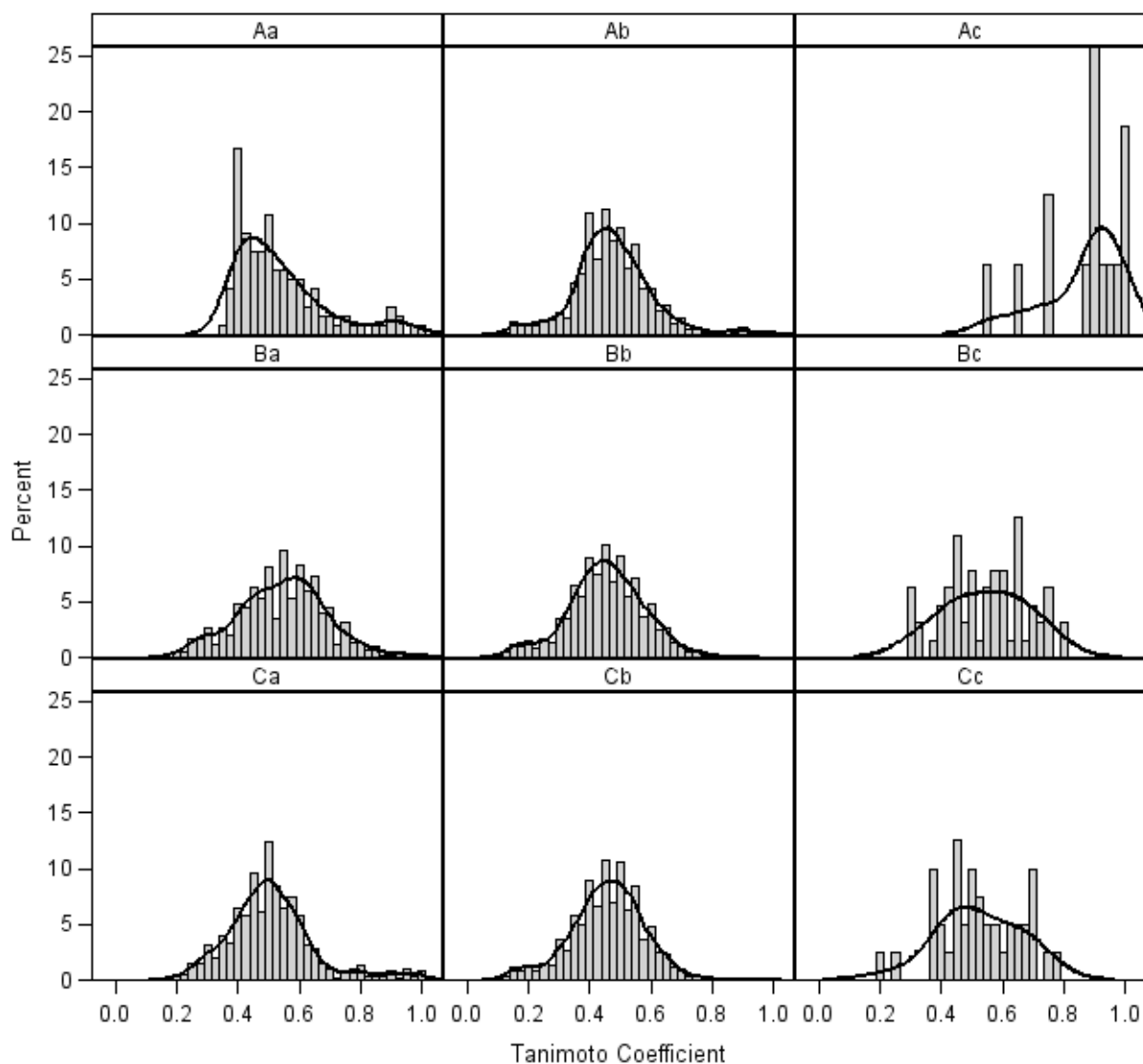


**Figure 5.3. Hit rate of 5-HT1A binders on diverse screening libraries using different  $Z_{\text{cutoff}}$  values.**

The compounds within  $Z_{\text{cutoff}}$  0.5 were further predicted by  $k$ NN consensus models. 445 compounds from the TimTec ActiTarg-G library, 487 from the TimTec Diversity Set 10K, 2,177 from the ASINEX Synergy GPCR CNS library and 782 from the ASINEX Diversity Set-Platinum 5K were predicted as binders by at least one of the  $k$ NN consensus models. To narrow the hit list and obtain the higher confidence level for each prediction, both the consensus score and model coverage were taken into account. In particular, only the hits with average class numbers between 1.0 and 1.1 and the model coverage over 50% were selected. We found that there were 64 compounds from the TimTec AntiTarg-G library and 40 from the TimTec Diversity Set 10K that satisfied both criteria. As for ASINEX libraries, there were still hundreds of compounds that met those strict criteria, so we will not take those into consideration at this time.

Several structural classes were observed by screening different libraries according to the Tanimoto coefficients (Tc) values. Notably, many of the 64 virtual hits from the TimTec AntiTarg-G library were found to be structurally similar to binders used in model building, while the 40 virtual hits from the TimTec Diversity Set 10K displayed highly different structural profiles. The pairwise similarity measured by Tc values was also compared between virtual hits versus virtual hits, hits versus their nearest neighbor within the modeling set compounds, virtual hits versus modeling set compounds, and modeling set compounds versus themselves (**Figure 5**). It is clearly seen that the virtual hits from the TimTec AntiTarg-G library showed structural profiles with a much lower similarity to the known 5-HT1A binders than Prestwick virtual hits. The average Tc value between TimTec Anti-Targ-G library hits and their nearest neighbors in the modeling set was 0.6 compared to 0.9 for the hits screened from Prestwick. For our virtual hits screened from the TimTec Diversity Set

10K, the Tc value between hits and their nearest neighbors in modeling set is as low as 0.45, suggesting that they are highly structurally different. While these hits are also predicted to be 5-HT1A binders with a high confidence by our consensus models as well as random forest and support vector machine, it would be interesting and exciting to test them experimentally, in hope of revealing new scaffold of 5-HT1A binders.



**Figure 5.4. The structural similarity analysis of virtual hits screened from different libraries.**

A: Tanimoto Coefficient (Tc) between Prestwick virtual hits and themselves (Aa), Prestwick virtual hits and modeling set compounds (Ab), and Prestwick virtual hits and their nearest neighbor compounds in the modeling set (Ac). B: Tc between TimTec AntiTarg-G library virtual hits and themselves (Ba), Target library virtual hits and modeling set compounds (Bb), and Target library virtual hits and their nearest neighbor compounds in the modeling set (Bc). C: Tc between TimTec Diversity library virtual hits and themselves (Ca), Diversity library virtual hits and modeling set compounds (Cb), and Diversity library virtual hits and their nearest neighbor compounds in the modeling set (Cc).

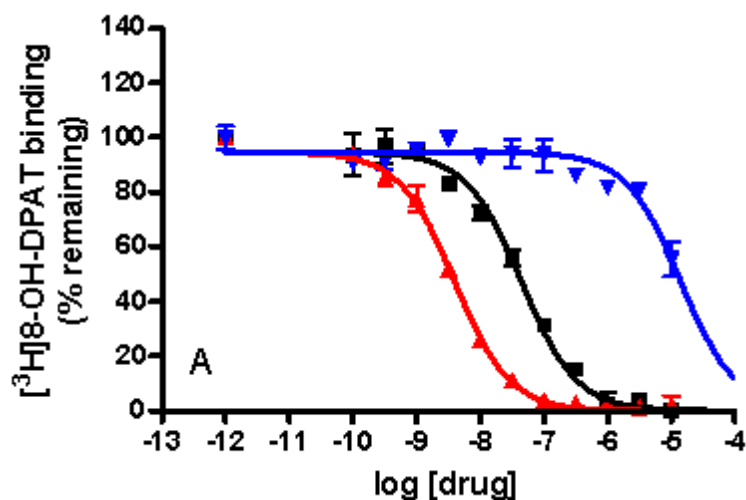


For all virtual hits chosen by *k*NN, 15 chemicals were further selected for experimental testing, including 5 compounds from Prestwick library, 5 from TimTec AntiTarg-G library, and 5 from TimTec Diversity Set 10K. The following criteria were met when selection was performed: 1) High confidence of consensus prediction by RF and SVM; 2) Low structural similarity between hits and the 5-HT1A binders we already known; 3) Convenient commercial availability.

### 5.3.2.5 Experimental Validation

The validations on our *in silico* hits by the NIMH PDSP were satisfying and yielded many true hits as 5-HT1A binders. We should stress that only binary QSAR models were used for screening so no estimate of exact binding affinities ( $K_i$  values) had been made. Nine out of fifteen *in silico* hits have the percentage of inhibition at or higher than 50% (i.e. Mesoridazine, Clozapine, Risperidone and Fluphenazine from PCL; ST030580 from GPCRs targeted library; ST023860, ST074311 and ST057540 from diversity library) and six of them even higher than 95%. For these compounds, the  $IC_{50}$  values were obtained from non-linear regression of radioligand competition binding isotherms, from which the final  $K_i$  (nM) values were calculated using the Cheng-Prusoff equation. The five *in silico* hits from PCL showed the highest success rate (80%), though most of them were similar to the modeling set compounds ( $T_c$  ranged from 0.80 to 0.99, with an average  $T_c$  value of 0.86) and no novel core scaffolds were found. They were also found to be less interesting from the point of view of drug repurposing. Mesoridazine and fluphenazine belong to the typical antipsychotics while clozapine and risperidone are atypical antipsychotics; all four compounds had been employed in the treatment of schizophrenia and bipolar disorder in clinics.

To our surprise, only one *in silico* hit from the GPCRs targeted library had been proved to be active ( $K_i = 243.8$  nM). This compound, ST030580, showed quite different rings arrangement from its nearest neighbor in the modeling set while maintaining the azaspiro-bicyclic structural element. Among the three confirmed hits from the TimTec Diversity Set 10K, compound ST057540 (also known as Lysergol ( $[(8\alpha)\text{-}6\text{-methyl-}9,10\text{-didehydroergolin-}8\text{-yl]methanol}$ )) yielded 98.20% binding inhibition against 5-HT1A receptor and its  $K_i$  value is 2.3 nM (**Figure 6**). Furthermore, the Tc between this compound and its nearest neighbor in the modeling set (ID: 27405, with dibenzo[de,g]quinolone structure) is only 0.69, indicating the structural distinctions in general. Lysergol is an alkaloid of the ergoline family that occurs as a minor constituent in some species of fungi, and is sometimes utilized as an intermediate in the manufacturing of some ergoloid medicines (e.g., nicergoline). This compound qualifies for all of the “Lipinski Rule of Five”, with a LogP value of 1.76<sup>103, 104</sup>, which is considered to be ideal for both oral absorption and CNS penetration. It was also predicted to have very low probability of rapid biodegradation by EPI-Suite<sup>105, 106</sup>. Lysergol does not have a known pharmacological action or a precursor relationship to LSD, and its pharmacological indication remained to be further explored. Two other active hits, compounds ST023860 and ST074311, also show relatively different scaffolds in comparison to modeling set compounds with Tc of 0.75 and 0.69 respectively. The findings were encouraging and some novel scaffolds identified are currently under patent application.

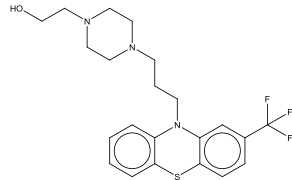
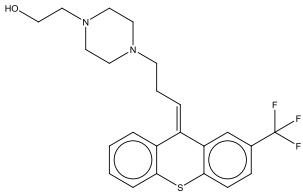
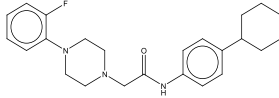
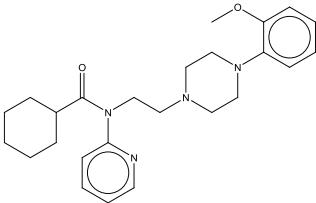
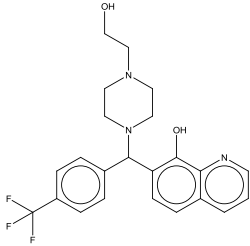
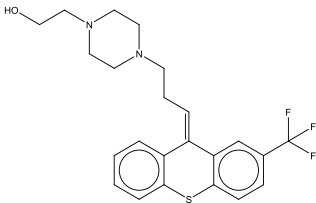


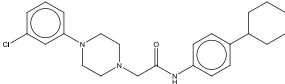
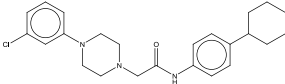
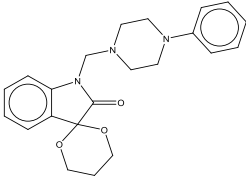
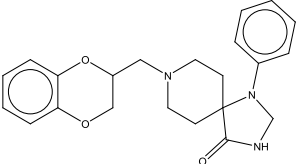
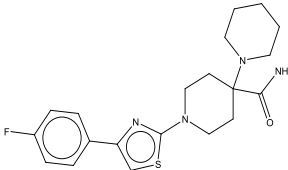
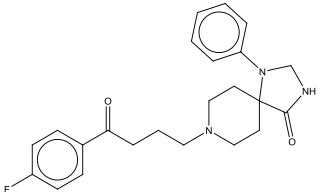
**Figure 5.5. The full dose response curves for hit compounds and the positive control.** Hit compounds ST057540 ( $K_i = 2.3$  nM) and ST074311 ( $K_i = 8,194$  nM) are represented in red and blue triangles, and the positive control, Methysergide ( $K_i = 26$  nM), is in black squares. The full dose response curves show the results in human 5-HT<sub>1A</sub> receptor radioligand binding assay, with [<sup>3</sup>H]-8-OH-DPAT used as the radioligand at the concentration of 0.5 nM in the standard binding buffer.

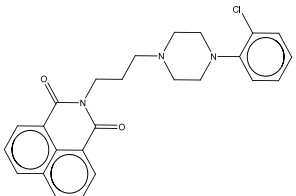
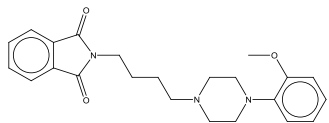
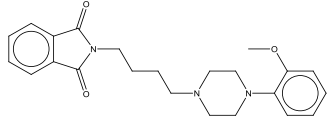
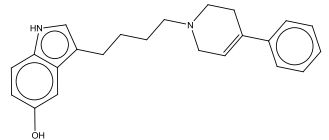
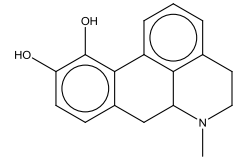
In summary, the above results once again proved the predictive power of our binary *k*NN, RF and SVM classification QSAR models built from 5-HT1A binders/non-binders. These studies illustrate that the validated QSAR workflow, as employed in this paper, could be used as a general tool for identifying promising hits by the means of virtual screening of various types of chemical libraries.

**Table 5.2.** The experimental test for the ten virtual hits of 5-HT<sub>1A</sub> binders identified by virtual screening.

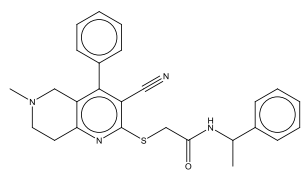
Structure	No.	Name or ID	Source library	Most similar compound in modeling set	Tc	kNN score	RF score	SVM score	Percent of inhibition	Exp. IC <sub>50</sub> (nM)
	1	Mesoridazine	Prestwick library		0.80	1.01	1.06	1.00	> 95%	33.1
	2	SKF-75670	Prestwick library		0.80	1.02	1.04	1.00	N/A	> 10,000
	3	Clozapine	Prestwick library		0.83	1.04	1.07	1.00	> 95%	104.8
	4	Risperidone	Prestwick library		0.99	1.00	1.02	1.00	N/A	427.5

		5	Fluphenazine	Prestwick library		0.89	1.00	1.13	1.00	> 95%	145.7
		6	ST016950	TimTec GPCR Targeted library		0.66	1.07	1.10	1.00	5.5% <sup>a</sup>	N/A
		7	ST014829	TimTec GPCR Targeted library		0.73	1.06	1.27	1.00	39.20%	N/A

	8	ST007472	TimTec		0.76	1.03	1.09	1.00	21.80%	N/A
			GPCR							
			Targeted library							
	9	ST030580	TimTec		0.84	1.02	1.09	1.00	95.50%	243.8 <sup>a</sup>
			GPCR							
			Targeted library							
	10	ST041900	TimTec		0.74	1.07	1.09	1.00	4.70%	N/A
			GPCR							
			Targeted library							

	11	ST023860	TimTec Diversity Set 10K		0.75	1.02	1.12	1.00	95.60%	159.0
	12	ST007110	TimTec Diversity Set 10K		0.84	1.08	1.14	1.06	11.10%	N/A
	13	ST074311	TimTec Diversity Set 10K		0.69	1.10	1.14	1.00	70.80%	6261.0
Patent under application	14	ST057540	TimTec Diversity Set 10K		0.69	1.05	1.14	1.00	98.20%	3.4

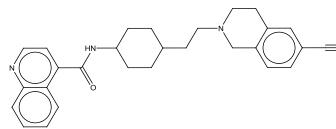




15

ST066677

TimTec  
Diversity  
Set 10K



0.53

1.06

1.13

1.00

49.60%

N/A

<sup>a</sup>: The full IC<sub>50</sub> curve was generated in further experiment and the Ki value was determined.

## 5.4 Discussions

We should emphasize that our model validation is a critically inherent feature of our QSAR modeling workflow. This issue of model validation had been given a lot of attention by the QSAR research community <sup>107</sup>. Until recently, most practitioners merely presumed that internally cross-validated models built from available training set data should be externally predictive. We and others have demonstrated that internal validation techniques such as leave-one-out (LOO) or even leave-many-out (LMO) cross-validation applied to the training set is insufficient to ensure the external predictive power of QSAR models <sup>15, 48</sup>. Thus, we used five-fold cross external validation sets in this study as well as the Y-randomization test to ensure the robustness and predictive power of *k*NN models. Needless to say, the use of externally validated models and applicability domains is especially critical when the models are employed in virtual screening.

Another important feature of many current biomolecular datasets, especially those generated as a result of High Throughput Screening (HTS), is the imbalance between “actives” and “inactives.” While in this study there are more actives, in many other cases instances of inactives will predominate; for example, the hit rates in assays deposited in PubChem by the NIH screening centers forming the Molecular Library Screening Center Network (MLSCN) are very low, in most cases not exceeding 0.5% <sup>108</sup>. The imbalanced datasets pose a significant problem for classification QSAR modeling because models that correctly predict the same fraction of objects in each class will have different objective function values. To circumvent this problem in this study, we assigned different weights of objective functions to the underrepresented class (non-binders) versus the other one (binders) for model building. The classification models built for the unbalanced subset with different

weight objective functions were shown to predict compounds in this external dataset as binders versus non-binders with very high accuracy.

Moreover, another unique 69 5-HT1A binders from different resources (i.e. WOMBAT database) were further validated by our consensus models for independent external validation. All three QSAR methods (*k*NN, RF and SVM) can accurately annotate the majority of compounds, with CCR<sub>ex</sub> ranged from 0.94 to 0.96. This additional independent external validation proposed in our study is unique, and has not yet been used elsewhere before. The success of this approach strongly suggested our QSAR models to be predictive, robust, and ready for virtual screening.

Finally, model-based virtual screening was performed on various databases with different characters, including two drug-like libraries, two GPCR-targeted libraries, and two diversity libraries. Both the global similarity search (using AD) and the subsequent QSAR model predictions confirmed our expectations that drug-like libraries and GPCR-targeted libraries had a much higher hit rate than diversity libraries, when the same cutoff values were applied, though the reason why ASINEX libraries had an extraordinarily high hit rate remained unclear. When prediction of QSAR models were made and the pairwise similarity analysis was further performed, it once again confirmed our hypothesis that those virtual hits from drug-like libraries had much higher structural similarity with our modeling set compounds than hits from GPCR-targeted libraries and diversity libraries. After experimental validation, 60% of the compounds suggested by our QSAR models were confirmed to be 5-HT1A binders; however, it was interesting to know that the experimental hit rate of the diversity library is much higher than the GPCR-targeted library, and the most potent 5-HT1A binder (inhibitor) was screened from diversity library, sharing a very low structural similarity with

5its nearest neighbor compound in the modeling set. These interesting findings verified that model-based virtual screening outperformed the simple similarity search, and also challenged our conventional opinions about structure-activity relationships (SAR), suggesting that it is not always true that more similar structures will lead to more similar chemical properties. Moreover, it is once again confirmed that by taking the advantage of various computational tools, such as QSAR modeling, more novel compounds could be revealed with diverse scaffolds.

## 5.5 Conclusions

Our studies demonstrate that classification QSAR models built with Dragon descriptors can accurately differentiate true 5-HT1A binders from non-binders. A special QSAR modeling scheme was employed for this imbalanced dataset and the models were rigorously validated using both internal (multiple training/test set divisions and Y-randomization) as well as external (five-fold cross external validation sets) validation approaches. We have demonstrated that this strategy afforded multiple QSAR models with high internal and external predictive power. As part of our QSAR modeling workflow, the predictors were further utilized for mining the WOMBAT hits (69 literature extracted compounds tested for 5-HT1A binding). We found that our validated models agreed highly with the experimental annotation of 69 compounds as 5-HT1A binders as reported in various literatures (extracted through WOMBAT database). On the other hand, our models used in the most conservative way (i.e., in consensus fashion and with the strictest applicability domain criteria) did identify 43 putative 5-HT1A binders among the TimTec AntiTarg-G library and TimTec Diversity Set 10K. Ten of them were tested experimentally in Dr. Roth's lab at UNC and all showed inhibition activities at a single concentration for percentage of

inhibition. Interestingly, the five virtual hits identified from the TimTec Diversity Set 10K showed higher 5-HT1A binding affinity than the other five from the TimTec AntiTarg-G library. One compound (compound ST057540) was found to have the highest  $K_i$  of 2.3nM, while the Tanimoto coefficients between this compound and its nearest neighbor in the modeling set (ID: 27405) was as low as 0.52. The results of our studies suggest that at least in some cases when a sufficient amount of data on true binders vs. nonbinding compounds is available, QSAR modeling approaches could be used successfully to complement (and possibly educate based on QSAR model interpretation) the conventional scoring functions used in three-dimensional docking studies. Furthermore, as we have demonstrated in this paper, QSAR models can be successfully used not only to discriminate binders vs. non-binders but most importantly, for finding promising hits by the means of virtual screening of chemical libraries.

The heatmap of the self-similarity matrix for 5-HT1A modeling set, distributions of models for Y-randomization tests, experimental data of 5-HT1A screening hits binding affinities, chemical structures and  $pIC_{50}$  values for 5-HT1A modeling dataset and screening hits, purity data for target compounds, and others supplementary data indicated in the text are available in the Appendix section.

## CHAPTER 6

### SUMMARY AND FUTURE DIRECTIONS

Combinations of structure-based and ligand-based approaches are being used to assist the medicinal chemists in identifying and designing ligands that could pharmacologically modulate the target of interest. Computer-Aided Drug Design methods can be categorized based on whether the three-dimensional (3D) structure of the target protein is available. If a crystal structure of target protein or receptor is accessible, structure-based drug design approaches, such as *de novo* design, docking-scoring, structure-based pharmacophoric search, could be used. Those compounds with high structural and physic-chemical complementarities to the active site are ranked according to their scores prioritized for experimental tests. If the structure of the target protein is not known, which is a more common case, ligand-based drug design methods are used, which only relies on knowledge of other molecules that bind to the biological target of interest. In this case, the chemical structures ( $m$  molecules) are represented by strings of numeric characters calculated by different types of descriptors ( $n$  descriptors). The generated  $m \times n$  matrix is then analyzed by diverse data analysis approaches to predict new molecules.

Many methods for multi-target predictions mainly consider of the computational efficiency; however, when simple similarity search is used for modeling, with the aim of fast prediction, the prediction accuracy cannot always be achieved at the same time. This feature, alternatively, is the most important aspect about which researches should care. Furthermore,

most current approaches of ligand-based drug discovery mainly focus on optimizing the computational algorithms to improve the efficiency and/or accuracy of virtual screening; however, the success of ligand-based drug design relies not only on the effectiveness and robustness of the underlying algorithms, but much more importantly, on the quality of the data for model building. Although numerous chemical probe databases have emerged recently, seldom evaluation of data quality and reliability was performed.

## **6.1 Summary and Future Directions of Chapter 2**

In this chapter, various state-of-the-art data analysis tools in cheminformatics for predicting compounds' biological properties are introduced. For supervised statistical learning, QSAR methods are discussed, including diverse algorithms such as *k*NN, RF and SVM. The concept of applicability domains and external validation tests is also covered. For the unsupervised statistical learning techniques, the underlying algorithms for two multi-profile prediction approaches, the Similarity Ensemble Search (SEA) and Prediction of Activity Spectra for Substances' (PASS) are introduced. These methods have been intensively compared by 7 cases of biological receptors, in terms of both internal recovery rate as well as external prediction accuracy.

The results showed that the internal recovery rate of SEA was about or less than 60%, while QSAR always achieved 100%. For external compounds' prediction, the results further highlighted the use of QSAR over the SEA prediction method. SEA can only predict less than half of the external validation set compounds, while QSAR achieved almost 100% for the three GPCR targets. PASS has been validated to have moderate prediction accuracy, both for internal recovery rate and external prediction accuracy.

For future works, more biological targets besides GPCR receptors will be tested for the comparison of these three methodologies. Moreover, the server for multi-target QSAR prediction is aimed to launch, when more readily developed QSAR models become available.

## 6.2 Summary and Future Directions of Chapter 3

The Molecular Libraries Program (MLP), a NIH Roadmap Initiative, aims to enhance chemical biology data through High Throughput Screening (HTS) to obtain chemical probes effective at modulating specific biological processes or disease states. PubChem is an open-access data repository system, acting as the portal site for MLP. To evaluate the quality of some biological activities deposited in PubChem, we have conducted *in silico* modeling studies for 5-Hydroxytryptamine Receptor Subtype 1A (5-HT1A) ligands. Our studies demonstrated that classification QSAR models can accurately differentiate true 5-HT1A binders from non-binders in reliable data sources, such as PDSP and WOMBAT. By contrast, we failed to generate qualified models using datasets deposited in PubChem (PubChem Assay id (AID) 613, 718, 755). The combinatorial QSAR modeling scheme is employed for all three 5-HT1A datasets and the models are rigorously validated using both internal (multiple training/test sets, Y-randomization test) as well as external (five-fold cross validation) validation. We have demonstrated that this strategy afforded QSAR models with high internal and external predictive power. Moreover, the predictors are further utilized for cross-validation of the 5-HT1A datasets from different sources. We find that the prediction results of our validated models highly agree with the experimental annotation of 69 5-HT1A binders as reported in WOMBAT database. Furthermore, the nine PubChem binders, which were identified by cross-validated QSAR models as false positive, were sent for experimental testing. The experimental results showed 100% agreement with our models' consensus



predictions, and confirmed that those compounds are mislabeled in data from PubChem confirmatory assays.

In conclusion, applied to reliable data sources from PDSP and WOMBAT databases, we have built validated robust and externally predictive QSAR models for 5-HT<sub>1A</sub> receptor ligands using *k*NN, RF and SVM methods. Nine false positive PubChem 5-HT<sub>1A</sub> binders identified by our QSAR models were further confirmed with 100% accuracy by the experimental test, suggesting that our models were also powerful enough to detect high noise to signal data sources.

For future works, more data depositions in PubChem will be validated, for the purpose of identifying universally high noise-to-signal data sources, and thus improving our database quality by publishing our results and warning such misleading efforts.

### **6.3 Summary and Future Directions of Chapter 4**

Antibiotic resistance is an escalating problem requiring the discovery of novel antibiotic classes acting on nonclassical cellular targets. Targeting the nonessential genes, for example RecA, offers possible attractive solution. In this study, we have developed combinatorial Quantitative Structure-Activity Relationship (QSAR) model for hundreds of chemically diverse RecA inhibitors and their structurally similar inactive compounds resulting from high-throughput screening and the subsequent confirmatory binding assays. The initial attempts to classify 53 RecA inhibitors out of over 3,000 non-inhibitors met with only limited success, due to the fact that activity cliff exists in several highly similar compound pairs, identified by pair-wise Tanimoto Coefficient (Tc) analysis. Then, a new dataset, containing 145 RecA inhibitors and 26,132 non-inhibitors, was created after both

data curation of those activity cliff pairs and incorporation of more up-to-date experimental testing results. The new dataset was clustered into three groups according to structure similarity, and then the variable selection  $k$ NN, RF and SVM, were employed for model building within each group using 2D topological Dragon chemical descriptors. Highly predictive QSAR models were generated and the  $CCR_{\text{evs}}$  values were much higher than the  $CCR_{\text{evs}}$  for model building without clustering. With two differently defined applicability domain thresholds, all validated QSAR models were employed concurrently for virtual screening (VS) of an in-house compound collection including 9.5 million molecules. VS resulted in 31 structurally unique consensus hits that were considered novel putative RecA inhibitors, with novel structural features that were not present in the original data set. 11 of *in silico* hits with novel scaffolds were tested experimentally and five of them were confirmed active against RecA, with  $IC_{50}$  values ranging from 5 to 28  $\mu\text{M}$ . Overall, this study illustrates the power of the combinatorial QSAR-VS method as a general approach for the effective identification of structurally novel bioactive compounds.

For future studies, pharmacophore modeling of existing and newly discovered RecA inhibitors will be conducted, with hopes of optimizing the potency of current RecA inhibitors. Moreover, lead compounds of high interest will be tested *in vivo*, for their toxicity studies, as well as the investigation of their efficacies in animal models.

## 6.4 Summary and Future Directions of Chapter 5

The 5-Hydroxy Tryptamine receptor subtype 1A (5-HT1A) is highly expressed in the *raphe* nuclei region and limbic structures; for that reason 5-HT1A has been an attractive target to treat mood disorders such as anxiety and depression. Our studies demonstrated that

combinatorial classification QSAR models built with Dragon descriptors can accurately differentiate true 5-HT1A binders from non-binders. A special QSAR modeling scheme was employed for this imbalanced dataset and the models were rigorously validated using both internal (multiple training/test set divisions and Y-randomization) as well as external (five-fold cross external validation sets) validation approaches. We have demonstrated that this strategy afforded multiple QSAR models with high internal and external predictive power. As part of our QSAR modeling workflow, the predictors were further utilized for validating the WOMBAT hits, with results that highly agree. On the other hand, our models used in the most conservative way (i.e., in consensus fashion and with the strictest applicability domain criteria) identified 120 putative 5-HT1A binders by virtually screened the drug-like libraries, GPCR-targeted libraries, and diversity libraries. After experimental validation on commercially available compounds, 60% of the compounds suggested by our QSAR models were confirmed to be 5-HT1A binders; however, it was interesting to learn that the experimental hit rate of the diversity library is much higher than the GPCR-targeted library, and the most potent 5-HT1A binder (inhibitor) was screened from the diversity library, sharing a very low structural similarity with its nearest neighbor compound in the modeling set. These interesting findings verified that model-based virtual screening outperformed the simple similarity search, and also challenged our conventional opinions about structure-activity relationships (SAR), suggesting that it is not always true that more similar structures will lead to more similar chemical properties. Moreover, it is once again confirmed that by taking advantage of the various computational tools, such as QSAR modeling, more novel compounds could be revealed with diverse scaffolds.

Furthermore, as we have demonstrated in this paper, QSAR models can be successfully used not only to discriminate binders versus non-binders but most importantly, for finding promising hits by the means of virtual screening of chemical libraries

## Appendix:

**Table A1.** Five-fold external cross validation test statistics for 5-HT1A binders and non-binders from PDSP

QSAR Methods	External Sets	Prediction CCR	Confusion Matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP <sup>a</sup>	TN <sup>a</sup>	FP <sup>a</sup>	FN <sup>a</sup>	SE <sup>a</sup>	SP <sup>a</sup>	EN(1) <sup>a</sup>	EN(2) <sup>a</sup>
<i>k</i> -Nearest Neighbor	1	0.86	19 <sup>b</sup>	14	18	11	3	2	0.90	0.79	1.62	1.77
	2	0.61	20	13	15	6	7	5	0.75	0.46	1.16	1.30
	3	0.77	22	11	20	7	4	2	0.91	0.64	1.43	1.75
	4	0.86	20	13	19	10	3	1	0.95	0.77	1.61	1.88
	5	0.68	23	10	22	4	6	1	0.96	0.40	1.23	1.80
Random Forest	1	0.80	20	14	16	11	3	4	0.80	0.79	1.58	1.59
	2	0.68	20	13	15	8	5	5	0.75	0.62	1.32	1.42
	3	0.84	22	11	21	8	3	1	0.95	0.73	1.56	1.88
	4	0.74	20	13	19	7	6	1	0.95	0.54	1.35	1.83
	5	0.83	23	10	22	7	3	1	0.96	0.70	1.52	1.88
Support Vector Machine	1	0.87	20	14	19	11	3	1	0.95	0.79	1.63	1.88
	2	0.68	20	13	18	6	7	2	0.90	0.46	1.25	1.64
	3	0.95	22	11	22	10	1	0	1.00	0.91	1.83	2.00
	4	0.76	20	13	18	8	5	2	0.90	0.62	1.40	1.72
	5	0.76	23	10	21	6	4	2	0.91	0.60	1.39	1.75

<sup>a</sup> N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate.

<sup>b</sup> Some N(1) inhibitors of and N(2) non-binders were out of application domain of all consensus models, thus having no prediction. Only data for compounds found within the AD were used for statistical summaries.

**Table A2.** Five-fold external cross validation test statistics for 5-HT1A binders from WOMBAT and non-binders from PDSP.

QSAR Methods	External Sets	Prediction CCR	Confusion Matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP <sup>a</sup>	TN <sup>a</sup>	FP <sup>a</sup>	FN <sup>a</sup>	SE <sup>a</sup>	SP <sup>a</sup>	EN(1) <sup>a</sup>	EN(2) <sup>a</sup>
<i>k</i> -Nearest Neighbor	1	0.93	15	12	14	11	1	0	0.93	0.92	1.93	1.86
	2	0.92	13	12	12	11	2	1	0.92	0.91	1.69	1.85
	3	0.91	15	11	15	9	2	0	1.00	0.82	1.69	2.00
	4	0.96	13	13	13	12	1	0	1.00	0.92	1.86	2.00
	5	0.96	13	13	13	12	1	0	1.00	0.92	1.86	2.00
Random Forest	1	0.96	15	12	15	11	1	0	1.00	0.92	1.85	2.00
	2	0.88	13	12	12	10	2	1	0.92	0.83	1.69	1.83
	3	0.92	15	11	14	10	1	1	0.93	0.91	1.82	1.86
	4	0.96	13	13	13	12	1	0	1.00	0.92	1.86	2.00
	5	0.85	13	13	11	11	2	2	0.85	0.85	1.69	1.69
Support Vector Machine	1	0.89	15	12	13	11	1	2	0.87	0.92	1.82	1.75
	2	1.00	13	12	13	12	1	0	1.00	1.00	1.85	2.00
	3	0.89	15	11	13	10	1	2	0.87	0.91	1.81	1.74
	4	0.81	13	13	10	11	2	3	0.77	0.85	1.67	1.57
	5	0.77	13	13	10	10	3	3	0.77	0.77	1.54	1.54

<sup>a</sup> N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate.

**Table A3.** Five-fold external cross validation test statistics for 5-HT1A agonists/antagonists from PubChem.

QSAR Methods	External Sets	Prediction CCR	Confusion Matrix						Statistics			
			N(1) <sup>a</sup>	N(2) <sup>a</sup>	TP <sup>a</sup>	TN <sup>a</sup>	FP <sup>a</sup>	FN <sup>a</sup>	SE <sup>a</sup>	SP <sup>a</sup>	EN(1) <sup>a</sup>	EN(2) <sup>a</sup>
<i>k</i> -Nearest Neighbor	1	0.61	6	14	3	10	4	3	0.50	0.71	1.27	1.18
	2	0.67	11	10	6	8	2	5	0.55	0.80	1.46	1.28
	3	0.72	10	11	8	7	4	2	0.80	0.64	1.38	1.52
	4	0.55	7	13	4	7	6	3	0.57	0.54	1.11	1.11
	5	0.43	11	10	3	6	4	8	0.27	0.60	0.81	0.90
Random Forest	1	0.46	6	14	3	6	8	3	0.50	0.43	0.93	0.92
	2	0.57	11	10	6	6	4	5	0.55	0.60	1.15	1.13
	3	0.70	10	11	4	11	0	6	0.40	1.00	2.00	1.25
	4	0.68	8	13	6	8	5	2	0.75	0.62	1.32	1.42
	5	0.53	11	10	4	7	3	7	0.36	0.70	1.10	1.05

<sup>a</sup> N(1) = number of inhibitors, N(2) = number of non-binders, TP = true positive (inhibitors predicted as inhibitors), FP = false positives (non-binders predicted as inhibitors), FN = false negatives (inhibitors predicted as non-binders), TN = true negative (non-binders predicted as non-binders), SE = sensitivity = TP/N(1), SP = specificity = TN/N(2), EN - the normalized enrichment, EN(1) = (2TP \* N(2))/(TP \* N(2) + FP \* N(1)), EN(2) = (2TN \* N(1))/(TN \* N(1) + FN \* N(2)), and CCR = correct classification rate.

**Table A4.** The number of screening compound within different  $Z_{\text{cutoff}}$  values for 5-HT1A virtual screening.

Screening Databases	Number of Compounds	$Z_{\text{cutoff}}$	Compounds in AD	Compounds in AD (%)
Prestwick	1,552	-0.5	209	13.5%
		0	304	19.6%
		0.5	458	29.5%
World Drug Index	53,382	-0.5	1334	2.5%
		0	3295	6.2%
		0.5	7371	13.8%
TimTec GPCRTargeted Library	2,300	-0.5	31	1.3%
		0	151	6.6%
		0.5	542	23.6%
Asinex GPCR Targeted Library	3,233	-0.5	144	4.5%
		0	890	27.5%
		0.5	2279	70.5%
TimTec Diversity Library	10,000	-0.5	46	0.5%
		0	222	2.2%
		0.5	803	8.0%
Asinex Diversity Library	5,072	-0.5	39	0.8%
		0	267	5.3%
		0.5	811	16.0%



## REFERENCES

1. Kaiser, J. Science resources. Chemists want NIH to curtail database. *Science* **2005**, 308, 774.
2. Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput. Aided Mol. Des* **2009**, 23, 195-198.
3. Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. Chemoinformatics in Drug Discovery . 2005. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, FRG.  
Ref Type: Electronic Citation
4. PDSP. <http://pdsp.med.unc.edu/> . 2010.  
Ref Type: Electronic Citation
5. Mandy, R. Gleevec: Highlighting the Power of Rational Drug Design. The Journal of Young Investigators 5[3]. 9-16-2005.  
Ref Type: Electronic Citation
6. Talete Dragon Descriptors. <http://www.talete.mi.it/index.htm> . 2011.  
Ref Type: Electronic Citation
7. eduSoft LC. MolconnZ. 2007.  
Ref Type: Computer Program
8. MOE. Chemical Computing Group [2007.09]. 2008.  
Ref Type: Electronic Citation
9. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, 42, 1273-1280.
10. Martin, Y. C. A practitioner's perspective of the role of quantitative structure-activity analysis in medicinal chemistry. *J. Med. Chem.* **1981**, 24, 229-237.
11. Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. Quantitative structure-activity relationship modeling of dopamine D(1) antagonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, 42, 3217-3226.

12. Tropsha, A.; Zheng, W. Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. *Curr. Pharm. Des* **2001**, *7*, 599-612.
13. Wang, J. X.; Dipasquale, A. J.; Bray, A. M.; Maeji, N. J.; Spellmeyer, D. C.; Geysen, H. M. Systematic study of substance P analogs. II. Rapid screening of 512 substance P stereoisomers for binding to NK1 receptor. *Int. J. Pept. Protein Res.* **1993**, *42*, 392-399.
14. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197-206.
15. Golbraikh, A.; Tropsha, A. Beware of  $q(2)!$  *Journal of Molecular Graphics & Modelling* **2002**, *20*, 269-276.
16. 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.
17. 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.
18. Hajjo, R.; Grulke, C. M.; Golbraikh, A.; Setola, V.; Huang, X. P.; Roth, B. L.; Tropsha, A. Development, validation, and use of quantitative structure-activity relationship models of 5-hydroxytryptamine (2B) receptor ligands to identify novel receptor binders and putative valvulopathic compounds among common drugs. *J. Med. Chem.* **2010**, *53*, 7573-7586.
19. PASS. <http://195.178.207.233/PASS/index.html> . 2008.  
Ref Type: Electronic Citation
20. Ajay A unified framework for using neural networks to build QSARs. *J. Med. Chem.* **1993**, *36*, 3565-3571.
21. Hirst, J. D.; King, R. D.; Sternberg, M. J. Quantitative structure-activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput. Aided Mol. Des* **1994**, *8*, 405-420.
22. Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of linear and nonlinear QSAR data using neural networks. *J. Med. Chem.* **1994**, *37*, 3758-3767.
23. Sternberg, M. J.; Lewis, R. A.; King, R. D.; Muggleton, S. Modelling the structure and function of enzymes by machine learning. *Faraday Discuss.* **1992**, 269-280.

24. Zhang, S.; Golbraikh, A.; Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J. Med. Chem.* **2006**, *49*, 2713-2724.
25. 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.
26. A.Liaw; M.Wiener. Classification and Regression by randomForest. R News 2[3], 18-22. 2002.  
Ref Type: Electronic Citation
27. Meyer, D.; Leisch, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55*, 169-186.
28. Zheng, W.; Cho, S. J.; Waller, C. L.; Tropsha, A. Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: a novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 738-746.
29. Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238-1244.
30. de Cerqueira, L. P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245-1254.
31. Han, L.; Wang, Y.; Bryant, S. H. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinformatics* **2008**, *9*, 401.
32. Bordas, B.; Komives, T.; Szanto, Z.; Lopata, A. Comparative three-dimensional quantitative structure-activity relationship study of safeners and herbicides. *J. Agric. Food Chem.* **2000**, *48*, 926-931.
33. Fan, Y.; Shi, L. M.; Kohn, K. W.; Pommier, Y.; Weinstein, J. N. Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *J. Med. Chem.* **2001**, *44*, 3254-3263.
34. Girones, X.; Gallegos, A.; Carbo-Dorca, R. Modeling antimalarial activity: application of Kinetic Energy Density Quantum Similarity Measures as descriptors in QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1400-1407.
35. Randic, M.; Basak, S. C. Construction of high-quality structure-property-activity regressions: the boiling points of sulfides. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 899-905.

36. Suzuki, T.; Timofei, S.; Iuoras, B. E.; Uray, G.; Verdino, P.; Fabian, W. M. Quantitative structure-enantioselective retention relationships for chromatographic separation of arylalkylcarbinols on Pirkle type chiral stationary phases. *J. Chromatogr. A* **2001**, 922, 13-23.
37. Moron, J. A.; Campillo, M.; Perez, V.; Unzeta, M.; Pardo, L. Molecular determinants of MAO selectivity in a series of indolylmethylamine derivatives: biological activities, 3D-QSAR/CoMFA analysis, and computational simulation of ligand recognition. *J. Med. Chem.* **2000**, 43, 1684-1691.
38. Recanatini, M.; Cavalli, A.; Belluti, F.; Piazzzi, L.; Rampa, A.; Bisi, A.; Gobbi, S.; Valenti, P.; Andrisano, V.; Bartolini, M.; Cavrini, V. SAR of 9-amino-1,2,3,4-tetrahydroacridine-based acetylcholinesterase inhibitors: synthesis, enzyme inhibitory activity, QSAR, and structure-based CoMFA of tacrine analogues. *J. Med. Chem.* **2000**, 43, 2007-2018.
39. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003**, 17, 241-253.
40. Dalpiaz, A.; Bertolasi, V.; Borea, P. A.; Nacci, V.; Fiorini, I.; Campiani, G.; Mennini, T.; Manzoni, C.; Novellino, E.; Greco, G. A concerted study using binding measurements, X-ray structural data, and molecular modeling on the stereochemical features responsible for the affinity of 6-arylpyrrolo[2,1-d][1,5]benzothiazepines toward mitochondrial benzodiazepine receptors. *J. Med. Chem.* **1995**, 38, 4730-4738.
41. Kubinyi, H. Quantitative Structure-Activity Relationships - 12th European Symposium Molecular modeling and prediction of bioactivity. *IDrugs*. **1998**, 1, 781-786.
42. Norinder, U.; Rivera, C.; Unden, A. A quantitative structure-activity relationship study of some substance P-related peptides. A multivariate approach using PLS and variable selection. *J. Pept. Res.* **1997**, 49, 155-162.
43. Zefirov, N. S.; Palyulin, V. A. QSAR for boiling points of "small" sulfides. Are the "high-quality structure-property-activity regressions" the real high quality QSAR models? *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1022-1027.
44. Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002**, 5, 231-243.
45. Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003**, 46, 3013-3020.

46. 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.
47. Tropsha, A. Recent Trends in Quantitative Structure-Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*; Abraham, D. Ed.; John Wiley & Sons, Inc.: New York, 2003; pp 49-77.
48. Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Qsar & Combinatorial Science* **2003**, *22*, 69-77.
49. Breiman L. Random forests. *Machine Learning* *41*, 5-32. 2011.  
Ref Type: Electronic Citation
50. Ho, T. K. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. *Pattern Analysis and Applications* [5], 102-112. 2002.  
Ref Type: Electronic Citation
51. Random Forest. [http://en.wikipedia.org/wiki/Random\\_forest](http://en.wikipedia.org/wiki/Random_forest) . 2010.  
Ref Type: Electronic Citation
52. Random Forest. [http://en.wikipedia.org/wiki/Random\\_forest](http://en.wikipedia.org/wiki/Random_forest) . 2010.  
Ref Type: Electronic Citation
53. A.Liaw; M.Wiener. Classification and Regression by randomForest. *R News* 2[3], 18-22. 2002.  
Ref Type: Electronic Citation
54. Vapnik, V. *The nature of statistical learning theory*; Springer-Verlag: New York, 1995.
55. Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J Med. Chem.* **2005**, *48*, 7322-7332.
56. Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design (Methods and Principles in Medicinal Chemistry, Vol 2)*; Waterbeemd, H. v. d. Ed.; Wiley-VCH Verlag GmbH: Weinheim (Germany), 1995; pp 309-318.
57. Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462*, 175-181.

58. PASS. <http://195.178.207.233/PASS/index.html> . 2008.  
Ref Type: Electronic Citation
59. Poroikov, V.; Akimov, D.; Shabelnikova, E.; Filimonov, D. Top 200 medicines: can new actions be discovered through computer-aided prediction? *SAR QSAR Environ. Res.* **2001**, *12*, 327-344.
60. MLI. <http://mli.nih.gov/> . 2006.  
Ref Type: Electronic Citation
61. Schreiber, S. L.; Nicolaou, K. C.; Davies, K. Diversity-oriented organic synthesis and proteomics. New frontiers for chemistry & biology. *Chem. Biol.* **2002**, *9*, 1-2.
62. Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846-854.
63. Dealing with a data dilemma. *Nat. Rev. Drug Discov.* **2008**, *7*, 632-633.
64. PDSP. <http://pdsp.med.unc.edu/> . 2010.  
Ref Type: Electronic Citation
65. Shapiro, D. A.; Renock, S.; Arrington, E.; Chiodo, L. A.; Liu, L. X.; Sibley, D. R.; Roth, B. L.; Mailman, R. Aripiprazole, a novel atypical antipsychotic drug with a unique and robust pharmacology. *Neuropsychopharmacology* **2003**, *28*, 1400-1411.
66. Roth, B. L.; Baner, K.; Westkaemper, R.; Siebert, D.; Rice, K. C.; Steinberg, S.; Ernsberger, P.; Rothman, R. B. Salvinorin A: A potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 11934-11939.
67. Olah M; Mracec M; Ostopovici L; et al. WOMBAT: world of molecular bioactivity, in Chemoinformatics in Drug Discovery. New York: Wiley-VCH [Oprea TI edition], 223-239. 2004.  
Ref Type: Electronic Citation
68. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31-36.
69. Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press: New York, 1976.
70. Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; Wiley: New York, 1986.
71. Randi, M. On Characterization on Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

72. Kier, L. B. A shape index from molecular graphs. *Quant. Struct. -Act. Relat.* **1985**, *4*, 109-116.
73. Kier, L. B. Inclusion of symmetry as a shape attribute in kappa-index analysis. *Quant. Struct-Act. Relat.* **1987**, *6*, 8-12.
74. Kier, L. B.; Hall, L. H. An Electrotopolological State Index for Atoms in Molecules. *Pharmaceutical Res.* **1990**, *7*, 801.
75. Kier, L. B.; Hall, L. H. An Index of Electrotopolological State of Atoms in Molecules. *J. Math. Chem.* **1991**, *7*, 229.
76. Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopolological State*; Academic Press: 1999.
77. 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.
78. 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.
79. Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494-3504.
80. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003**, *17*, 241-253.
81. Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002**, *5*, 231-243.
82. Malone, D. L.; Genuit, T.; Tracy, J. K.; Gannon, C.; Napolitano, L. M. Surgical site infections: reanalysis of risk factors. *J. Surg. Res.* **2002**, *103*, 89-95.
83. Singleton, S. F.; Roca, A. I.; Lee, A. M.; Xiao, J. Probing the structure of RecA-DNA filaments. Advantages of a fluorescent guanine analog. *Tetrahedron* **2007**, *63*, 3553-3566.
84. Lee, A. M.; Ross, C. T.; Zeng, B. B.; Singleton, S. F. A molecular target for suppression of the evolution of antibiotic resistance: inhibition of the Escherichia coli RecA protein by N(6)-(1-naphthyl)-ADP. *J. Med. Chem.* **2005**, *48*, 5408-5411.

85. Wigle, T. J.; Sexton, J. Z.; Gromova, A. V.; Hadimani, M. B.; Hughes, M. A.; Smith, G. R.; Yeh, L. A.; Singleton, S. F. Inhibitors of RecA Activity Discovered by High-Throughput Screening: Cell-Permeable Small Molecules Attenuate the SOS Response in Escherichia Coli. *J. Biomol. Screen.* **2009**.
86. Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177-182.
87. World Drug Index (WDI). <http://www.daylight.com/products/wdi.html> . 2007.  
Ref Type: Electronic Citation
88. Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002**, *5*, 231-243.
89. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003**, *17*, 241-253.
90. Hajjo, R.; Grulke, C. M.; Golbraikh, A.; Setola, V.; Huang, X. P.; Roth, B. L.; Tropsha, A. Development, validation, and use of quantitative structure-activity relationship models of 5-hydroxytryptamine (2B) receptor ligands to identify novel receptor binders and putative valvulopathic compounds among common drugs. *J. Med. Chem.* **2010**, *53*, 7573-7586.
91. 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.
92. ChemAxon. *JChem.* 2010.  
Ref Type: Computer Program
93. Hajjo, R.; Grulke, C. M.; Golbraikh, A.; Setola, V.; Huang, X. P.; Roth, B. L.; Tropsha, A. Development, validation, and use of quantitative structure-activity relationship models of 5-hydroxytryptamine (2B) receptor ligands to identify novel receptor binders and putative valvulopathic compounds among common drugs. *J. Med. Chem.* **2010**, *53*, 7573-7586.
94. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* **2003**, *17*, 241-253.
95. Prestwick Chemical Library. Prestwick Chemical . 1-26-2011.  
Ref Type: Electronic Citation
96. World Drug Index. Thomson Reuters . 2011.  
Ref Type: Electronic Citation



97. TimTec ActiTarg-G Library. TimTec . 2011.  
Ref Type: Electronic Citation
98. ASINEX GPCR-targeted library. ASINEX . 2011.  
Ref Type: Electronic Citation
99. TimTec Diversity Screening Library. TimTec . 2011.  
Ref Type: Electronic Citation
100. ASINEX Diversity Set. ASINEX . 2011.  
Ref Type: Electronic Citation
101. 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.
102. MACCS Structural Keys. MDL Ltd.:San Leandro, CA . 1992.  
Ref Type: Electronic Citation
103. Lysergol Information on ChemSpider. ChemSpider . 2011.  
Ref Type: Electronic Citation
104. LogP prediction. ACD/Labs . 2011.  
Ref Type: Electronic Citation
105. Lysergol Information on ChemSpider. ChemSpider . 2011.  
Ref Type: Electronic Citation
106. EPI-Suite Introduction. EPA . 2011.  
Ref Type: Electronic Citation
107. Jorgensen, W. L.; Tirado-Rives, J. QSAR/QSPR and Proprietary Data. *J Chem. Inf. Model.* **2006**, *46*, 937.
108. Oprea, T. I.; Tropsha, A.; Faulon, J. L.; Rintoul, M. D. Systems chemical biology. *Nat. Chem. Biol.* **2007**, *3*, 447-450.