

***IN SILICO* STRATEGIES TO STUDY  
POLYPHARMACOLOGY OF G-PROTEIN-COUPLED  
RECEPTORS**

Rima Hajjo

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill  
in partial fulfillment of the requirements for the degree of Doctor of Philosophy  
in the UNC Eshelman School of Pharmacy  
(Division of Medicinal Chemistry and Natural Products).

Chapel Hill  
2010

Approved by:

Alexander Tropsha, PhD

Bryan L. Roth, MD, PhD

Rihe Liu, PhD

David Lawrence, PhD

Weifan Zheng, PhD

Xiang Wang, PhD

©2010  
Rima Hajjo  
ALL RIGHTS RESERVED

## ABSTRACT

RIMA HAJJO: *In Silico* Strategies to Study Polypharmacology of G-Protein-Coupled Receptors  
(Under the direction of Alexander Tropsha)

The development of drugs that simultaneously target multiple receptors in a rational way (i.e., ‘magic shotguns’) is regarded as a promising approach for drug discovery to treat complex, multi-factorial and multi-pathogenic diseases. My major goal is to develop and employ different computational approaches towards the rational design of drugs with selective polypharmacology towards guanine nucleotide-binding protein (G-protein)-coupled receptors (GPCRs) to treat central nervous system diseases. Our methodologies rely on the advances in chemocentric informatics and chemogenomics to generate experimentally testable hypotheses that are derived by fusing independent lines of evidence. We posit that such hypothesis fusion approach allows us to improve the overall success rates of *in silico* lead identification efforts. We have developed an integrated computational approach that combines Quantitative Structure-Activity Relationships (QSAR) modeling, model-based virtual screening (VS), gene expression analysis and mining of the biological literature for drug discovery.

The dissertation research described herein is focused on: (1) The development of robust data-driven Quantitative Structure-Activity Relationship (QSAR) models of single target GPCR datasets that will amount to the compendium of GPCR predictors: the GPCR QSARome; (2) The development of robust data-driven QSAR models for families of GPCRs

and other trans-membrane molecular targets (i.e., sigma receptors) and the application of models as virtual screening tools for the quick prioritization of compounds for biological testing across receptor families; (3) The development of novel integrative chemocentric informatics approaches to predict receptor-mediated clinical effects of chemicals. Results indicated that our computational efforts to establish a compendium of computational predictors and devise an integrative chemocentric informatics approach to study polypharmacology *in silico* will eventually lead to useful and reliable tools aimed at identifying and enriching chemical libraries with compounds that have the desired activities for more than one molecular target of interest.



To mom, dad, Mahmoud, and Ghassan:

The dearest people to my heart and soul, they are the whole World to me, they are my family,  
my friends, and everything pleasant.

To Ali and Iman:

My little angels, God's best gift to me, the greatest source for my happiness, hope and  
determination.

And in the memory of my beloved grandfather (Ibrahim Hajjo):

A person whom I truly loved from the bottom of my heart! A person whose love,  
compassion, and kindness can never be forgotten.

## ACKNOWLEDGEMENTS

Over the few years spent in this great country I was blessed and fortunate to get to know the most amazing people one could find in a lifetime. These people were a major source for my happiness, hard work and faith in a better tomorrow. I wish I can thank them here in my simple words:

I am very grateful to my advisor Dr. Alexander Tropsha. I am thankful for his efforts in recruiting me to UNC and for accepting me among his great group. I was fortunate enough to join his group and learn directly from him after I read much about his work while I was working at the University of Jordan. However, I could have never succeeded in this field without the great help, support and encouragement I got from Alex. I also owe him a lot of what I became today and of what I learned about science, life, culture, religion, politics, and not to forget Russian proverbs! One of my favorite was and will always be: “Train hard, fight easy...and win; Train easy, fight hard...and die.” He was really always encouraging me and inspiring me with that!

I am grateful for Dr. KH Lee for his cooperation in recruiting me and giving me the chance to join his lab. I also thank all the members of the Lee’s group for their nice treatment and friendship, especially to Dr. Qian Shi for her enormous help with organic chemistry issues that I encountered during the course of my studies.

I am grateful to Dr. Bryan Roth for his invaluable advice, critique and continuous help with testing our computational hypotheses through the PDSP, and later in analyzing the results. Dr. Roth’s distinguished knowledge and expertise in the area of molecular pharmacology and neuroscience has been a great asset for my research. I am also thankful to all PDSP members, especially, Dr. Vincent Setola, Dr. XP Huang and Jon Evans.

I am thankful for Dr. Justin Lamb from the Broad Institute for his continuous advice. I also thank him for offering me the opportunity to join his group.

I also thank my committee members: Dr. Rihe Liu, Dr. David Lawrence, Dr. Weifan Zheng and Dr. Xiang Wang. I had a great time rotating in Dr. Liu's lab and learning directly from him. I always had many questions and he's been always glad to answer and explain things to me. I learned a great deal about available biotechnologies. I am also very thankful to Steve and Alex from his group for the great help and pleasant times in the Liu's lab. I also had a great experience to learn directly from Weifan's wisdom and good planning. Weifan provided me with great advice how to plan my work and succeed in completing it on a timely matter with all the difficulties that surrounded me. I owe Weifan a lot and he's always going to be one of my favorite people. I also thank Simon for his help and advice on many computational issues and for being a good friend of mine who can take all my jokes! I really miss his presence these days at GMB but I wish him all the best in his new faculty position.

I am very grateful to Chris Grulke for his incomparable help and support. I really have no words to thank him enough. I owe him a lot of what I learned in this lab. He has also been one of my best friends that I can talk freely with about all issues from science, and programming languages to philosophy and religion.

I am also grateful to Dr. Alexander Golbraikh for his enormous help and hands-on training. I thank him for all the time and effort he spent with me to help me excel in my computational skills. He's an amazing mathematician but I never got all the things he was talking about!

Many thanks also for Dr. Denis Fourches for his help and advice on many scientific and personal issues. I learned a lot from him about science, life, culture and much more. I really value his wisdom and honest opinions. His presence at GMB was really missed!

I am also thankful to all my friends the MML-ers: Jui-Hua, Tang, Tanarat, Aleck, Hao, Clark, Stephen, Jon, Liying, Ashutosh, Mihir, Nancy, Paula, Man Luo, Guiyu, Theo, Tony, Eugene, Denise, Andy and Kun. Also Many Thanks for Daniel Samarov for his great help with statistical analysis for genomic data, I really appreciated that. I also can't forget here two of my favorite: Berk and Dr. Karthikeyan; I am especially thankful for Berk for teaching me a lot about computers and nice design and for Karthik to keep reminding me about publications!

I also want to thank the nicest people I met here who took by my hands, helped me adjust to my new life here and later became my second family and my best friends: Diane, Phillip, Janis, Raed. I also can't thank enough one of the nicest people I have gotten to know; whom then became my sister friend "Reema". I was blessed to find her here beside me all the time. Life would have been so boring and rough without having her and her family in our lives. I love her from the bottom of my heart and I wish her and her family all the best in this World.

I am also grateful to my home university, the University of Jordan, for the PhD scholarship they provided me with. I am very thankful my MSc advisor, Dr Fatima Afifi for everything I learned from her, I thank her for being a role model how a mom and a scientist could be. I also thank Dr. Ali Qaisi for being such a great dean for the School of Pharmacy.

And also thank TransTech Phamra for the internship opportunity and for offering me a job after I graduate. I am especially thankful to Dr. Robert Andrews, Dr. Mohan Rao, and



Dr. Nidhi Singh. I thank them for everything I learned from them and the pleasant times I spent with them. I can't wait to start working with you.

My biggest thanks to my family who made this dream come true. I feel I ran out of words. I can't find accurate words to describe their sacrifice, love and compassion. I can't find fair words to thank them enough for their enormous love and support. I am really nothing without you; I miss you all so much and hope to be back home soon. Greatest thanks to my greatest loving mom and the kindest husband ever; you both made it happen!

I am also grateful to my most beloved ones, Ali and Iman. They filled my life with joy and happiness the moment they arrived here and lived with me. I hope the future is holding the best for us. I hope that the coming days and years will allow me to give you much more. I love more than anything else in the entire World.

I thank God for all of this and all the blessings he's given me in my life.

## TABLE OF CONTENTS

ABSTRACT.....	iii
DEDICATION.....	v
ACKNOWLEDGEMENTS.....	vi
LIST OF TABLES.....	xv
LIST OF FIGURES.....	xviii
ABBREVIATIONS.....	xx
Chapter	
I. INTRODUCTION.....	1
Drugs Acting at Multiple Targets and Selective Polypharmacology as an Important Drug Discover Approach.....	1
G Protein-Coupled Receptors as Molecular Targets to Study Polypharmacology.....	3
Quantitative Structure Activity Relationships Modeling.....	4
In silico Receptoromics to Study Polypharmacology.....	5
Chemocentric Informatics.....	9
‘Omics’ Data Types and Repositories.....	11
The Multidimensional Chemocentric Space.....	14
Hypothesis Fusion.....	16
Overview of Chapter 2: Materials and Methods.....	16
Overview of Chapter 3: <i>In silico</i> Receptoromics.....	17
Overview of Chapter 4: The 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.....	18

Overview of Chapter 5: An Integrative Chemocentric Informatics Approach to Drug Discovery .....	20
Overview of Chapter 6: Conclusions and Future Studies.....	22
<b>II. MATERIALS AND METHODS.....</b>	<b>23</b>
Databases and Datasets.....	23
Dataset Curation.....	25
Computational Methods.....	29
Molecular Descriptors.....	31
Machine Learning Methods.....	34
Balancing Datasets Using Similarity Searching.....	40
Model Selection and Validation.....	40
Virtual Screening.....	44
Applicability Domain.....	45
Consensus Prediction.....	46
Integrative Chemocentric Informatics Approach.....	46
Experimental Methods.....	47
<b>III. IN SILICO RECEPTOROMICS: QSAR MODELING OF RECEPTOR SUBTYPES AND FAMILIES, MODEL APPLICATION FOR VIRTUAL SCREENING, AND EXPERIMENTAL VALIDATION.....</b>	<b>49</b>
Introduction.....	49
Materials and Methods.....	53
Databases and Datasets.....	53
Preprocessing of the Datasets.....	53
Dataset Division for Model Building and Validation.....	54
Computational Methods.....	56

Applicability Domain.....	57
Robustness of QSAR Models.....	58
Virtual Screening and Consensus Prediction Thresholds.....	59
Experimental Validation in Radiologand Binding Assays.....	61
Results and Discussion.....	61
QSAR Modeling for Receptor Subtypes.....	61
QSAR Modeling to Discriminate Actives vs. Inactives for individual Receptor Families.....	64
Descriptor Analysis.....	75
Virtual Screening and Experimental Validation.....	89
Biological Relevance.....	96
Conclusions.....	98
<b>IV. THE 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.....</b>	<b>100</b>
Introduction.....	100
Materials and Methods.....	107
Dataset.....	107
Preprocessing of the Dataset.....	108
Dataset Division for Model Building and Validation.....	108
Computational Methods.....	109
Molecular Descriptors.....	110
Balancing the Dataset Using Similarity Searching.....	111

QSAR Methods.....	112
Robustness of QSAR Models.....	112
Applicability Domains of <i>k</i> NN QSAR Models.....	113
Consensus Prediction.....	114
Virtual Screening and Compound Selection for Experimental Validation.....	115
Results and Discussion.....	115
Combinatorial QSAR Modeling of 5-HT <sub>2B</sub> Actives vs. Inactives.....	115
Comparison of Binary QSAR Approaches for Classifying 5-HT <sub>2B</sub> Actives vs. Inactives.....	121
Model Validation by Predicting Drugs Known to be 5-HT <sub>2B</sub> Actives and Valvulopathogens.....	123
Model Validation by Predicting an External Set.....	124
The Importance of Variable Selection.....	125
Virtual Screening of the World Drug Index Database to Identify Putative 5-HT <sub>2B</sub> Ligands.....	128
Experimental Validation.....	132
Conclusions.....	146
<b>V. AN INTEGRATIVE CHEMOCENTRIC INFORMATICS APPROACH TO DRUG DISCOVERY BASED ON STRUCTURAL HYPOTHESIS FUSION: IDENTIFICATION AND EXPERIMENTAL VALIDATION OF SELECTIVE ESTROGEN RECEPTOR MODULATORS AS LIGANDS OF 5-HYDROXYTRYPTAMINE-6 RECEPTORS.....</b>	<b>149</b>
Introduction.....	149
Materials and Methods.....	153
Integrative Chemocentric Informatics Approach.....	153
Databases and Datasets.....	158
Computational Methods.....	158

(1) QSAR Modeling and QSAR-based Virtual Screening.....	158
(2) Biological Network Mining.....	164
(3) Hypothesis Fusion.....	166
Experimental Validation in Radiologand Binding Assays.....	166
Results and Discussion.....	167
QSAR Modeling of 5-HT <sub>6</sub> R Binders and Non-Binders.....	168
QSAR-Based Virtual Screening.....	170
Searching the Connectivity Map for Potential Anti-Alzheimer's Agents.....	173
Hypothesis Generation: Integrating and Fusing Independent Hypotheses from QSAR-Based VS and cmap Analysis.....	179
Scoring and Fusing Structural Hypotheses .....	180
Hypothesis Testing: Evaluation of Computational Hits at Human Cloned 5-HT <sub>6</sub> Receptors.....	189
SERMs Identified as 5-HT <sub>6</sub> R Ligands.....	199
Raloxifene Identified as a 5-HT <sub>6</sub> R Ligand and Agent with Potential Utility in Alzheimer's Disease.....	200
Predict and Validate Polypharmacology of SERMs.....	201
Conclusions.....	212
VI. FUTURE DIRECTIONS.....	214
Summary.....	214
<i>In silico</i> Receptoromics.....	216
Chemocentric Informatics Approach.....	219
Conclusion.....	220
APPENDICES.....	221
REFERENCES.....	229

## LIST OF TABLES

Table		
2.1	Confusion matrix for binary classification model.....	43
3.1	QSAR Models for Selected GPCRs.....	63
3.2	Performance of <i>k</i> NN classification methods to classify actives <i>vs.</i> inactives across six receptor families.....	65
3.3	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across 5-HT receptor family.....	67
3.4	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across adrenergic alpha receptor family .....	68
3.5	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across dopamine receptor family.....	69
3.6	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across histamine receptor family.....	70
3.7	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across muscarinic receptor family.....	71
3.8	Performance of SVM classification methods to classify actives <i>vs.</i> inactives across Sigma receptor family.....	72
3.9	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify binders <i>vs.</i> non-binders against the 5-HT receptor family.....	78
3.10	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify binders <i>vs.</i> non-binders against the adrenergic alpha receptor family.....	80
3.11	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify binders <i>vs.</i> non-binders against the dopamine receptor family.....	81
3.12	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify binders <i>vs.</i> non-binders against the histamine receptor family.....	83
3.13	Top twenty most frequently used Dragon descriptors in	

	validated <i>k</i> NN-Dragon models to classify binders vs. non-binders against the muscarinic receptor family.....	85
3.14	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify binders vs. non-binders against the sigma receptor family.....	87
3.15	Eleven VS hits selected from WDI chemical library.....	91
3.16	Experimental validation results for the 11 computational hits predicted to be ligands to six families of receptors as a result of QSAR-based mining of the WDI chemical screening library.....	93
4.1	Chemical structures of marketed drugs known as 5-HT <sub>2B</sub> receptor agonists and associated with VHD.....	102
4.2	Performance of <i>k</i> NN classification methods to classify actives vs. inactives based on external validation set statistics.....	117
4.3	Top twenty highly weighted Dragon descriptors by DWD for 5-HT <sub>2B</sub> actives vs. inactives.....	120
4.4	Comparison between different <i>k</i> NN-Dragon QSAR models generated with or without variable selection.....	126
4.5	Top twenty most frequently used Dragon descriptors in validated <i>k</i> NN-Dragon models to classify 5-HT <sub>2B</sub> actives vs. inactives.....	127
4.6	Experimental validation results for the ten computational hits predicted as 5-HT <sub>2B</sub> ligands as a result of QSAR-based mining of the WDI chemical screening library.....	133
4.7	Virtual screening recovery results using Tanimoto similarities and 166 MACCS keys.....	139
4.8	Virtual screening recovery results using Euclidean distances and 903 Dragon descriptors.....	141
4.9	Nearest neighbor compounds from the active compounds in the dataset and the ten experimentally validated VS hits.....	143
5.1	Top twenty negative connections from the cmap with Alzheimer's disease gene signature S1.....	177
5.2	Top twenty negative connections from the cmap with Alzheimer's disease gene signature S2.....	178



5.3	Final thirty nine computational hits from QSAR-based VS and cmap.....	184
5.4	Therapeutic classes of the thirty nine final computational hits from QSAR-based VS and cmap.....	185
5.5	Experimental validation results for the thirteen computational hits predicted as 5-HT <sub>6</sub> R ligands and had negative connections with Alzheimer's disease gene signatures.....	191
5.6	The significance of the tested hits in relation to cognition, neuroprotection and anti-Alzheimer's effects.....	198
5.7	Predicting polypharmacology of SERMs using receptor family-based QSAR.....	202
5.8	Tanimoto similarities between SERMs based on MACCS structural keys.....	204
5.9	<i>K<sub>i</sub></i> estimates for SERMs (i.e., clomiphene, raloxifene, tamoxifen and toremifene) at a large panel of cloned receptors.....	209

## LIST OF FIGURES

Figure		
1.1	Concept of QSAR-based in silico receptoromics enabling a virtual network pharmacology approach.....	8
1.2	The connectivity map concept.....	13
1.3	Chemocentric informatics multidimensional space .....	15
2.1	General workflow of chemical dataset curation.....	27
2.2	Flowchart of predictive QSAR modeling framework based on validated QSAR models.....	30
3.1	Comparison of CCR values for the external validation set for different QSAR models developed to classify binders vs. non-binders across six receptor families.....	74
3.2	The heatmap of descriptor frequencies across receptor families.....	77
3.3	The heatmap of <i>k</i> NN CS WDI compounds across six receptor Families.....	97
4.1	The workflow for QSAR model building and validation as applied to the 5-HT <sub>2B</sub> dataset.....	110
4.2	Comparison of CCR values for the external validation set for different QSAR models developed to classify actives vs. inactives.....	122
4.3	Steps of the virtual screening of the WDI database to identify putative 5-HT <sub>2B</sub> ligands.....	131
4.4	Competition binding at 5-HT <sub>2B</sub> receptors.....	135
4.5	The heatmap of the Tanimoto distances between the 146 5-HT <sub>2B</sub> actives WDI compounds.....	140
4.6	The heatmap of the Euclidean distances between the 146 5-HT <sub>2B</sub> actives WDI compounds.....	142
5.1	Study design for the integrative informatics approach .....	155
5.2	The workflow for QSAR model building, validation and virtual screening as applied to 5-HT <sub>6R</sub> dataset.....	161

5.3	Comparison of the QSAR approaches to classify 5-HT <sub>6</sub> R binders vs. non-binders based on CCR <sub>evs</sub> .....	169
5.4	A representation for QSAR-based virtual screening steps of two chemical databases.....	172
5.5	Querying the connectivity map with Alzheimer's disease gene signatures.....	174
5.6	The workflow for fusing hypotheses from QSAR modeling and cmap negative connections.....	181
5.7	Plots for <i>k</i> NN scores vs. cmap connectivity scores for 39 final common hits from QSAR-based VS and map.....	187
5.8	Competition binding isotherms at 5-HT <sub>6</sub> R for several predicted actives.....	195
5.9	Chemical protein interaction networks for SERMs.....	205
5.10	The heatmap of binding affinities ( <i>K<sub>i</sub></i> ) for several SERMs across a panel of GPCRs and other transmembrane molecular targets.....	211

## ABBREVIATIONS

5-HT <sub>2B</sub>	5-Hydroxytryptamine subtype 2B receptors
5-HT <sub>6R</sub>	5-Hydroxytryptamine-6 receptors
AD	Applicability Domain
C <sub>max</sub>	Maximum Plasma Concentration
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
CCR <sub>rand</sub>	Correct Classification Rate of the models generated with randomized activities of training set compounds and using the external validation set
CV	Cross Validation
DWD	Distance Weighted Discrimination
E	Enrichment
En	Normalized Enrichment
ER	Estrogen Receptors
FDA	U.S. Food and Drug Administration
FN	False Negative
FP	False Positive
GPCR	G Protein-Coupled Receptors

HTS	High Throughput Screen
<i>k</i> NN	<i>k</i> Nearest Neighbor
LOO-CV	Leave-One-Out Cross Validation
MFD	Most Frequent Descriptors
MOE	Molecular Operating Environment
MORE	Multiple Outcomes of Raloxifene Evaluation
MZ	MolConnZ descriptors
PDSP	NIMH Psychoactive Drug Screening Program
QSAR	Quantitative Structure Activity Relationships
SA	Simulated Annealing
SE	Sensitivity
SERM	Selective Estrogen Receptor Modulator
SG	Subgraph
SP	Specificity
T <sub>c</sub>	Tanimoto Coefficient
TP	True Positive
TN	True Negative
VHD	Valvular Heart Disease
VS	Virtual Screening
WDI	World Drug Index

# CHAPTER 1

## INTRODUCTION

### **Drugs Acting at Multiple Targets and Selective Polypharmacology as an Important Drug Discovery Approach**

Over the past decade, there has been a decline in the number of new drugs reaching the market and being used as effective therapeutics. Many reasons have been suggested to explain this decline in drug development productivity (Bakker et al. 1999, Cascante et al. 2002, Hood & Perlmutter 2004, van der Greef & McBurney 2005). This led to the suggestion that there may be real issues with the core assumptions that framed drug discovery approaches in the past two decades (Hopkins 2008). One of the key goals of the rational drug design has been the discovery of very selective ligands acting via individual molecular targets. In fact, a highly selective ligand for a given target does not always result in clinically efficacious drug because of the redundancy in our biological systems (Frantz 2005, Mencher & Wang 2005). Therefore, this one-molecule, one-target approach, which has led to the discovery of many blockbuster drugs and will probably remain popular for many years, might not be suitable for treating complex multifactorial diseases such as neuropsychiatric and neurodegenerative diseases.

Polypharmacology (Frantz 2005, Hampton 2004, Keith et al. 2005, Mencher & Wang 2005, Roth & Kroeze 2006, Wermuth 2004) or the selective promiscuous modulation of several molecular targets has been proposed as a promising approach for drug discovery to treat complex diseases. Recent studies provided a substantial evidence that compound promiscuity is the main reason for the great efficacy of a significant number of approved drugs (Hampton 2004, Hopkins et al. 2006, Keith et al. 2005, Mencher & Wang 2005, Roth et al. 2004). The growing understanding of the complexity of biological networks and the robustness and redundancy of biological systems challenges the current approaches of single target drug discovery including *in silico* approaches (Hopkins 2007, Hopkins 2008, Roth et al. 2000, Roth & Kroeze 2006). Nowadays, medicinal chemists are becoming more interested in identifying polypharmacological drugs (i.e., ‘magic shotguns’) (Armbruster & Roth 2005, Roth et al. 2004) that can bind moderately to several targets in a disease-protein network and affect the overall outcome significantly (Hopkins et al. 2006, Hopkins 2009, Roth et al. 2004).

Thus, the main goal of polypharmacology is to identify a compound with a desired biological profile across multiple targets whose combined modulation will perturb a disease state (Hopkins 2008). The history of drug development of antipsychotics is a clear example of the migration from single target approach centered on dopamine D<sub>2</sub> to a multiple target strategy involving D<sub>2</sub> blockade and the recruitment of a wide array of other receptor activities. The recent generation of antipsychotics which are called “atypical”, approved by the U.S Food and Drug Administration (FDA) for use in the treatment of schizophrenia, acute mania, bipolar mania, psychotic agitation, bipolar maintenance, and other indications,

have interactions with dopaminergic, serotonergic, histaminergic, cholinergic and adrenergic receptors (Roth et al. 2004).

However, this approach faces many challenges such as the need to process, understand, and utilize the available information about drug interactions with multiple biological targets. Currently, most of the successful polypharmacological drugs on the market have been discovered by serendipity (Roth et al. 2004). Currently, there is no systematic strategy to design and optimize multi-target drugs in traditional drug discovery approaches. Thus, the integration of *in silico* methods, combined with ligand biological profiles against protein assays and gene expression arrays, can provide researchers with a novel toolbox to assess polypharmacology (Cavalli et al. 2008).

### **G Protein-Coupled Receptors as Molecular Targets to Study Polypharmacology**

G Protein-Coupled Receptors (GPCRs) are promising targets for the discovery of novel drugs. They constitute the largest family of membrane proteins that mediate most cellular responses to hormones and neurotransmitters and are also responsible for vision, olfaction and taste (Rosenbaum et al. 2009). The entire family of currently known and verified human GPCRs includes at least 799 unique full-length members (Gloriam et al. 2007). GPCRs are involved in a multitude of biological responses in all organs and systems including the central and peripheral nervous systems. In the latter, particularly important functions include neurotransmitter release, cell-to-cell communication, modulation of learning and memory, response to psycho-active substances, regulation of neuronal growth and differentiation and of glial responses. However, ligands for GPCRs comprise structurally very different compounds and often the ligands interact with more than one GPCR, i.e., they



are promiscuous. Thus, this group of receptors was chosen to study polypharmacology and to fulfill the research aims of my thesis projects.

GPCRs present a wide range of opportunities as therapeutic targets in areas including cancer, cardiac dysfunction, diabetes, central nervous system disorders, obesity, inflammation, and pain. Consequently, GPCRs are major components of pipelines in small and large pharmaceutical companies, and many drug discovery projects in academia and industry focus exclusively on these receptors. But the path to novel GPCR-targeted medicines is not routine. Most GPCR-modulating drugs on the market weren't initially targeted to a specific protein but were developed on the basis of functional activity observed in an assay. That they activated or inhibited a GPCR specifically was only later discovered. Post- Human Genome Project, however, targets are the starting points for most drug discovery endeavors. And there is still much to be learned about how GPCRs work and how they can be selectively modulated.

### **Quantitative Structure Activity Relationships Modeling**

QSAR studies rely heavily upon statistics to derive models that relate the biological activity of a series of compounds to one or more molecular properties that can be easily measured or calculated. Modern QSAR approaches are characterized by the use of multiple descriptors of chemical structure combined with the application of both linear and non-linear optimization approaches, and a strong emphasis on rigorous model validation to afford robust and predictive QSAR models. The goal of QSAR modeling is to establish a trend in the descriptor values, which parallels the trend in biological activity. All QSAR approaches imply, directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures

are expected to have similar biological activities. However the definition of similarity is not always simple as it depends on descriptors, variable selection, and similarity functions.

The differences in various QSAR methodologies can be understood in terms of the types of target property values, descriptors, and optimization algorithms used to relate descriptors to the target properties and generate statistically significant models. Target properties can generally be of three types: continuous (i.e., real values covering certain range, e.g., IC<sub>50</sub> values, or binding constants), categorical related (i.e., classes of target properties covering certain range of values, e.g., active and inactive compounds, frequently encoded numerically for the purpose of the subsequent analysis as one (for active) or zero (for inactive)), and categorical unrelated (i.e., classes of target properties that do not relate to each other in any continuum, e.g., compounds that belong to different pharmacological classes, or compounds that are classified as drugs vs. non-drugs). The corresponding methods of data analysis are referred to as classification or continuous property QSAR. The examples of leading QSAR approaches, their applications, and developing trends in the field can be found in recent reviews (Fan et al. 2001, Girones et al. 2000, Randic & Basak 2000).

### ***In silico* Receptoromics to Study Polypharmacology**

One way to identify polypharmacological chemical compounds is to virtually screen all potential molecular targets of interest for interactions with these chemicals (Roth 2005). Here in, we suggest that one approach to enable virtual screening of the receptorome would be to generate a compendium of computational predictors (e.g., QSAR models or structure based models). Subsequently, these models can be used for the virtual screening of chemical libraries to identify new ligands for the different molecular targets and predict polypharmacological matrices for these all chemicals included in these databases. Ultimately,

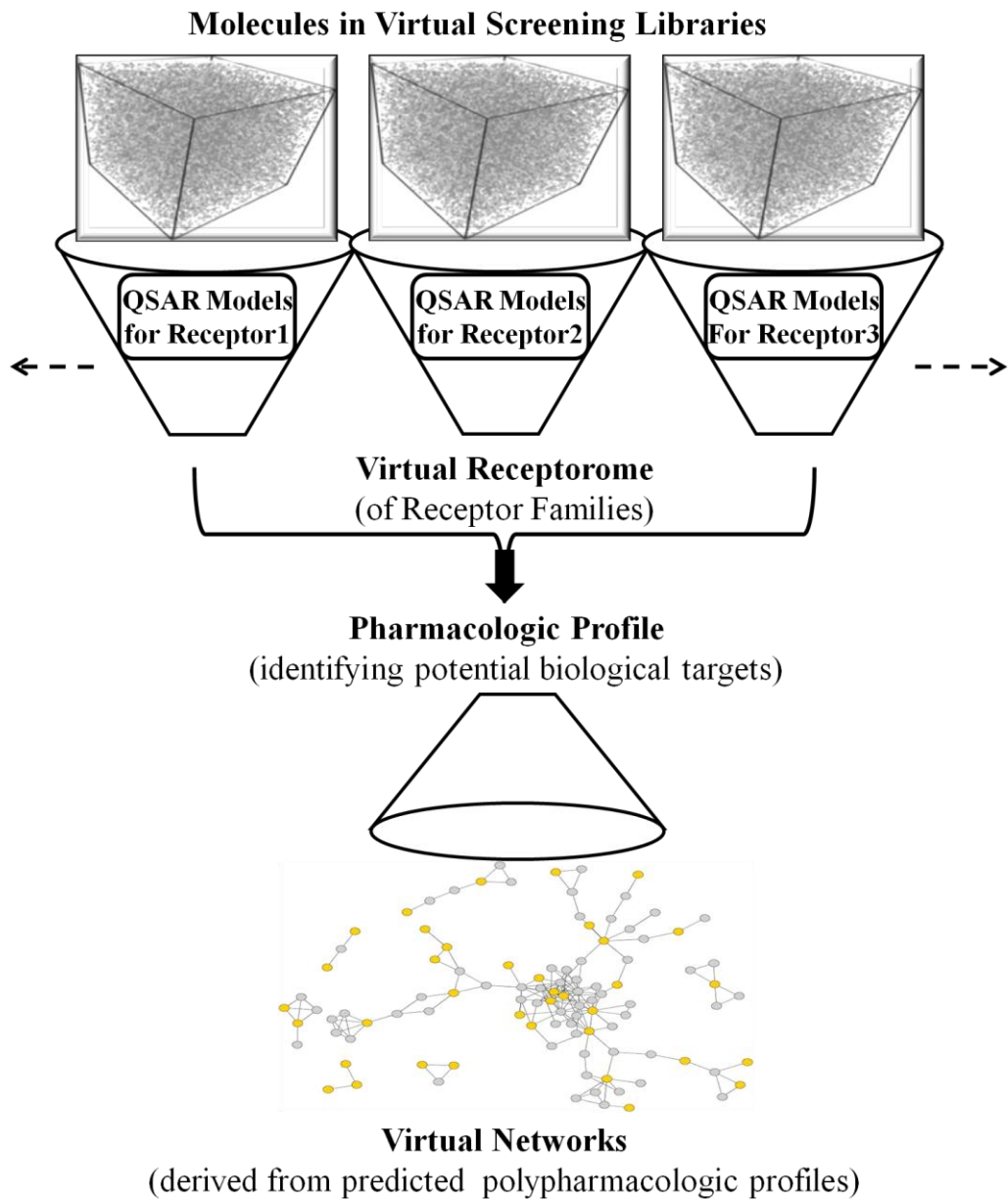
all models can be consolidated in a database of models that can be used for parallel virtual screening of chemical compounds, including current medications, for their abilities to interact with all known members of the receptorome.

*In silico* screening approaches are routinely employed nowadays in academic, governmental and commercial sectors and they have become largely applied in drug discovery (Armbruster & Roth 2005, Bajorath 2002, Bajorath 2005, Becker 2004, Becker et al. 2004, Evers & Klebe 2004, Kitchen et al. 2004, Klabunde et al. 2009). Research conducted in our group has demonstrated that the generation of QSAR models and subsequent model-based virtual screening of chemical libraries has led to the identification of chemically diverse molecules with high success rates in experimental validation tests (Hsieh et al. 2008, Oloff et al. 2005, Peterson et al. 2009, Shen et al. 2002, Shen et al. 2004, Tang et al. 2009, Tropsha 2006, Tropsha & Pearlman 2000, Tropsha & Wang 2006). This approach to drug discovery comprises the following steps: (1) defining the target(s) of interest, (2) extracting relevant structure activity data from the biological literature and specialized databases, (3) dataset curation, (4) compound representation by suitable chemical descriptors, (5) model generation and validation, (6) the application of validated QSAR models for virtual screening (VS) of chemical databases to predict binders and, if possible agonists and antagonists. Often, the interpretation of chemical descriptors found significant for the success of QSAR models can reveal important structural requirements for ligand binding and activity. For the most part, this approach has been applied to datasets of compounds tested in individual assays characterizing their interaction with a single molecular target.

Theoretically, QSAR models explore information restricted to the experimental knowledge of chemical structures and biological activities of ligands. Hence, this approach is

especially important when the X-ray crystal structures for the biological targets of interest (e.g., most GPCRs and trans-membrane proteins) are unavailable. By applying QSAR modeling approach to a large number of datasets, we can accumulate a compendium of QSAR models representing a variety of different biological targets, and subsequently establish a virtual receptorome system to screen molecules simultaneously against an array of available models. Ultimately, we can use these models to obtain a list of common matching hits among several receptor families and could link the hits to all predicted biological targets, thereby enabling an *in silico* identification of biological networks (i.e., virtual networks formed from the ‘predicted’ chemical-molecular target activity profiles across a multitude of molecular targets) which will possibly be influenced by these compounds (concept explained in Fig. 1.1). This approach can also identify selective compounds for each receptor family after excluding common hits. An example of a similar approach to broad *in silico* profiling of compound libraries is given by the method PASS (Brady & Stouten 2000), which currently allows *in silico* screening against a large panel of target proteins. However, the datasets behind PASS models are not publicly available, which makes it difficult to employ and validate alternative techniques to the same datasets.

**Figure 1.1.** Concept of QSAR-based *in silico* receptoromics enabling a virtual network pharmacology approach.



Lately, different computational groups have attempted to predict polypharmacological effects of chemical molecules (Freyhult et al. 2005, Lapinsh et al. 2002, Becker et al. 2004). As an example on one of the most recent efforts, one group at Indiana University used PubChem activity data available at for multiple assays to develop a network representation of the assay collection and then applied a bipartite mapping between this network and various biological networks (Chen et al. 2009). They claimed that their method of mapping to a drug-target network permitted the prioritization of new selective compounds, while mapping to other biological networks enabled them to observe interesting target pairs and their associated compounds in the context of biological systems.

It is likely that our studies and these other efforts described above will eventually result in useful and reliable tools aimed at enriching chemical libraries in compounds that have affinities for more than a single molecular target. We think that a combination of these methods will be more powerful than a single method alone as our expertise in the computational field has indicated time after time.

### **Chemocentric Informatics**

Target-oriented drug discovery has become one of the most popular modern drug discovery approaches (Connor et al. 2010, Nicholson et al. 2004, Petak et al. 2010, Raamsdonk et al. 2001, Yang et al. 2010). Target-oriented approaches rely on established functional associations between activation or inhibition of a molecular target and a disease. Modern genomics approaches including gene expression profiling, genotyping, genome-wide association, and mutagenesis studies continue to serve as useful sources of novel hypotheses linking genes (proteins) and diseases and providing novel putative targets for drug discovery.

Recently, functional genomics approaches have been increasingly augmented by chemical genomics (Brenner 2004, Darvas et al. 2004, Nislow & Giaever 2003, Salemme 2003, Zheng & Chan 2002b, Zheng & Chan 2002a), i.e., large scale screening of chemical compound libraries in multiple biological assays (Campbell et al. 2010, Hamadeh et al. 2010, Kiessling & Splain 2010, Ogorevc et al. 2010, Wagner & Clemons 2009). Chemical genomics studies yield data indicating that both physical and functional interactions exist between chemicals and their biological targets. Such data (either obtained in chemical genomics centers or collected and curated from published literature) is deposited in many public and private databases such as the NIMH Psychoactive Drug Screening Program (PDSP) (PDSP 2009), PubChem (PubChem 2009), ChEMBL (ChEMBL 2010), WOMBAT (Olah et al. 2007) and others (see Oprea and Tropsha (Oprea & Tropsha 2006) for a recent review).

Various *in silico* techniques have been exploited for analyzing target-specific biological assay data. A recent publication by Kortagere and Ekins (Kortagere & Ekins 2010) could serve as a good summary of most common target-oriented computational drug discovery approaches including: (1) structure based virtual screening (docking and scoring) using either experimentally characterized (with X-ray or NMR) or predicted by homology modeling structure of the target protein, (2) chemical similarity searching using known active compounds as queries, (3) pharmacophore based modeling and virtual screening, (4) quantitative structure activity relationship (QSAR) modeling, and (5) network or pathway analysis.

## **‘Omics’ Data Types and Repositories**

Data resulting from large-scale gene or protein expression or metabolite profiling (often collectively referred to as 'omics' approaches) (Burgun & Bodenreider 2008, Kandpal et al. 2009, Polychronakos 2008, Vangala & Tonelli 2007) can be explored not only for specific target identification but also in the context of systems pharmacology to identify networks of genes (or proteins) that may collectively define a disease phenotype. For example, ‘omics’ data can be used to ask what genes or proteins, or post-translationally modified states of proteins are over- (or under-) expressed in patients suffering from a particular disease. These types of data can be found in a number of public repositories such as the Gene Expression Omnibus (GEO) (Edgar et al. 2002, Barrett & Edgar 2006), GEOmetadb (Zhu et al. 2008b), the Human Metabolome Database (HMDB) (Wishart 2007, Wishart et al. 2009), Kinase SARfari (Kinase SARfari 2010), the Connectivity Map (cmap) (Lamb et al. 2006), the Comparative Toxicogenomics Database (CTD) (Davis et al. 2009), STITCH (Kuhn et al. 2009, Kuhn et al. 2008), GenBank (Burks et al. 1991, Burks et al. 1990), and others.

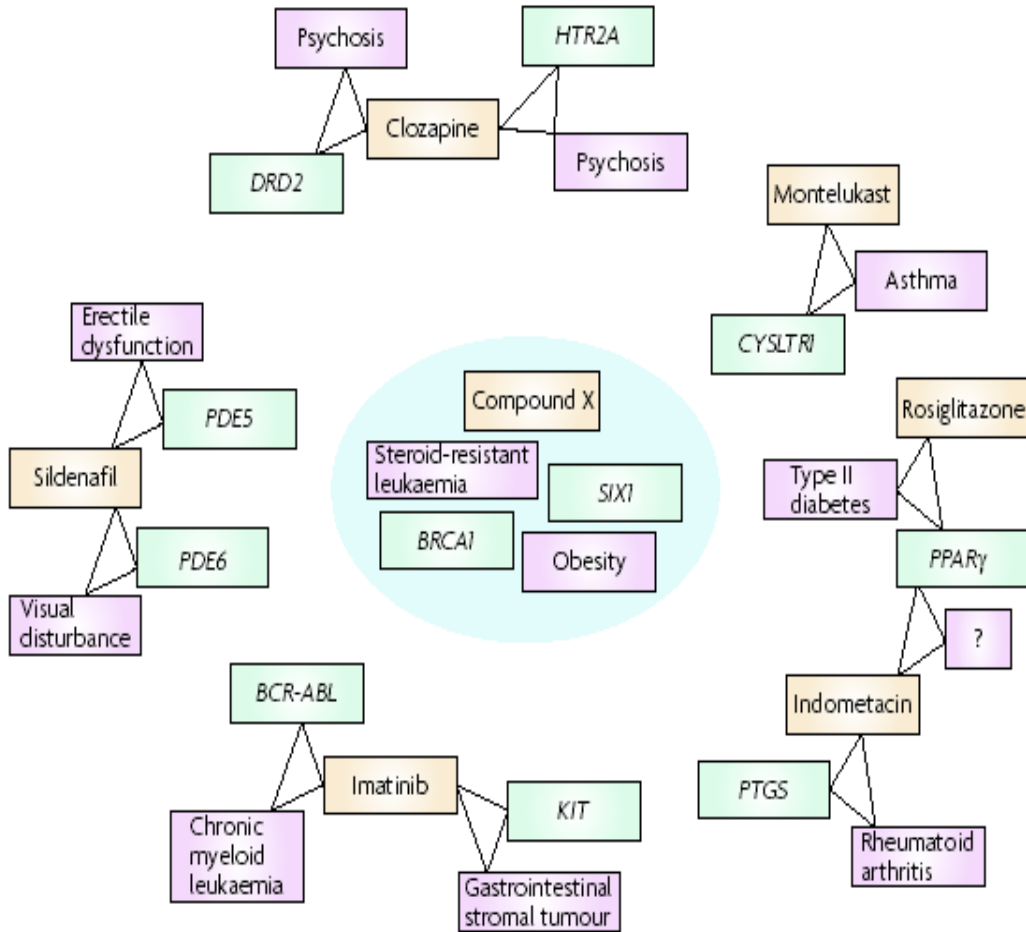
Insights into disease pathology and underlying mechanisms can be revealed by the disease ‘gene signature’, i.e., those genes whose expression varies consistently between patients and healthy individuals (controls) (Palfreyman et al. 2002). Gene-expression profiling has been often applied to elucidate the mechanisms underlying the roles of biological pathway in a disease (DeRisi et al. 1997, Lamb et al. 2003), reveal arcane subtypes of a disease (Golub et al. 1999, Perou et al. 2000), and predict cancer prognosis (Pomeroy et al. 2002, van, V et al. 2002). At the same time, the treatment of cultured human cells with chemical compounds that target a disease can produce a drug related ‘gene signature’, i.e.,



differential expression profile of genes in response to the chemical (Altar et al. 2009, Ogden et al. 2004, Palfreyman et al. 2002, Le-Niculescu et al. 2007). Recently, a group of scientists at the Broad Institute have established the Connectivity Map (cmap) database (see Fig. 1.2 for concept) to catalog the biological responses of a large number of diverse chemicals in terms of their gene expression profiles (Lamb et al. 2006). It has been shown that examining the correlations between gene expression profiles characteristic of a disease and those modulated by drugs may lead to novel hypotheses linking chemicals to either etiology or treatments for a disease (Garman et al. 2008, Golub et al. 1999, Hassane et al. 2008, Hieronymus et al. 2006, Lamb et al. 2006, Riedel et al. 2008, Setlur et al. 2008, Zimmer et al. 2008, Zimmer et al. 2010).

The cmap database provides an unusual but intriguing example of what we shall call a *chemocentric* ‘omics’ database and methodology for generating independent and novel drug discovery hypotheses. Indeed, there exists a wealth of information buried in the biological literature and numerous specialized chemical databases (ChEMBL 2010, 2004, PDSP 2009, PubChem 2009, Olah et al. 2007) linking chemical compounds and biological data (such as targets, genes, experimental biological screening results; cf. Baker and Hemminger (Baker & Hemminger 2010)). The chemocentric exploration of these sources, either individually or in parallel opens up vast possibilities for formulating novel drug discovery hypotheses concerning the predicted biological or pharmacological activity of investigational chemical compounds or known drugs. The integration and cross-validation of such independent structural hypotheses can increase their level of confidence and can be referred to as structural hypothesis fusion.

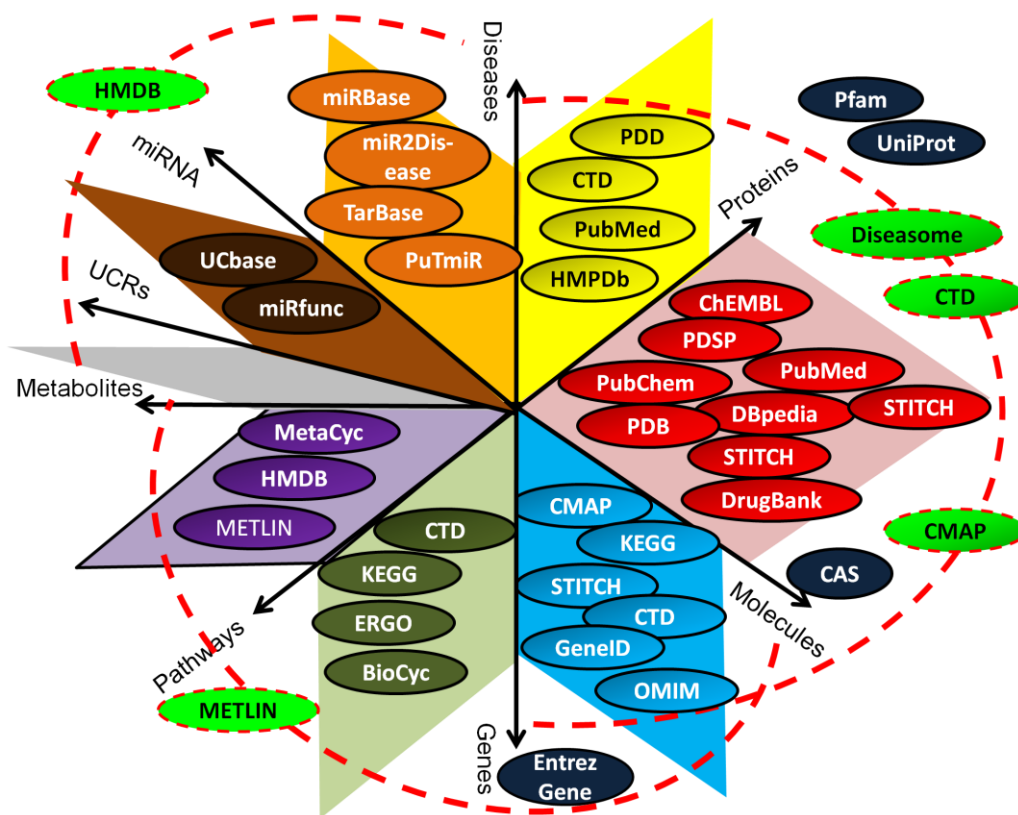
**Figure 1.2.** The connectivity map concept. Functional relationships between a drug (yellow), a gene (green) and a disease (pink) constitute the nodes in this map. Adopted from Lamb, J. *Nature*, 7, 54-60.



## **The Multidimensional Chemocentric Space**

In order to design novel chemocentric informatics approaches to tackle the problems of conventional drug discovery projects, we should first understand the chemocentric space and how to divide this space into smaller subspaces that could be dealt with separately. The chemocentric space is a complex space defined by multiple interconnected dimensions. This is schematically represented in Fig.1.3, where molecules, proteins, genes, pathways, and diseases are some of the different dimensions represented within the chemocentric space. Different types of experimental data, deposited in a multitude of databases, allow for connecting pairs of dimensions in this representation. For example, pharmacological data on the binding affinities of chemicals to particular proteins allow the mapping of the molecule-protein space (see Fig.1.3), one of the two-dimensional subspaces defining this multidimensional chemocentric space.

**Figure 1.3.** Chemocentric informatics multidimensional space with examples on online databases populated with relevant data concerning the two dimensions forming each subspace. Dashed red lines show examples of connecting other dimensions and forming new subspaces that are not clearly drawn here.



## **Hypothesis Fusion**

Data Fusion is the process of combining multiple data in order to produce new information that improves the performance of a system (i.e., *in silico* model or predictor). Data can come from one or many sources. Sources may be similar, or dissimilar. Data fusion may be useful for several objectives such as detection, recognition, identification, tracking, change detection, decision making, etc. Similarly, hypothesis fusion is the process of fusing different hypotheses derived independently from different data types. In all cases, efficient fusion schemes may have significant advantages such as: (1) improved confidence in decisions due to the use of complementary information, (2) improved performance to countermeasures (e.g., outliers in QSAR studies) and (3) Improved performance in adverse experimental conditions. This fusion approach was first developed for applications in signal processing (Klien 1999) and later on was applied in VS efforts to enable better decisions to which small number of molecules should go further for biological testing (Sukumar et al. 2008, Whittle et al. 2006a, Whittle et al. 2006b).

Herein, we used hypothesis fusion to cross-examine structural hypotheses that have been generated from different sets of data and using different machine learning algorithms and then applied for virtual screening of chemical libraries with those hypotheses derived from biological network mining efforts. Finally, we accepted accept common hits only based on chemical structure identity.

## **Overview of Chapter 2: Materials and Methods**

In Chapter 2 we discussed the major computational approaches applied in our studies including: (1) QSAR model development, (2) QSAR-based VS, (3) Comparison studies with simple similarity searches to evaluate the performance of our QSAR methods, (4) Devising a

novel chemocentric informatics approach to study polypharmacology, and (5) Independent structural hypothesis fusion.

### **Overview of Chapter 3: *In silico* Receptoromics**

Studies described in this chapter were designed to improve the rigor of QSAR modeling techniques to analyze GPCR datasets including the computational data modeling as well as novel approaches to tackle difficult problems in modern QSAR modeling such as dealing with unbalanced datasets, and applying different validation methods (e.g., consensus prediction, applicability domain, consensus predictions thresholds, and external sets) to improve the predictive power of models. Models were generated for many GPCRs with a special focus on a few receptors that are highly implicated in drug design efforts for neuropsychiatric and neurodegenerative diseases. Anti-target GPCRs (e.g., 5-HT<sub>2B</sub> receptors) were given special emphasis as well and will be discussed separately as a model QSAR study in Chapter 4.

The growing understanding within molecular pharmacology of the complexity of biological networks and the robustness and redundancy of biological systems, however, challenges the current approaches of single target drug discovery including *in silico* approaches (Hopkins 2007, Hopkins 2008, Roth et al. 2000, Roth & Kroeze 2006). Nowadays, medicinal chemists are becoming more interested in identifying polypharmacological drugs that can bind moderately to several targets in a disease-protein network and affect the overall outcome significantly (Hopkins et al. 2006, Hopkins 2009, Roth et al. 2004). Herein, we suggest for the first time receptor-family-based QSAR models as computationally inexpensive tools for the quick prioritization of polypharmacological hits for further experimental testing against a large panel of receptors. Additionally, the generated

receptor family-based models will be highly valuable due to both their biological relevance and statistical significance.

Family based models seem promising tools to statically increase the rigor of the generated QSAR models for the following reasons: (1) increased size of datasets, (2) increased diversity of datasets, (3) improved applicability domains of the models, (4) suitability of the dataset for applying MTL to uncover hidden cross-family relationships and family specific chemical features. As a consequence, it is very likely that these models will increase the potential of QSAR models to indentify novel leads that are difficult to uncover otherwise.

#### **Overview of Chapter 4: The 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**

In this study, we have applied a combinatorial QSAR approach to a dataset of 800 compounds experimentally annotated as 5-Hydroxytryptamine (2B) (5-HT<sub>2B</sub>) receptor agonists, antagonists and inactives resulting in statistically validated and externally predictive models. We will discuss this study in details as an example on our QSAR performed herein. Specifically, we used three different classification methods: *k* nearest neighbor (*k*NN), classification based on association (CBA), and distance weighted discrimination (DWD) and four different descriptor types (Dragon, MolconnZ, MOE and subgraphs) to generate classification QSAR models to discriminate between 5-HT<sub>2B</sub> actives (agonists and antagonists) from inactives. Predictive models with

classification accuracies as high as 0.80 for actives *vs.* inactives, as estimated on external validation sets, were obtained.

Classification models for actives *vs.* inactives were further validated by predicting an external validation set obtained after we completed the modeling studies. The high accuracy of prediction for the second external validation set proved that our models were indeed rigorous. Therefore, we posited that our studies afforded a robust computational tool to predict potential 5-HT<sub>2B</sub> activity and consequently prioritize hits for testing in functional 5-HT<sub>2B</sub> assays to predict valvulopathic side effects of drugs and drug candidates that act as 5-HT<sub>2B</sub> agonists. We suggested that this computational predictor could be used to eliminate high risk compounds at the early stages of the drug development process. To illustrate this point, we have used this predictor retrospectively to evaluate the valvulopathic potential of two drugs withdrawn from the U.S. market for this reason, *i.e.*, fenfluramine and dexrofenfluramine. Both drugs were not included in our modeling set and both were indeed predicted with high confidence as actives for binding to 5-HT<sub>2B</sub> receptors.

Encouraged by our model validation results, we have applied these models for virtual screening of the 59000 compounds in WDI database. Our classification strategies identified 122 potential 5-HT<sub>2B</sub> ligands. Ten structurally diverse VS hits were experimentally tested at PDSP. Nine compounds were experimentally confirmed as 5-HT<sub>2B</sub> ligands thereby demonstrating a very high success rate of 90%.

The predictor developed in this report is similar in its potential use to other predictors of drug liability such as carcinogenicity and mutagenicity that are widely used



in pharmaceutical industry. For instance, the TOPKAT program available in the Discovery Studio (Discovery Studio 2008), is a QSAR-based system that generates and validates accurate, rapid assessments of various types of chemical toxicity solely from a chemical's molecular structure. In contrast, our predictor is a unique specialized tool for the prediction of 5-HT<sub>2B</sub> activity and therefore prioritizing compounds for functional testing against 5-HT<sub>2B</sub> receptors to assess their valvulopathic potential. Therefore, this predictor can be used, along with other computational chemical health risk assessment tools, to evaluate compounds' safety at early stages of the drug development. It can be used as well to verify that all drugs available on the market are free from possibly fatal valvulopathic risk. This predictor is publicly available at the ChemBench server established in the Laboratory for Molecular Modeling ([chembench.mml.unc.edu](http://chembench.mml.unc.edu)).

## **Overview of Chapter 5: An Integrative Chemocentric Informatics Approach to Drug Discovery Based on Structural Hypothesis Fusion**

Herein, we describe a novel integrative chemocentric informatics approach to drug discovery that combines structural hypotheses generated from independent analysis of both traditional target-specific assay data and those resulting from large scale genomics and chemical genomics studies. Herein, we have focused on the Alzheimer's disease as one of the most debilitating neurodegenerative diseases with complex etiology and polypharmacology. We have considered and cross-examined two independent but complimentary approaches to the discovery of novel putative anti-Alzheimer's drugs. First, we have employed a traditional target-oriented cheminformatics approach to discovering anti-Alzheimer's agents. We have built QSAR models of ligands binding to 5-hydroxytryptamine-6 receptor (5-HT<sub>6</sub>R), a potential target for the cognitive enhancement in

Alzheimer's disease (Geldenhuys & Van der Schyf 2009); it has been shown that 5-HT<sub>6</sub>R antagonists can improve memory and cognition in animal models of impaired cognition (Holenz et al. 2006). We have then used models developed with the rigorous predictive QSAR modeling workflow established and implemented in our laboratory (Tropsha 2010) for virtual screening (VS) of the World Drug Index database (WDI) (Daylight 2004) and DrugBank (Wishart et al. 2006, Wishart et al. 2008) to identify putative cognition enhancing agents with potential utility as anti-Alzheimer's agents as compounds predicted to interact with 5-HT<sub>6</sub>R. Second, we have explored (chemo)genomic data available from the cmap project (Lamb et al. 2006) to link chemical compounds and the Alzheimer's disease without making explicit hypotheses about target-specific mechanisms of action, i.e., treating Alzheimer's disease as a complex polypharmacological disease.

We then cross-examined and combined common hits regarded as structural hypotheses resulting from both approaches (i.e., hypothesis fusion) towards common integrated higher-confidence hypotheses supported by two independent lines of computationally-based evidence. Thirteen common hits were tested in 5-HT<sub>6</sub>R binding assays at the NIMH Psychoactive Drug Screening Program (PDSP) and ten were confirmed experimentally as having activity. Unexpectedly, we found that the confirmed actives included several selective estrogen receptor modulators (SERMs) suggesting that they may be potential 5-HT<sub>6</sub>R actives as well as cognitive enhancing agents in Alzheimer's disease. Indeed, we have identified clinical evidence in biomedical literature in support of this hypothesis. We believe that approaches discussed in this study can be applied to a large variety of systems to identify novel drug-target-disease associations.

## **Overview of Chapter 6: Conclusions and Future Studies**

We think it is likely that our computational efforts described herein and other efforts by different groups to study polypharmacology *in silico* will eventually result in useful and reliable tools aimed at enriching chemical libraries in compounds that have affinities for more than a single desired molecular target. We think that a combination of these methods will be more powerful than a single method alone as our expertise in the computational field has indicated time after time. However, our studies revealed some limitations of the current available methods that could be improved dramatically in the near future with the availability of more specialized databases, better disease signatures, and full matrices of tested chemical-molecular target interactions.

## CHAPTER 2

### MATERIALS AND METHODS

#### Databases and Datasets

**PDSP  $K_i$ -DB.** PDSP  $K_i$ -DB (PDSP 2009) (<http://pdsp.med.unc.edu/pdsp.php>) includes published binding affinities ( $K_i$ ) of drugs and chemical compounds for receptors, neurotransmitter transporters, ion channels, and enzymes. It currently lists more than 47000  $K_i$  values for more than 700 molecular targets.  $K_i$ -DB represents a curated, fully searchable database of both published data and data internally-derived from the NIMH-PDSP. The experimental data for Alzheimer's disease related target 5-HT<sub>6</sub>R were extracted from the PDSP  $K_i$ -DB available in the public domain. The complete 5-HT<sub>6</sub>R dataset included binding affinity data for 250 compounds

**World Drug Index.** The world drug index (WDI) (Daylight 2004) is an authoritative database for marketed and developmental drugs providing information about internationally recognized drug names, synonyms, trade names, trivial names, trial preparation codes, compound structures, and activity data. Herein, we used WDI for QSAR-based VS to identify putative 5HT<sub>6</sub>R ligands.

**DrugBank.** DrugBank (Wishart et al. 2008) (<http://www.drugbank.ca>) is a unique bioinformatics and cheminformatics resource that combines detailed drug data (i.e., chemical, pharmacological, and pharmaceutical) with comprehensive drug target information (i.e., sequence, structure, and pathway). Currently, the database contains nearly 4800 drug

entries. Herein, we used DrugBank for virtual screening using QSAR models to identify putative 5-HT<sub>6</sub>R ligands among known drugs

**PubChem.** PubChem (PubChem 2009) (<http://pubchem.ncbi.nlm.nih.gov/>) is a public repository of chemical structures and their activities obtained from a variety of biological assays. The PubChem compound repository presently contains more than 25 million unique structures with biological property information provided for many of the compounds. Herein, we used PubChem obtain all chemical structures for our datasets in SDF file format.

**Connectivity Map.** The connectivity map (cmap) (Lamb et al. 2006, Lamb 2007) (<http://www.broadinstitute.org/cmap/>) is a unique database for using genomics in drug discovery framework. It provides researchers with a systematic solution for the discovery of the functional connections between drugs, genes, and diseases. The database (cmap build 02) currently houses 7056 genome-wide expression profiles representing 6100 individual treatment instances with 1309 bioactive small molecules (i.e., drugs and other biologically active compounds). All gene expression profiles included in the cmap were derived from treating cultured human cells (MCF7, PC3, HL60, SKMEL5, HepG2, SHSY5Y) with chemical compounds.

**ChemoText.** ChemoText (Baker & Hemminger 2010) is an in-house repository of chemical entities, and activity terms (indicating biological effects) extracted from annotations provided in Medline records. This resource has different applications in drug discovery projects. First, we can use ChemoText in a discovery-mode to formulate independent hypotheses about chemical-disease associations according to Swanson's ABC rule (Swanson 1990). Secondly, we can use it as an information retrieval tool to gather relevant data about chemical-protein (or gene)-disease connections derived from biomedical literature. In this

study, we used ChemoText to retrieve all available biological information about the final computational hits predicted by our integrative approach. This analysis helped us in assessing the novelty of the produced hypotheses and in validating some of them.

**STITCH.** STITCH (Kuhn et al. 2008, Kuhn et al. 2009) (<http://stitch.embl.de/>) is a tool for searching chemical and protein interaction networks. It integrates information from metabolic pathways, crystal structures, binding experiments, and drug-target relationships. Inferred information from phenotypic effects, text mining, and chemical structure similarity is used to predict relations between chemicals. The database contains interaction information for over 68000 chemicals, including 2200 drugs, and connects them to 1.5 million genes across 373 genomes. In this study we used STITCH to analyze chemical-protein networks for some computational hits predicted by our QSAR-based VS or the integrative approach to be discussed later.

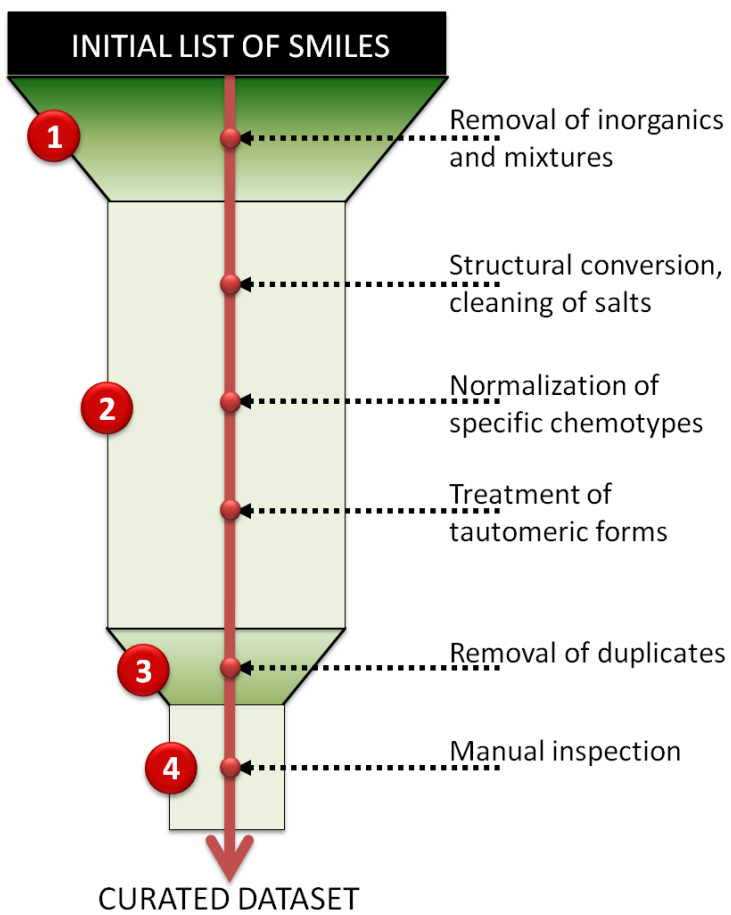
**NetAffx.** NetAffx (Cheng et al. 2004, Liu et al. 2003) (<http://www.affymetrix.com>) gene ontology mining tool is a web-based, interactive tool that permits traversal of the gene ontology graph in the context of microarray data. It accepts a list of Affymetrix probe sets and renders a gene ontology graph as a heat map colored according to significance measurements. It also details and annotates probe sets on Affymetrix GeneChip microarrays. In this study we used NetAffx to populate our disease gene signatures with Affymetrix U133A probe sets.

### **Dataset curation**

Data curation is a mandatory step in data analysis that must be performed before proceeding with any modeling project. Several case studies have been reported by our lab where chemical curation of the original “raw” dataset/database resulted in a significant

improvement of the outcome of the modeling study (especially, QSAR analysis). The different steps that will be used for cleaning chemical records in datasets and databases include: the removal of a fraction of data that cannot be appropriately handled by conventional cheminformatics techniques such as inorganic compounds, counter-ions and mixtures; structure validation; ring aromatization; normalization of specific chemo-types; curation of tautomeric forms; addition or deletion of hydrogen atoms; and the deletion of duplicates (see Fig. 2.1) (Fourches et al. 2010).

**Figure 2.1.** General workflow of chemical dataset curation developed in our lab (Fourches et al. 2010).





For the purposes of this work, all datasets will be curated as follows: First, all molecules will be “washed” using both the Wash Molecules application in MOE (MOE 2008) (v.2007.09) and ChemAxon Standardizer included in the ChemAxon JChem package (JChem 2009). The MOE Wash application normalizes chemical structures by carrying out a number of operations including 2D depiction layout, hydrogen correction, salt and solvent removal, chirality and bond type normalization, tautomer generation, adjustment and enumeration of protonation states. Second, duplicate chemical structures will be removed using the Sort and Remove Duplicates functionalities in MOE: keeping one chemical structure only if both activities in both cases are the same and removing both if activities were different. Activities might differ due to different stereochemistry that is not considered in our modeling studies where we only use 2D descriptors that cannot differentiate between stereoisomers. Finally, a careful manual inspection step should not be neglected as a last step in data curation. Some of the common errors identified during the manual cleaning procedure may include: wrong structures, incomplete normalization of chemical bonds, some duplicates may still be present despite the use of automated software to remove them, wrong charges, presence of explicit hydrogens in a hydrogen depleted structures, incorrect bonds, etc.

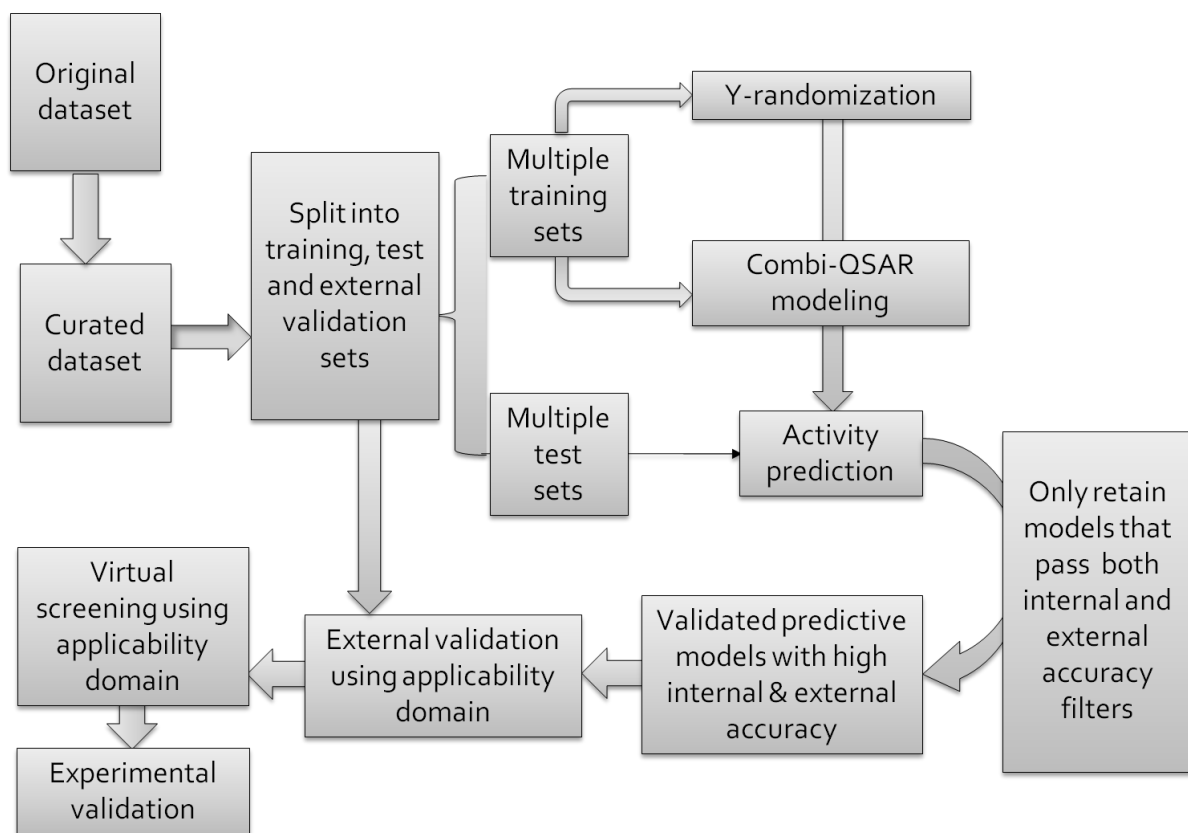
## Computational Methods

We used the combinatorial QSAR approach using different sets of molecular descriptors and applying several machine learning methods to establish the correlation between structural descriptors and biological activities. We also devised a novel chemocentric informatics approach the course of these studies.

### Combinatorial QSAR Approach

To achieve QSAR models of the highest internal, and most importantly, external accuracy, we apply a combi-QSAR approach (de Cerqueira et al. 2006, Kovatcheva et al. 2004), which explores all possible combinations of various descriptor types and optimization methods along with external model validation. All modeling attempts are conducted according to our predictive QSAR modeling workflow (Fig. 2.2) (Tropsha 2010). For purposes of this research, descriptors types mentioned earlier and different QSAR methods will be explored. We envision QSAR as a highly experimental area of statistical data modeling where it is impossible to decide *a priori* as to which particular QSAR modeling method will prove most successful. Each combination of descriptor sets and optimization techniques is likely to capture certain unique aspects of the structure-activity relationship. Since our ultimate goal is to use the resulting models as reliable activity (property) predictors, combi-QSAR will increase our chances for success.

**Figure 2.2.** Flowchart of predictive QSAR modeling framework based on validated QSAR models.



## Molecular Descriptors

Molecular descriptors are numerical values that characterize properties of molecules. They vary in complexity of encoded information and in computation time. For purposes of this research we will be using the so called 2D molecular descriptors due to their relative simplicity of calculation, lack of dependence on conformation, and demonstrated ability to compete, if not outperform, 3D descriptors in chemical similarity and QSAR studies. Herein, five sets of 2D molecular descriptors will be used: 2D Dragon (Dragon 2007), MolConnZ (MZ) (MolconnZ 2006), Molecular Operating Environment (MOE) (MOE 2008), subgraph descriptors (SG) (Khashan et al. 2005) developed in this laboratory and MACCS structural keys (MDL Ltd 1992). Each type of these descriptors will be used separately with different machine learning methods in the context of our combi-QSAR strategy.

**DRAGON Descriptors.** The Dragon Professional version 5.4 software (Dragon 2007) was used to calculate 2D descriptors. These included topological descriptors, constitutional descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge indices, eigenvalue-based indices, functional group counts, atom-centered fragments and molecular properties. The initial descriptor set was reduced by eliminating the constant and near-constant variables using built-in functions within the software. The pairwise correlations for all descriptors were examined and one of the two descriptors with the correlation coefficient  $R^2$  of 0.95 or higher was excluded. The calculation procedures for these descriptors, with related literature references, are reported by Todeschini and Consonni (Todeschini & Consonni 2000). Finally, the remaining

descriptors were normalized by range-scaling so that their values were distributed within the interval 0-1.

**MolConnZ Descriptors.** The MolConnZ (MZ) software (MolconnZ 2006) available from EduSoft affords the computation of a wide range of topological indices of molecular structure. These indices include, but are not limited to, the following descriptors: valence, path, cluster, path/cluster and chain molecular connectivity indices (Kier & Hall 1976, Kier & Hall 1986, Randic 1975), kappa molecular shape indices (Kier 1985, Kier 1987), topological (Hall & Kier 1990) and electrotopological state indices (Hall et al. 1991a, Hall et al. 1991b, Kellogg et al. 1996, Kier & Hall 1999), differential connectivity indices (Kier & Hall 1986, Kier & Hall 1991), graph's radius and diameter (Petitjean 1992), Wiener (Wiener 1947) and Platt (Platt 1947) indices, Shannon (Shannon & Weaver 1949) and Bonchev-Trinajstić (Bonchev et al. 1981) information indices, counts of different vertices, counts of paths and edges between different types of vertices (<http://www.edusoft-lc.com/molconn/manuals/400>). Descriptors with zero values or zero variance were removed; the remaining descriptors were normalized by range-scaling so that their values were distributed within the interval [0-1].

**MOE Descriptors.** MOE 2007.09 software (MOE 2008) was used to generate 2D MOE descriptors. These included physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity (Kier & Hall 1976, Kier & Hall 1986, Randic 1975) and kappa shape indices (Kier 1985, Kier 1987), adjacency and distance matrix descriptors (Balaban 1979, Balaban 1982, Petitjean 1992, Wiener 1947), pharmacophore feature descriptors, and partial charge descriptors (MOE 2008). Descriptors with zero

values or zero variance were removed; the remaining descriptors were normalized by range-scaling so that their values were distributed within the interval [0-1].

**Subgraph Descriptors (SG).** Frequent subgraph mining of chemical structures is a novel approach to generating fragment descriptors that was developed recently in our group (Khashan et al. 2005). SG descriptors are derived from each dataset, i.e., not pre-defined which gives the advantage of finding important chemical fragments that may have not been defined *a priori* by other fragment descriptor generating methods. The fragments are derived based on recurring substructures found in at least a subset of molecules (defined by a support value  $\sigma$ ) in the dataset. These recurring substructures can implicate chemical features responsible for compounds' biological activities.

First, chemical structures were converted into labeled, undirected graph representations where nodes were labeled by atom types and edges corresponded to chemical bonds. Fast frequent subgraph mining (FFSM) algorithm (Huan et al. 2003) was then used to find common frequent subgraphs for a given support value ( $\sigma$ ), which is one of the variables defined by the user that determines the size of the set of subgraphs generated using FFSM. Obviously, the larger is the value of the support, the smaller is the number of subgraphs descriptors. As the support value decreases, the number of subgraphs increases dramatically. Redundant subgraphs were identified and removed leaving only the so called "closed subgraphs". A subgraph  $SG_i$  is closed in a database if there exists no supergraph  $SG_j$  such that  $SG_i \subseteq SG_j$  and  $\sigma_{SG_i} = \sigma_{SG_j}$ . However, subgraph  $SG_i$  would not be deleted if it also occurs by itself (not as part of the  $SG_j$ ) in the graph database. Removing redundant subgraphs (fragments) reduces the number of subgraphs descriptors drastically and therefore makes the subsequent calculations more efficient.

The frequency of individual ‘closed subgraphs’ in each molecule of the dataset is calculated and used as the descriptor value for each molecule.

**MACCS Structural Keys.** 166 MACCS (MDL Ltd 1992) structural keys implemented in MOE 2007.09 software (MOE 2008) were used for purposes of simple similarity searching using an in-house written script and applying Tanimoto coefficients for similarity measures.

### **Machine Learning Methods**

Different machine learning algorithms will be used to correlate chemical descriptors with the corresponding biological activities. Correlation algorithms including *k*NN (Zheng & Tropsha 2000), super vector machine (SVM) (Cortes & Vapnik 1995), classification based on association (CBA) (Liu et al. 2001), distance weighted discrimination (DWD) (Marron et al. 2007), MOE (MOE 2008) binary QSAR, and MOE decision trees will be used in this research in combination with the descriptor types mentioned earlier. There is no single machine learning method that can claim to be uniformly superior to any other. Hence, the implementation of combi-QSAR (Kovatcheva et al. 2004, de Cerqueira et al. 2006) (i.e. ensemble machine learning), a set of modeling techniques (different molecular descriptors and different correlation algorithms) whose individual decisions are combined in some way (typically by weighted or un-weighted voting) will be employed to improve the performance of the overall modeling system and the success rates of *in silico* lead identification.

***k* Nearest Neighbors QSAR.** The *k* nearest neighbors (*k*NN) QSAR method (Zheng & Tropsha 2000) is based on the *k* nearest neighbors principle and the variable selection procedure. It employs the leave-one-out (LOO) cross-validation (CV) procedure and a simulated-annealing algorithm (Kirkpatrick et al. 1983, Metropolis et al. 1953) to

optimize variable selection. The procedure starts with the random selection of a predefined number of descriptors from all descriptors. If the number of nearest neighbors  $k$  is higher than one, the estimated activities  $\hat{y}_i$  of compounds excluded by the LOO procedure are calculated using the following formula:

$$\hat{y}_i = \frac{\sum_{j=1}^k y_j w_{ij}}{\sum_{j=1}^k w_{ij}} \quad (\text{Eq. 2.1})$$

where  $y_j$  is the activity of the  $j$ -th compound. Weights  $w_{ij}$  are defined as:

$$w_{ij} = \left( 1 + \frac{d_{ij}}{\sum_{j'=1}^k d_{ij'}} \right)^{-1} \quad (\text{Eq. 2.2})$$

and  $d_{ij}$  is Euclidean distances between compound  $i$  and its  $j$ -th nearest neighbor. However, if the number of nearest neighbors  $k$  is equal to one, then the estimated activity  $\hat{y}_i$  of the compound will be equal to the activity of this one nearest neighbor.

The  $k$ NN classification algorithm employs an LOO cross-validation procedure on the training set and a simulated annealing algorithm in order to select subsets of descriptors, which lead to the highest LOO cross-validation correct classification rate (CCR). The procedure starts with the random selection of a predefined subset of descriptors from all descriptors. If the number of nearest neighbors  $k$  is higher than one, estimated activities  $\hat{y}_i$  of compounds excluded by LOO procedure are calculated using the following formula:



$$\hat{y}_i = \text{Rounding}\left(\frac{\sum_{j=1}^k y_j w_{ij}}{\sum_{j=1}^k w_{ij}}\right) \quad (\text{Eq. 2.3})$$

where  $y_j$  is the binary activity of the  $j$ -th nearest neighbor. Weights  $w_{ij}$  are defined in Eq. (2), where  $d_{ij}$  is the Euclidean distance between compound  $i$  and its  $j$ -th nearest neighbor. If  $k=1$ , then  $\hat{y}_i = y_i$ .

The predicted values are then rounded to the closest integer. After each run, CCR and other statistical parameters are calculated as follows:

$$\text{Accuracy} = \frac{N_{\text{TruePositives}} + N_{\text{TrueNegatives}}}{N_{\text{TruePositives}} + N_{\text{FalseNegatives}} + N_{\text{TrueNegatives}} + N_{\text{FalsePositives}}} \quad (\text{Eq. 2.4})$$

$$\text{CCR} = \frac{(\text{Sensitivity} + \text{Specificity})}{2} \quad (\text{Eq. 2.5})$$

$$\text{Sensitivity} = \frac{N_{\text{TruePositives}}}{N_{\text{TruePositives}} + N_{\text{FalseNegatives}}} \quad (\text{Eq. 2.6})$$

$$\text{Specificity} = \frac{N_{\text{TrueNegatives}}}{N_{\text{TrueNegatives}} + N_{\text{FalsePositives}}} \quad (\text{Eq. 2.7})$$

$$\text{Precision} = \frac{N_{\text{TruePositives}}}{N_{\text{TruePositives}} + N_{\text{FalsePositives}}} \quad (\text{Eq. 2.8})$$

Then, a predefined small number of descriptors are randomly replaced by other descriptors from the original pool, and a new CCR value is obtained. If  $\text{CCR}(\text{new}) > \text{CCR}(\text{old})$ , the new set of descriptors is accepted; otherwise, if  $\text{CCR}(\text{new}) \leq \text{CCR}(\text{old})$ , the new set of descriptors is accepted with probability  $p = \exp((\text{CCR}(\text{new}) - \text{CCR}(\text{old}))/T)$ , or rejected with probability  $(1-p)$ , where  $T$  represents the simulated annealing temperature parameter. During this process,  $T$  is decreasing until a predefined threshold. Thus, the optimal (highest) CCR is achieved (Zheng & Tropsha 2000, Xiao et al. 2004). For the prediction, the final set

of selected descriptors is used, and expressions (2) and (3) with rounding the predicted activity to the closest integer are applied to predict activities of test set compounds. Prediction is unreliable, if a representative point of a compound is “too far” from the  $k$  nearest neighbors representing training set compounds with known activities. To limit the model’s applicability domain, we apply a distance cutoff value between a compound under prediction and its nearest neighbors of the training set.

**Support Vector Machines.** The description of the original support vector machines (SVM) algorithm could be found in many publications (Cortes & Vapnik 1995, Chang & Lin 2001). Briefly, molecular descriptors are first mapped onto a high dimensional feature space using various kernel functions. Then, SVM finds a separating hyperplane with the maximal margin in this high dimensional space in order to separate compounds with different activities. Models built with this machine learning technique allow the prediction of a target property using a set of descriptors solely calculated from the structure of a given compound. In this study, we used the WinSVM program developed in our group (freely available for academic laboratories upon request) implementing the open-source libSVM package (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>) (Chang & Lin 2001). The WinSVM program provides users with a convenient graphical interface to prepare input data; to split datasets into training and test sets; to set up parameters for SVM grid calculations, including iterative and simultaneous grid optimization of SVM parameters; to launch and follow calculation progress in a powerful graphical interface; to select models with the best prediction accuracy on both training and internal test sets; and to apply them to the external evaluation set as an ensemble consensus model. The program also allows one to visualize molecular structures and various plots, making the use of SVM easier and more appropriate for QSAR modeling

in order to obtain robust and predictive models and apply them to virtual libraries.

**Classification Based on Association.** Classification based on association (CBA) method integrates both classification rule mining (Breiman et al. 1984, Quinlan 1992), which aims to discover a small set of rules in the database that forms an accurate classifier, and association rule mining (Agrawal & Srikant 1994), which finds all the rules existing in the database that satisfy some minimum support and minimum confidence constraints. For association rule mining, the target of discovery is not predetermined, while for classification rule mining there is one and only one predetermined target. The integration is done by focusing on mining a special subset of association rules, called *class association rules* (CARs). An efficient algorithm is also used for building a classifier based on the set of discovered CARs.

The CBA algorithm (Liu et al. 1998, Liu et al. 1999) consists of two parts, a rule generator, which is based on the *a priori* algorithm for finding association rules, and a classifier builder. The candidate rule generator is similar to the *a priori* one. The difference is that CBA updates the support value in each step while the *a priori* algorithm only updates this value once. This allows us to compute the confidence of the *ruleitem*. A *ruleitem* is of the form:  $\langle \text{condset}, y \rangle$  where *condset* is a set of items,  $y \in Y$  is a class label. The support count of the *condset* (called *condsupCount*) is the number of cases in the dataset (*D*) that contain the *condset*.

Next, a classifier is built from CARs. To produce the best classifier out of the whole set of rules would involve evaluating all the possible subsets of it on the training data and selecting the subset with the right rule sequence that gives the least number of errors. There are  $2^m$  such subsets, where *m* is the number of rules. It is a heuristic

algorithm. Given two rules,  $r_i$  and  $r_j$ ,  $r_i$  precedes  $r_j$  if (1) the confidence of  $r_i$  is greater than that of  $r_j$ , or (2) their confidences are the same, but the support of  $r_i$  is greater than that of  $r_j$ , or (3) both the confidences and the supports of  $r_i$  and  $r_j$  are the same, but  $r_i$  is generated earlier than  $r_j$ . If  $R$  is a set of generated rules (i.e. CARs) and  $D$  the training data, the basic idea of the algorithm is to choose a set of high precedence rules in  $R$  to cover  $D$ . The classifier follows this format:  $\langle r_1, r_2, \dots, r_n, \text{default\_class} \rangle$ , where  $r_i \in R$ . In classifying an unseen case, the first rule that satisfies the case will classify it. If there is no rule that applies to the case, it takes on the default class.

The descriptors used with CBA need to be discrete in nature (Liu et al. 1998) as is the case with SG descriptors but not Dragon, MolConnZ or MOE. Hence, this method was only used with SG descriptors using CBA (v2.1) software (Liu et al. 2001).

**Distance Weighted Discrimination.** Distance weighted discrimination (DWD) was initially proposed by Marron and Todd (Marron et al. 2007) with the goal of improving the performance of SVM (Cristianini & Shawe-Taylor 2000, Vapnik 1995) in high dimensional low sample size (HDLSS) contexts. The main idea is to improve upon the criterion used for “separation of classes” in SVM. SVM has data piling problems along the margin, because it is maximizing the minimum distance to the separating plane, and there are many data points that achieve the minimum. A natural improvement is to replace the minimum distance by a criterion that allows all the data to have an influence on the result. DWD does this by maximizing the sum of the inverse distances. This results in directions that are less adversely affected by spurious sampling artifacts. The major contribution of this new discrimination method is that it avoids the data piling problem, to give the anticipated improved generality. Like SVM, DWD is based on

computationally intensive optimization; however, while SVM uses well known quadratic programming algorithms, DWD uses interior-point methods for so-called Second-Order Cone Programming (SOCP) problems (Alizadeh & Goldfarb 2003). Detailed discussion of these issues may be found in Marron and Todd (Marron et al. 2007), which is available with the supporting information at <https://genome.unc.edu/pubsup/dwd/>. All DWD computations were performed using the DWD software (Marron 2002) written in Matlab (Mathworks 2010) and kindly provided by Dr. Marron.

### **Balancing Datasets Using Similarity Searching**

Some of the classification datasets we are dealing with are highly imbalanced, i.e. one of the classes (e.g., non-binders) is much larger or smaller than the other class (binders). However, highly imbalanced datasets might affect the predictive performance of QSAR models negatively. Therefore, only a subset of the larger class of approximately the same size of the smaller class will be used for model building with some modeling techniques like  $k$ -nearest neighbors ( $k$ NN) that are highly sensitive to imbalanced datasets. This subset will be selected to include compounds from the larger class that are most similar to the compounds in the smaller class.

### **Model Selection and Validation**

Following our predictive QSAR modeling workflow (Tropsha 2010) (cf. Fig. 2.2), all QSAR models generated to build regression models for binding affinities or to classify binders vs. non-binders were validated by predicting both test and external validation sets and applying different validation criteria.

**Dataset Division for Model Building and Validation.** All QSAR models generated in this research will be validated by predicting external validation sets generated by: (1)

Randomly extracting 20% of the dataset using an in-house script, or (2) Extracting 5 different external validation sets using external 5-fold cross validation (CV) (Hawkins et al. 2003, Kohavi 1995). Datasets employed in QSAR studies were first randomly divided into a modeling set and an external validation set (evs). Another level of internal validation was achieved by dividing the modeling set into multiple chemically diverse training and test sets using the Sphere Exclusion algorithm implemented in our laboratory (Golbraikh & Tropsha 2002b). These routines are always employed as a part of our predictive QSAR modeling workflow to emphasize the fact that training-set-only modeling is not sufficient to obtain reliable models that are externally predictive. It should be mentioned that only models that are highly predictive on the test sets will be retained for the consensus prediction of the external validation sets. Finally, only those models that are shown to be highly predictive on both external sets will be used in consensus fashion for virtual screening of external compound libraries.

**Model Acceptability Criteria for Rigorous Predictor Development.** Several publications by our group have recommended a set of statistical criteria which must be satisfied by a predictive model (Golbraikh & Tropsha 2002b, Golbraikh & Tropsha 2002a, Golbraikh et al. 2003, Tropsha 2006, Tropsha & Golbraikh 2007, Tropsha 2010). For continuous QSAR, criteria that we will follow in developing activity/property predictors are as follows: (i) correlation coefficient  $R$  between the predicted and observed activities; (ii) coefficients of determination (predicted versus observed activities  $R$ , and observed versus predicted activities  $R_0^2$  for regressions through the origin); (iii) slopes  $k$  and  $k'$  of regression lines through the origin. We consider a QSAR model predictive, if

the following conditions are satisfied (i)  $q^2 > 0.6$ ; (ii)  $R^2 > 0.6$ ; (iii)  $\frac{(R^2 - R_0^2)}{R^2} < 0.1$  and  $0.85 \leq k \leq 1.15$  or  $\frac{(R^2 - R_0^2)}{R^2} < 0.1$  and  $0.85 \leq k' \leq 1.15$ ; (iv)  $|R_0^2 - R_0'^2| < 0.3$  where  $q^2$  is the cross-validated correlation coefficient calculated for the training set, but all other criteria are calculated for the test set.

For classification and category QSAR, a  $2 \times 2$  confusion matrix can be defined (see Table 3) in the case when compounds belong to two classes (e.g., active and inactive compounds), where  $N(1)$  and  $N(0)$  are the number of compounds in the data set that belongs to classes (1) and (0) respectively. TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. The following classification accuracy characteristics associated with confusion matrices are widely used in QSAR studies: sensitivity ( $SE = TP/N(1)$ ), specificity ( $SP = TN/N(0)$ ), and enrichment  $E = TP \cdot N / [(TP + FP) \cdot N(1)]$ . In this study, we have employed normalized confusion matrices. A normalized confusion matrix can be obtained from the non-normalized one by dividing the first column by  $N(1)$  and the second column by  $N(0)$ . Normalized enrichment is defined in the same way as  $E$  but is calculated using a normalized confusion matrix:  $En = 2TP \cdot N(0) / [TP \cdot N(0) + FP \cdot N(1)]$ .  $En$  takes values within the interval of  $[0, 2]$  (Golbraikh et al. 2002, Kovatcheva et al. 2004). The prediction accuracy (correct classification rate,  $CCR_{\text{train}}$ ) is calculated as in Equation 2.5.

**Table 2.1.** Confusion matrix for binary classification models.

	Observed class (1)	Observed class (0)	Total
Predicted class (1)	TP	FP	TP+FN
Predicted class (0)	FN	TN	FN+TN
Total	$N_{(1)} = TP+FN$	$N_{(0)} = FP+TN$	$N=TP+FP+FN+TN$



**Robustness of QSAR Models.** Y-randomization test is a widely used validation technique to ensure the robustness of a QSAR model (Wold & Eriksson 1995). This test will be used to evaluate all generated QSAR models in our research efforts. Applying this test includes (i) randomly shuffling the dependent-variable vector, Y-vector (class labels or actual activity values) of training sets and (ii) rebuilding models with the randomized activities (or class labels) of the training set. All calculations are repeated several times using the original independent-variable matrix. It is expected that the resulting QSAR classification models, built with randomized activities for the training set, should generally have low CCRs for training, test, and external validation sets. It is likely that sometimes, though infrequently, high CCR values may be obtained due to a chance correlation or structural redundancy of the training set. However, if some QSAR classification models obtained in the Y-randomization test have relatively high CCR it implies that an acceptable QSAR classification model cannot be obtained for the given dataset by the particular modeling method used. Y-randomization test will be applied to all datasets considered in this research, and the test will be repeated five times in each case.

### **Virtual Screening**

Our main goal from model building studies is the prediction of the target properties of the compounds in chemical libraries allowing for their immediate prioritization for subsequent experimental validation. Therefore, we are seeking to develop and deliver highly efficient yet accurate QSAR-based virtual screening and scoring protocols that will significantly increase the experimentally validated hit rate of virtual screening. Robust and externally predictive QSAR models generated for selected GPCRs were used for VS of chemical databases to predict new ligands for these targets. The predicted ligands could

become important biological probes or drug candidates. The final compendium of models could be potentially useful for predicting biological profiles and side effects of drugs. Our search methodologies will be based on chemical similarity estimated in two different ways: (1) global similarity based on all descriptors calculated for the modeling set and acts as a primary filter that will assure a some level of similarity of the predicted compounds to the modeling set, (2) model-based similarity where the predicted compounds possess the chemical features (descriptors) that have been chosen by the QSAR model after the variable selection process. Model-based similarity implies that few chemical features in the chemical structure in fact control the modeled biological property.

**Applicability Domain.** Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, if it is highly dissimilar from all compounds of the modeling set, reliable prediction of its activity is unlikely to be reached. The concept of applicability domain AD, previously implemented and widely used in our laboratory (Zhu et al. 2008a, Zhang et al. 2008a, Tropsha 2003), was applied to avoid unreliable prediction. In this study, we defined AD as a distance threshold  $D_T$  between a compound under prediction and its closest nearest neighbors of the training set. It was calculated as follows:

$$D_T = \bar{y} + Z\sigma \quad (\text{Eq. 2.9})$$

where,  $\bar{y}$  is the mean Euclidean distance between each compound and its  $k$ -nearest neighbors in the model space of the training set (i.e.,  $k$  is the parameter optimized during QSAR model generation, and the distances are calculated using descriptors selected by the optimized model only),  $\sigma$  is the standard deviation of these Euclidean distances, and  $Z$  is an arbitrary parameter. We set the default value of this parameter  $Z$  at 0.5, which formally places the

allowed distance threshold at the mean plus one-half of the standard deviation. We also defined the AD in the entire descriptor space, i.e., global AD. In this case, the same formula (9) is used,  $k=1$ ,  $Z=0.5$ , and Euclidean distances were calculated using all descriptors. Thus, if the distances of the external compound from its  $k$  nearest neighbors (see above) in the training set within either the entire descriptor space or the selected descriptor space exceeded these thresholds, no prediction was made.

**Consensus Prediction.** Our experience suggests that consensus prediction of the target property for external compounds, i.e., when the compound activity is calculated by averaging values predicted by all individual models that satisfy our acceptability criteria, always provides the most stable and accurate solution (de Cerqueira et al. 2006, Zhu et al. 2008a, Kovatcheva et al. 2004). The assumption being that averaging predicted activities of a compound over multiple predictive models cancels out the errors of prediction. In this research we will be averaging the predictions for each compound by majority voting for QSAR models, using all models passing the validation criteria ( $CCR_{\text{train}}$ ,  $CCR_{\text{test}}$  and  $CCR_{\text{ex}} \geq 0.70$  for classification models and  $q^2$  and  $R^2 \geq 0.70$  for continuous models). In order to determine the confidence in the obtained predictions we need to define a consensus score (CS) for each of the predicted hits first. The consensus score can be defined as the average predicted value of the target property by all models used for prediction. In this research we investigated the performance of different consensus scores when prioritizing hits for experimental testing.

### **Integrative Chemocentric Informatics Approach**

We have devised an integrative workflow focused on the discovery of new drug candidates and finding new uses for existing drugs by fusing predictions generated from

different data types and methods. Currently, the workflow (will be discussed in details in Chapter 5 as a new integrative methodology based on hypothesis fusion devised herein) incorporates three major components: (1) a module for QSAR-based VS of chemical libraries to identify new ligands for target proteins, (2) a network-mining module to identify small molecule therapeutics for specific diseases without necessarily knowing the underlying target-specific mechanism; this module explicitly relies on cmap(Lamb et al. 2006, Lamb 2007), an external online database ([www.broadinstitute.org/cmap/](http://www.broadinstitute.org/cmap/)) that links the effects of different drugs and diseases using gene expression profiles, and (3) ChemoText (Baker & Hemminger 2010), an in-house repository of relationships between chemicals, diseases, proteins, and biological processes. The first two modules have been employed extensively for studies reported herein. This new approach will be discussed in details in chapter 5.

## **Experimental Method**

**Radioligand Binding Assays.** This screen was performed by the National Institute of Mental Health Psychoactive Drug Screening Program (PDSP). Radioligands were purchased by PDSP from Perkin-Elmer or GE Healthcare. Competition binding assays were performed using transfected or stably expressing cell membrane preparations as previously described (Roth et al. 2002, Shapiro et al. 2003) and are available online (<http://pdsp.med.unc.edu>). All experimental details are available online (<http://pdsp.med.unc.edu/UNC-CH%20Protocol%20Book.pdf>).

**Chemistry.** Chemical compounds predicted as hits from the virtual screening were obtained from commercial suppliers according to their availability. All compounds were ordered to have  $\geq 95\%$  purity. Additionally, all compounds were subjected to purity assessment using LC/MS by the Center for Integrative Chemical Biology and Drug

Discovery at UNC-Chapel Hill. LC/MS spectra of all compounds were acquired from an Agilent 6110 Series system with UV detector set to 220 nm. Samples were injected (5 uL) onto an Agilent Eclipse Plus 4.6 x 50 mm, 1.8 uM, C18 column at room temperature. A linear gradient from 10% to 100% B (MeOH + 0.1% Acetic Acid) in 5.0 min was followed by pumping 100% B for another 2 minutes with A being H<sub>2</sub>O + 0.1% acetic acid. The flow rate was 1.0 mL/min.

## CHAPTER 3

### ***IN SILICO* RECEPTOROMICS: QSAR MODELING OF RECEPTOR SUBTYPES AND FAMILIES, MODEL APPLICATION FOR VIRTUAL SCREENING, AND EXPERIMENTAL VALIDATION OF COMPUTATIONAL HITS**

#### **Introduction**

*In silico* screening approaches are routinely employed nowadays in academic, governmental and commercial sectors and they have become broadly applied techniques in drug discovery (Armbruster & Roth 2005, Bajorath 2002, Bajorath 2005, Becker 2004, Becker et al. 2004, Evers & Klebe 2004, Kitchen et al. 2004, Klabunde et al. 2009). Research conducted in our group demonstrated that the generation of Quantitative Structure Activity Relationship (QSAR) models and subsequent model-based virtual screening of chemical libraries has led to the identification of chemically diverse molecules with high success rates in experimental validation tests (Hsieh et al. 2008, Oloff et al. 2005, Peterson et al. 2009, Shen et al. 2002, Shen et al. 2004, Tang et al. 2009, Tropsha 2006, Tropsha & Pearlman 2000, Tropsha & Wang 2006). This approach to drug discovery comprises the following steps: (1) defining the target(s) of interest, (2) extracting relevant structure activity data from the biological literature and specialized databases, (3) dataset curation, (4) compound representation by suitable chemical descriptors, (5) model generation and validation, (6) the application of validated QSAR models for mining chemical databases to predict binders and, if possible agonists and antagonists. Often, the interpretation of chemical descriptors found significant for the success of QSAR models can reveal important structural requirements for

ligand binding and activity. For the most part, this approach has been applied to datasets of compounds tested in individual assays characterizing their interaction with a single molecular target.

The growing understanding within molecular pharmacology of the complexity of biological networks and the robustness and redundancy of biological systems challenges the current approaches of single target drug discovery including *in silico* approaches (Hopkins 2007, Hopkins 2008, Roth et al. 2000, Roth & Kroeze 2006). Nowadays, medicinal chemists are becoming more interested in identifying polypharmacological drugs that can bind moderately to several targets in a disease-protein network and affect the overall outcome significantly (Hopkins et al. 2006, Hopkins 2009, Roth et al. 2004). Herein, we suggest receptor-family-based QSAR models as computationally inexpensive tools for the quick prioritization of polypharmacological hits for further experimental testing against a large panel of receptors.

Theoretically, QSAR models explore information restricted to the experimental knowledge of chemical structures and biological activities of ligands. Hence, this approach is especially important when the X-ray crystal structures for the biological targets of interest (e.g., most GPCRs and trans-membrane proteins) are unavailable. By applying QSAR modeling approach to a large number of datasets, we can accumulate a compendium of QSAR models representing a variety of different biological targets, and subsequently establish a virtual receptorome system to screen molecules simultaneously against an array of available models. Ultimately, we can use these models to obtain a list of common matching hits among several receptor families and could link the hits to all predicted biological targets, thereby enabling an *in silico* identification of biological networks that will possibly be

influenced by these compounds (concept explained in Fig. 1.1). This approach can also identify selective compounds for each receptor family after excluding common hits. An example of a similar approach to broad *in silico* profiling of compound libraries is given by the method PASS (Brady & Stouten 2000), which currently allows *in silico* screening against a large panel of target proteins. However, the datasets behind PASS models are not publicly available, which makes it difficult to employ and validate alternative techniques to the same datasets.

In this research, we have focused on G-protein Coupled Receptors (GPCRs) and related trans-membrane proteins (i.e., Sigma receptors) as promising targets for the drug discovery projects targeting polypharmacology. GPCRs constitute the largest family of membrane proteins that mediate most cellular responses to hormones and neurotransmitters and are also responsible for vision, olfaction and taste (Rosenbaum et al. 2009). The total number of currently known and verified human GPCRs consists of at least 799 unique full-length members (Gloriam et al. 2007). GPCRs have been involved in a multitude of biological responses in all organs and systems including the central and peripheral nervous systems. In the latter, particularly important functions include neurotransmitter release, cell-to-cell communication, modulation of learning and memory, response to psycho-active substances, regulation of neuronal growth and differentiation and of glial responses.

However, ligands for GPCRs comprise structurally very diverse compounds and often the ligands interact with more than one GPCR, i.e., they are promiscuous. Furthermore, most highly effective polypharmacological compounds (e.g., clozapine) are highly promiscuous as well and consequently, they often have serious side effects. Although polypharmacology is desired for treating many diseases, highly promiscuous compounds are sometimes very toxic.



Therefore, we should be investigating new approaches to identify moderately promiscuous compounds that have decent binding affinities to few highly desired receptors that belong to the same family or different families of receptors. For example, a compound that acts as a 5-HT<sub>2C</sub> agonist, 5-HT<sub>6</sub> and H<sub>3</sub> antagonist might be a good anti-obesity lead/drug. However, such compound would be a major health hazard if it activates 5-HT<sub>2B</sub> receptor subtype since the latter activity often leads to undesired cardiovascular effects (Huang et al. 2009, Setola & Roth 2005, Setola & Roth 2008). Thus, predicting ligand's binding to serotonin and histamine families of receptors is very crucial for activity, while binding to other receptor families of receptors (or receptor subtypes) would make the drug less safe.

In this regard, family based models where a compound is regarded as active if it interacts with at least one member of the family, and inactive if it shows no (or poor) binding affinity against all members of the family, can be useful tools to predict selective polypharmacological profiles where promising compounds can be identified and experimentally validated for both efficacy and safety, and then may be, modified to achieve better activity profiles against the desired pharmacologic targets. Moreover, the use of family specific datasets will increase the statistical rigor of the generated QSAR models for the following reasons: (1) increased size of datasets, (2) increased chemical diversity of datasets, (3) larger applicability domains of the models.

In this study, we have developed QSAR classification models for ligands interacting with several receptor subtypes and families of GPCRs and other trans-membrane proteins as part of an ongoing project in our lab; the GPCR QSARome project. The modeled receptor families included 5-Hydroxytryptamine (5-HT), adrenergic alpha, dopaminergic, histamine, muscarinic, and sigma receptors achieving external classification accuracies as high as 95 %.

All models were subjected to rigorous internal and external validation. The results confirmed the high external prediction accuracy of our computational models, which led us to conclude that these models can be used reliably to screen chemical databases to identify putative binders across receptor families. Thus, the models were used for virtual screening (VS) of two commercially available databases: the World Drug Index (WDI) (Daylight 2004) and DrugBank (Wishart et al. 2006, Wishart et al. 2008). Eleven VS hits from the WDI were subjected to parallel binding assays against a panel of trans-membrane protein targets. Nine compounds were found to bind to at least one receptor subtype among the predicted families with binding affinities between 0.6 - 9000 nM. Thus, these models will be highly valuable to assess the potential of chemicals to bind several families of GPCRs in an effort to predict interesting polypharmacological profiles.

## **Materials and Methods**

### **Databases and Datasets**

Four databases described in details in chapter 2 were used for purposes of the work presented herein, namely: PDSP  $K_i$ -DB, WDI, DrugBank, and PubChem. The experimental data for receptor family datasets (i.e., for 5-HT, adrenergic alpha, dopamine, histamine, muscarinic and sigma receptors) were extracted from the PDSP  $K_i$ -DB available in the public domain. We used WDI and DrugBank chemical databases for QSAR-based VS to identify putative ligands for the studied receptor families, while we used PubChem obtain all chemical structures for our datasets in SDF file format.

### **Preprocessing of the Datasets**

We used a workflow for chemical data curation that was developed in our lab and published recently (Fourches et al. 2010) and described in details in Chapter 2. We assigned

the ‘activity’ class for each compound based on its  $K_i$  value(s) obtained from the PDSP and according to PDSP specifications as reported at the PDSP website (<http://pdsp.med.unc.edu/>). Compounds with  $K_i$  values less than 10  $\mu\text{M}$  were considered binders and assigned to class 1, whereas compounds with  $K_i$  values more than or equal to 10  $\mu\text{M}$  were considered non-binders and assigned to class 0.

Binders to 5-HT family consisted of compounds that bind to any of the following receptors: 5-HT<sub>1A</sub>, 5-HT<sub>1B</sub>, 5-HT<sub>1D</sub>, 5-HT<sub>1E</sub>, 5HT<sub>2A</sub>, 5-HT<sub>2B</sub>, 5-HT<sub>2C</sub>, 5-HT<sub>5A</sub>, 5-HT<sub>6</sub>, or 5-HT<sub>7</sub>. Binders to the adrenergic alpha receptors consisted of compounds that bind any of the following receptor subtypes: alpha1A, alpha1B, alpha 2A, alpha 2B, or alpha 2C. Binders to the dopamine family consisted of compounds that bind any of the following receptor subtypes: D<sub>1</sub>, D<sub>2</sub>, D<sub>3</sub>, D<sub>4</sub> or D<sub>5</sub>. Binders to the histamine family consisted of compounds that bind any of the following receptor subtypes: Binders to the muscarinic family consisted of compounds that bind any of the following receptor subtypes: M<sub>1</sub>, M<sub>2</sub>, M<sub>3</sub>, M<sub>4</sub>, or M<sub>5</sub>. Binders to the sigma family consisted of compounds that can bind either sigma 1 or sigma 2 (or both) receptors.

### **Dataset Division for Model Building and Validation**

Previously, we and other groups demonstrated that, generally, there is no correlation between the statistical parameters of QSAR models for the training set, such as leave-one-out (LOO) cross-validation  $R^2(q^2)$ , and the correlation coefficient  $R^2$  between predicted and observed activities of the test set. This statement is also true for classification QSAR models: high classification accuracy for the training and the test set usually do not correlate with each other. Thus, acceptable statistics for the training set only is insufficient to assume that the model also has high external predictive power and QSAR models should be rigorously

validated using external validation sets of compounds which were not used to train the models (Golbraikh & Tropsha 2002a). Following our predictive QSAR modeling workflow (Tropsha 2010) all QSAR models generated to classify binders *vs.* non-binders across the studied receptor families were validated by predicting both test and external validation sets. Each dataset was randomly split into 5 different subsets of nearly equal size to allow for external 5-fold cross validation (CV) (Hawkins et al. 2003, Kohavi 1995). In this protocol, each subset including 20% of the original dataset was systematically employed as the external validation set while the remaining 80% of the compounds constituted the modeling set.

Another level of internal validation was achieved by comparing model performance for training and test sets. Herein, all modeling sets (each including 80% of the original dataset) were additionally divided multiple times into chemically-diverse training and test sets using the Sphere Exclusion program developed in-house and described elsewhere (Golbraikh & Tropsha 2002b). The Sphere Exclusion algorithm divides the modeling set into multiple pairs of training and test sets to guarantee that at least in the entire descriptor space, (i) all representative points of the test set are close to at least one representative point of the training set, i.e. test set compounds are within the applicability domain defined by the training set; (ii) given the relative sizes of the training and test sets, the highest portion of the representative points of the training set are close to representative points of the test set; (iii) and the training set is a representative subset of the entire modeling set, i.e., there is no subset in the modeling set not represented by a similar compound in the training set. Here, the modeling sets were divided 28-39 times into training and test sets of different sizes.

Multiple QSAR models were developed using these training sets and validated using the corresponding test sets. Models with high prediction accuracy assessed by statistical criteria (cf. Chapter 2) were used for consensus prediction of external validation set compounds: each compound was predicted by all models for which it was within the applicability domain (refer to Chapter 2 for details), and the consensus predicted value for each compound was rounded to the closest integer (class). The predictivity of the models was evaluated by the consensus CCR for the external validation set. The model building and validation approach is illustrated schematically in Figure 2.2.

## **Computational Methods**

**Dragon Descriptors.** An ensemble of 929 molecular descriptors was computed using the Dragon Professional software (version 5.4) (Dragon 2007) for all compounds (with explicit hydrogen atoms) in our datasets. Descriptors included: 0D-constitutional descriptors (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-topological descriptors, 2D-edge adjacency indices, 2D-Burden eigenvalues, and various molecular properties such as octanol-water partition coefficient. Descriptors with low variance (standard deviation lower than 0.0001) or missing values were removed. Furthermore, if the correlation coefficient between any two descriptors exceeded 95%, one of them was removed. The final set used in this QSAR study included 298 descriptors. These descriptors were range-scaled, so that their values were within the interval [0, 1]. Definition and calculation procedures for Dragon descriptors and the related references are given in the Handbook of Molecular Descriptors (Todeschini & Consonni 2000).

**Machine Learning Methods.** Both kNN classification algorithm and SVM were used to generate QSAR models herein. All details about these two methods can be found in Chapter 2.

### **Selection and Validation of QSAR Models**

To evaluate the predictive power of the generated QSAR models, CCR (Equation 2.5) values for the training, test, and external evaluation set were calculated. We used sensitivity (SE) and specificity (SP) (Equations 2.6 and 2.7) as well. SE and SP reflect the accuracy of predicting the compounds of active and inactive classes, respectively. We considered a QSAR model to have an acceptable predictive power, if both of the following conditions were satisfied:

- (i) CCR for the LOO cross-validation of the training set (i.e.,  $CCR_{LOO}$ ) was at least 65%, and CCR for the test set (i.e.,  $CCR_{test}$ ) was also at least 65%;
- (ii) For both training and test sets, SE and SP (i.e.,  $SE_{train}$ ,  $SE_{test}$ ,  $SP_{train}$ ,  $SP_{test}$ ) were at least 60%.

### **Applicability domain**

Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, if it is highly dissimilar from all compounds of the modeling set, reliable prediction of its activity is unlikely to be reached. The concept of applicability domain (AD), previously implemented and widely used in our laboratory (Zhu et al. 2008a, Zhang et al. 2008a, Tropsha 2003), was applied to avoid unreliable prediction. In this study, we defined AD as a distance threshold  $D_T$  between a compound under prediction and its closest nearest neighbors of the training set according to

Equation 2.9 (cf. Chapter 2). In all these studies we set the default value of this parameter  $Z$  at 0.5, which formally places the allowed distance threshold at the mean plus one-half of the standard deviation. We also defined the AD in the entire descriptor space. In this case, the same formula (Eq. 2.9) is used, where  $Z$  equals to 0.5, using one nearest neighbor (i.e.,  $k=1$ ), and Euclidean distances were calculated using all descriptors. Thus, if the distances of the external compound from its  $k$  nearest neighbors (see above) in the training set within either the entire descriptor space or the selected descriptor space exceeded these thresholds, no prediction was made.

In receptor family modeling efforts, diversifying datasets will generally increase the distance cutoff (DT) that is calculated using Equation 2.9, because we are increasing the average distance between each compound and its nearest neighbors in the training set. We set the default value of this parameter  $Z$  at 0.5. In this way the distance of the external compound from its nearest neighbors in the training set might become below the threshold and consequently we will be able to predict more compounds using our family-based models that were initially very distant from the training set compounds (i.e., out of applicability domain).

### **Robustness of QSAR models**

Y-randomization (randomization of response) is a widely used approach to validate the robustness of QSAR models. It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of the model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower values of CCR for the training or the test set than the models built using training set with real activities, or at least these models should not satisfy some of the validation criteria mentioned above. If this condition is not satisfied, models built for this training set with real

activities are not reliable and should be discarded. This test was applied to all training sets obtained by data splits into training and test sets and it was repeated three times for each split.

### **Virtual Screening and Consensus Prediction Thresholds**

We screened both the WDI and DrugBank databases to derive a predicted polypharmacology matrices for all compounds included in these databases and to identify common and unique binders for the six studied receptor families. Dragon descriptors were generated for each compound in the databases and normalized based on the maximum and minimum values of each descriptor in the modeling set. Each validated *k*NN-Dragon model was then used to predict the activities of compounds that were within AD. The results for each individual prediction were combined into a consensus prediction: a consensus score (CS) was calculated for each compound that was within the ADs of multiple models. The consensus scores employed in this study take into account the total number of models used to predict the compound's activity class (binder or non-binder), and the number of models that predicted the compound to belong to a specific class. Since we define two classes of compounds, i.e., class 1 (binders) and class 0 (non-binders), some models may predict a compound to belong to class 0 and others may predict it to belong to class 1. As a result, a consensus score between 0 and 1 will be obtained for each of the predicted compounds.

Additionally, Different Consensus Prediction Thresholds (CPTs) were then used to improve prediction accuracy. To clarify, each individual model could only make binary predictions of compounds as either active (value of 1) or inactive (value of 0). However, since we integrated predictions from the ensemble of models (that passed the acceptance criteria), we could have a situation when different models disagree in their predictions. Thus,



the averaged (consensus) predicted activity for each compound may be in the range between 0 and 1. Formally, compounds with a predicted activity higher than or equal to 0.5 are classified as active and those lower than 0.50 are classified as inactive. Obviously, the closer the average predicted value to 1 or 0 is, the higher the concordance among all models is and the higher is our confidence in annotating compounds as active or inactive, respectively. Thus, two additional thresholds reflecting this concordance among predictions could be established as a different type of the model applicability domain: for instance, selecting only external compounds with predicted activity above 0.90 or below 0.10 would limit the selection of compounds from virtual screening library to a set with higher confidence (but of course reduces the total number of compounds in the predicted set). Therefore, CPTs were employed in this study to select compounds with high prediction confidence: for instance, CPT 0.9/0.1 means: (i) compounds with predicted activity higher than the upper threshold (0.9) were classified as actives; (ii) compounds with activity lower than 0.1 were classified as inactives; and (iii) compounds with the average predicted activity between the two thresholds were not assigned to any class (inconclusive). The inconclusive compounds were not included in models' prediction accuracy calculation. Two different CPTs were tested in this study; CPT1 0.90/0.10 and CPT2 0.50/0.49 to analyze their impacts on models' predictivity.

The percentages of models that were used to make prediction for each compound in the virtual screening database were recorded as well. We hypothesize that the higher percentage of models that give the same prediction for a compound, the more likely the compound actually possesses this predicted activity; the smaller the prediction variance across all models, the more confidence we have that the predicted biological activity for this compound is accurate. For these reasons, a compound was selected as a hypothetical hit, if

and only if (i) it was predicted by at least 50% of the selected models (i.e., it was found within the ADs of these models) and (ii) among those models, at least 90% of them predicted this compound as active for CPT1 (or at least 50 % of them predicted this compound as active for CPT2).

### **Experimental Validation in Radiologand Binding Assays**

Final common hit compounds from QSAR-based VS across six receptor families were purchased and submitted to PDSP for experimental target validation. The experimental details are described in Chapter 2.

## **Results and Discussion**

### **QSAR Modeling for Receptor Subtypes**

***k*NN-*D*ragon.** *k*NN-*D*ragon models (classification and regression models) were generated for several receptor subtypes included in Table 3.1. First, a validation set (20% of the dataset) was excluded from each datasets randomly. The compounds in the remaining modeling set (80% of the original dataset) were divided into multiple training and test sets (28-35 divisions) using the Sphere Exclusion method implemented in our laboratory (Golbraikh & Tropsha 2002b). Multiple QSAR models were generated independently for all training sets and applied to the test sets. We accepted classification models with CCR values for both the training and test set greater than 0.70. We also accepted models with  $q^2$  and  $R^2$  values greater than or equal to 0.70. These models were used for the prediction of external validation sets. Model statistics (i.e.,  $CCR_{\text{evs}}$ ,  $R^2$ ) based on external sets are provided in Table 3.1. Additionally, results of the Y-randomization test confirmed that *k*NN classification models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  values above or equal to 0.70 were robust. Additionally, regression models with  $q^2$  and  $R^2$  above or equal 0.70 were also robust. None of the models

with randomized class labels of the training set compounds had  $CCR_{\text{rand}}$  above 0.54 for any dataset, and none of the regression models generated with randomized activities for training set activities had  $R^2$  above 0.50.

**Table 3.1.** QSAR Models for Selected GPCRs

<b>GPCR</b>	<b>End Point</b>	<b>No. Cps</b>	<b>Model Type</b>	<b>Accuracy</b>
5-HT <sub>2A</sub>	Binder/Non-binder	105/61	Classification	CCR=0.94
5-HT <sub>2B</sub>	Binder/Non-binder	148/607	Classification	CCR=0.74
5-HT <sub>2B</sub>	Agonist/Antagonist	33/115	Classification	CCR=0.84
5-HT <sub>6</sub>	Binder/Non-binder	97/79	Classification	CCR=0.92
5-HT <sub>6</sub>	Binding affinities	60	Continuous	R <sup>2</sup> =0.59
5-HT <sub>7</sub>	Binder/Non-binder	72/68	Classification	CCR=0.84
5-HT <sub>7</sub>	Binding affinities	62	Continuous	R <sup>2</sup> =0.60
Alpha <sub>2A</sub>	Binding affinities	74	Continuous	R <sup>2</sup> =0.67
Alpha <sub>2B</sub>	Binding affinities	73	Continuous	R <sup>2</sup> =0.62
Alpha <sub>2C</sub>	Binding affinities	76	Continuous	R <sup>2</sup> =0.65
D <sub>1</sub>	Binder/Non-binder	56/44	Classification	CCR=0.90
D <sub>2</sub>	Binder/Non-binder	56/58	Classification	CCR=0.85
D <sub>3</sub>	Binder/Non-binder	57/49	Classification	CCR=0.88
D <sub>4</sub>	Binder/Non-binder	60/51	Classification	CCR=0.92
D <sub>5</sub>	Binder/Non-binder	51/54	Classification	CCR=0.97

## QSAR Modeling to discriminate Actives vs. Inactives for individual Receptor Families

***k*NN-**Dragon**.** *k*NN-Dragon models were generated for the six receptor families mentioned earlier. First, a validation set (20% of the dataset) was excluded from each datasets randomly. The compounds in the remaining modeling set (80% of the original dataset) were divided into multiple training and test sets (24-30 divisions) using the Sphere Exclusion method implemented in our laboratory (Golbraikh & Tropsha 2002b). Multiple QSAR models were generated independently for all training sets and applied to the test sets. We accepted models with CCR values for both the training and test set greater than 0.90. These models were used for the prediction of external validation sets. Model statistics based on external sets are provided in Table 3.2. Additionally, results of the Y-randomization test confirmed that *k*NN classification models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  values above or equal to 0.70 were robust. None of the models with randomized class labels of the training set compounds had  $CCR_{\text{rand}}$  above 0.54 for any dataset.

**Table 3.2.** Performance of *k*NN classification methods to classify actives vs. inactives across six receptor families based on external validation set statistics.

Family	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>eva</sub> <sup>d</sup>
5-HT	282	17	12	11	11	1	6	0.65	0.92	1.77	1.44	0.78
Alpha <sup>e</sup>	1126	11	15	9	15	0	2	0.82	1.00	2.00	1.69	0.91
D <sup>f</sup>	619	12	14	11	13	1	1	0.92	0.93	1.86	1.84	0.92
H <sup>g</sup>	121	11	11	10	10	1	1	0.91	0.91	1.82	1.82	0.91
M <sup>h</sup>	297	10	13	8	11	2	2	0.80	0.85	1.68	1.62	0.82
Sigma	104	8	14	7	12	2	1	0.88	0.86	1.72	1.75	0.87

<sup>a</sup>Num. Mod, number of models with  $CCR_{train}$  and  $CCR_{test} \geq 0.90$  ( $\geq 0.85$  for 5-HT); <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>eva</sub>, correct classification rate of the consensus models using the external validation set; <sup>e</sup>Alpha, adrenergic alpha; <sup>f</sup>D, dopamine; <sup>g</sup>H, histamine; <sup>h</sup>M, muscarinic.

**SVM-Dragon.** Models were built for the six receptor families mentioned earlier. First, the dataset was divided into five parts. 80% of the original dataset was used as modeling sets and an external set included the remaining 20% of the dataset. The compounds in all modeling sets were divided into multiple training and test sets (28-40 divisions) using the Sphere Exclusion method (Golbraikh & Tropsha 2002b). Multiple QSAR models were generated independently for all training sets and applied to the test sets. We accepted models with CCR values for both the training and test set greater than 0.70. Then, we applied the accepted models for the prediction of external sets. Model statistics based on external validation sets are provided in Tables 3.3-3.8. Additionally, results of the Y-randomization test confirmed that *k*NN classification models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  values above or equal to 0.70 were robust. None of the models with randomized class labels of the training set compounds had  $CCR_{\text{rand}}$  above 0.50 for any dataset.

**Table 3.3.** Performance of SVM classification methods to classify actives vs. inactives across 5-HT receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	4	22	7	20	5	2	2	0.91	0.71	1.52	1.77	0.81
F2	12	16	13	14	11	2	2	0.88	0.85	1.70	1.74	0.86
F3	75	16	13	15	12	1	1	0.94	0.92	1.85	1.87	0.93
F4	281	15	13	13	12	1	1	0.87	0.92	1.84	1.87	0.89
F5	71	16	12	13	12	1	2	0.81	1.00	1.81	1.78	0.91
Average								0.88	0.88	1.74	1.81	0.88

<sup>a</sup>Num. Mod, number of models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.



**Table 3.4.** Performance of SVM classification methods to classify actives vs. inactives across adrenergic alpha receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	463	12	15	12	14	1	0	1.00	0.93	1.88	2.00	0.97
F2	503	12	14	11	14	0	1	0.92	1.00	2.00	1.85	0.96
F3	458	10	16	10	13	3	0	1.00	0.81	1.68	2.00	0.91
F4	445	17	9	17	8	1	0	1.00	0.89	1.80	2.00	0.94
F5	471	11	15	11	14	1	0	1.00	0.93	1.88	2.00	0.97
Average								0.98	0.91	1.85	1.97	0.95

<sup>a</sup>Num. Mod, number of models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.

**Table 3.5.** Performance of SVM classification methods to classify actives vs. inactives across dopamine receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	336	13	13	12	13	0	1	0.92	1.00	2.00	1.86	0.96
F2	442	11	15	11	13	2	0	1.00	0.87	1.76	2.00	0.93
F3	516	12	14	12	14	0	0	1.00	1.00	2.00	2.00	1.00
F4	407	12	14	10	13	1	2	0.83	0.93	1.84	1.70	0.88
F5	391	10	16	9	16	0	1	0.90	1.00	2.00	1.82	0.95
Average								0.93	0.96	1.92	1.87	0.95

<sup>a</sup>Num. Mod, number of models with CCR<sub>train</sub> and CCR<sub>test</sub>  $\geq$  0.70; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.

**Table 3.6.** Performance of SVM classification methods to classify actives vs. inactives across histamine receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	270	10	13	10	13	0	0	1.00	1.00	2.00	2.00	1.00
F2	117	13	9	12	9	0	1	0.92	1.00	2.00	1.86	0.96
F3	131	11	12	9	11	1	2	0.82	0.92	1.82	1.67	0.87
F4	258	10	12	10	12	0	0	1.00	1.00	2.00	2.00	1.00
F5	397	14	8	10	7	1	4	0.71	0.88	1.70	1.51	0.79
Average								0.89	0.96	1.90	1.81	0.92

<sup>a</sup>Num. Mod, number of models with  $CCR_{train}$  and  $CCR_{test} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.

**Table 3.7.** Performance of SVM classification methods to classify actives vs. inactives across muscarinic receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	205	12	11	10	9	2	2	0.83	0.82	1.64	1.66	0.83
F2	236	8	15	5	10	5	3	0.63	0.67	1.30	1.28	0.65
F3	70	6	17	6	15	2	0	1.00	0.88	1.79	2.00	0.94
F4	332	3	19	3	18	1	0	1.00	0.95	1.90	2.00	0.97
F5	244	5	17	5	16	1	0	1.00	0.94	1.89	2.00	0.97
Average								0.89	0.85	1.70	1.79	0.87

<sup>a</sup>Num. Mod, number of models with  $CCR_{train}$  and  $CCR_{test} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.

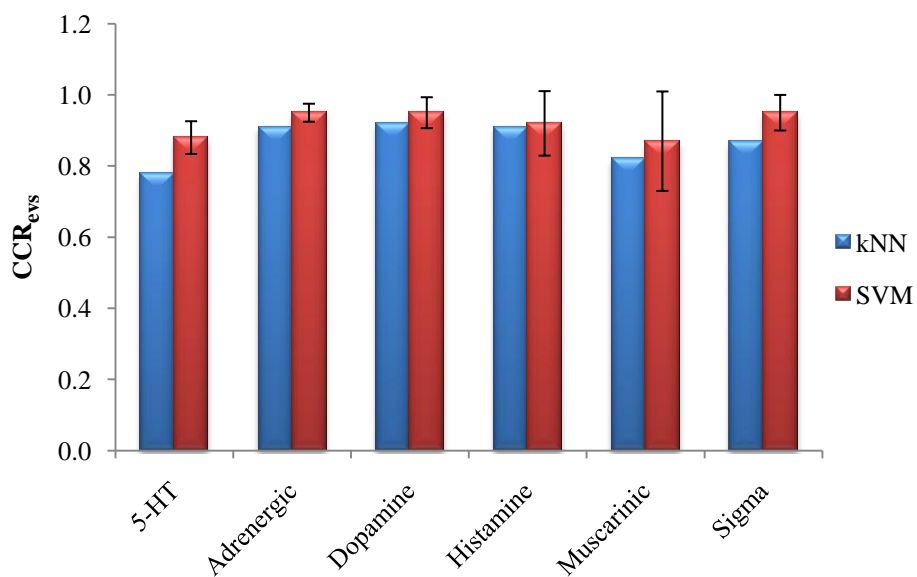
**Table 3.8.** Performance of SVM classification methods to classify actives vs. inactives across Sigma receptor family based on external validation set statistics.

Fold	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models				
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
F1	410	9	14	9	14	0	0	1.00	1.00	2.00	2.00	1.00
F2	290	9	14	9	13	1	0	1.00	0.93	1.87	2.00	0.96
F3	362	10	12	9	10	2	1	0.90	0.83	1.69	1.79	0.87
F4	380	6	16	6	15	1	0	1.00	0.94	1.88	2.00	0.97
F5	390	10	12	9	12	0	1	0.90	1.00	2.00	1.82	0.95
Average								0.96	0.94	1.89	1.92	0.95

<sup>a</sup>Num. Mod, number of models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set.

**Comparison between *k*NN and SVM Models.** The performance of *k*NN and SVM classification models for six receptor families, based on validation set statistics, is summarized in Figure 3.1. Both *k*NN and SVM combined with Dragon descriptors performed very well in classifying binders vs. non-binders across six receptor families based on external validation set statistics (Tables 3.1-3.8), yielding the highest  $CCR_{\text{evs}}$  of 0.95 for SVM classification models for adrenergic alpha, dopamine and sigma receptor families. Best performance for *k*NN models was for the dopamine receptor family with highest  $CCR_{\text{evs}}$  of 0.92 followed by *k*NN models for the adrenergic alpha and histamine receptor families with  $CCR_{\text{evs}}$  values of 0.91 in both cases.

**Figure 3.1.** Comparison of CCR values for the external validation set ( $CCR_{evs}$ ) for different QSAR models developed to classify binders vs. non-binders across six receptor families.



## Descriptor Analysis

Mechanistic interpretability is frequently regarded as very important feature of QSAR models. We generally argue that only models that have been extensively validated on external datasets and identified experimentally-confirmed hits should be subjected to interpretation. Furthermore, very few classes of models, specifically, those based on (multiple) linear regression and small number of descriptors can afford a relatively straightforward interpretation. The interpretation of multi-parametric statistical models developed with non-linear optimization algorithms (as in this study) should be attempted with great care because of strong and often poorly understood interplay between descriptors. Furthermore, although we could foresee that in some cases medicinal chemists may want to modify their candidate compounds to enhance or prevent ligand's binding to some receptor families of interest, the tools developed in this study are predominantly intended for virtual screening of libraries of drug candidates to predict compounds with interesting binding profiles and perhaps interesting polypharmacological effects. However, any compound designed by chemists could be passed through our models to predict its binding potential towards different receptor families.

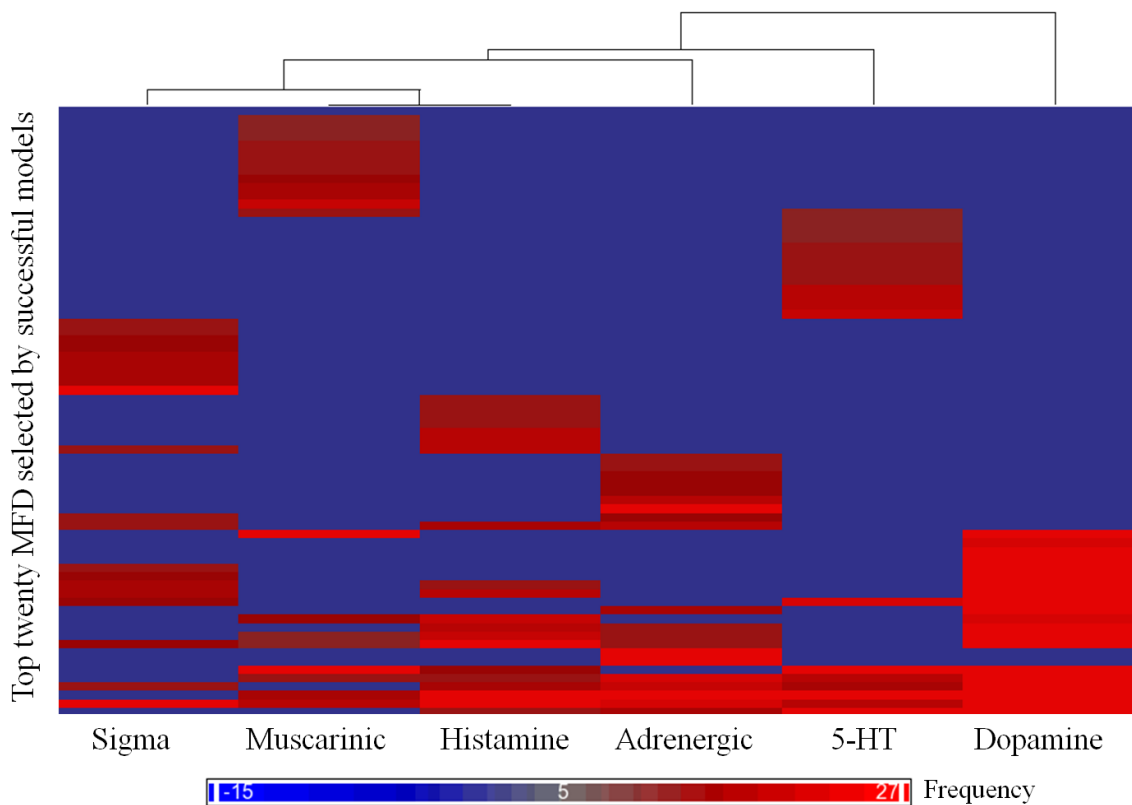
It needs to be pointed out that variable selection *k*NN QSAR optimizes the selection of a small number of descriptors to produce an acceptable QSAR model. By default, any successful QSAR model captures the correlation between variations in descriptor values and those in the target property. Thus, the significant correlation could be achieved with a small subset of descriptors. However, some other descriptors may serve as essential determinants of the compound pharmacological class but not be included in the model because of their low variances across the training set (cf. pharmacophoric groups that by default are the same for



all active compounds). Therefore, if one searches a database with a small number of variables selected by QSAR models, a similarity screen of the database using the entire pool of descriptors (global similarity) is necessary in addition to model-based activity prediction (i.e., so that not to miss any important structural features essential for biological activity).

We restricted the discussion in this paper to the most frequent descriptors found by all acceptable *k*NN models used for the prediction of external compounds to stress on the fact that the process of variable selection employed as part of model optimization has indeed converged on a small number of descriptors. From an initial pool of ca. 300 descriptors, only a small set of descriptors was selected for the acceptable QSAR models (see Fig. 3.2). Clearly, this small set of MFD was different for the different receptor families which is an indication that our models were distinct despite the fact that a large portion of the compounds contained in the different datasets was similar (i.e., many compounds were promiscuous with binding affinities to different receptor families). All details about top twenty MFD selected by successful models for all receptor families can be found in Tables 3.9-3.14.

**Figure 3.2.** The heatmap of descriptor frequencies across receptor families analyzed by hierarchical clustering of the pairwise similarities in descriptor frequencies using Euclidean distances and normalized frequencies of top twenty MFD. The bar-view is a key for coloring according to normalized descriptor frequency based on normalized descriptor frequencies where red color indicates most frequent descriptors while blue color denotes least frequent or unused descriptors.



**Table 3.9.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the 5-HT receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
N-074	140	R#N / R=N-	Atom-centered fragments
C-006	109	CH2RX	Atom-centered fragments
nArOR	102	Number of ethers (aromatic)	Functional group counts
nRNH2	72	Number of primary amines	Functional group counts
C-008	71	CHR2X	Atom-centered fragments
nNq	68	Number of quaternary N	Functional group counts
F-084	65	F attached to C1 (sp2)	Atom-centered fragments
nRCONR2	55	Number of tertiary amides (aliphatic)	Functional group counts
JGI3	53	Topological charge index of order 3	Topological charge indices
BELm3	49	Lowest eigenvalue n. 3 of Burden matrix	Burden eigenvalues
Depressant-50	47	Ghose-Viswanadhan-Wendoloski antidepressant-like index at 50%	Molecular properties
C-028	46	R--CR--X	Atom-centered fragments
nOHs	46	Number of secondary alcohols	Functional group counts
GATS7m	44	Geary autocorrelation - lag 7/weighted by atomic masses	2D autocorrelations
CIC0	44	Complementary information content (neighborhood symmetry of 0-order)	Information indices
H-047	43	H attached to C1(sp3) / CO(sp2)	Atom-centered fragments
C-005	43	CH3X	Atom-centered fragments
O-058	42	Corresponds to =O	Atom-centered fragments
GATS5e	42	Geary autocorrelation - lag 5/weighted by atomic Sanderson electronegativities	2D autocorrelations

MATS6p	40	Moran autocorrelation - lag 6 / weighted by atomic polarizabilities	2D autocorrelations
--------	----	---------------------------------------------------------------------------	---------------------

**Table 3.10.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the adrenergic alpha receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
nN+	401	Number of positively charged N	Functional group counts
C-006	358	CH <sub>2</sub> RX	Atom-centered fragments
nNq	309	Number of quaternary N	Functional group counts
H-047	257	H attached to C1(sp <sup>3</sup> ) / CO(sp <sup>2</sup> )	Atom-centered fragments
T(O..O)	254	Sum of topological distances between O..O	Topological descriptors
nTB	230	Number of triple bonds	Constitutional descriptors
nDB	217	Number of double bonds	Constitutional descriptors
O-058	215	Corresponds to =O	Atom-centered fragments
nRSR	211	Number of sulfides	Functional group counts
nArNR2	211	Number of tertiary amines	Functional group counts
F-084	205	F attached to C1 (sp <sup>2</sup> )	Atom-centered fragments
JGI3	191	Topological charge index of order 3	Topological charge indices
nO	175	Number of oxygen atoms	Constitutional descriptors
Depressant-50	173	Ghose-Viswanadhan-Wendoloski antidepressant-like index at 50%	Molecular properties
BLTF96	166	Verhaar model of fish baseline toxicity from MLOGP(mmol/l)	Molecular properties
MLOGP	164	Moriguchi octanol-water partition coefficient	Molecular properties
MLOGP2	163	Squared Moriguchi octanol-water partition coefficient	Molecular properties
GATS3m	162	Geary autocorrelation - lag 3/weighted by atomic masses	2D autocorrelations
N-071	159	Ar-NAI2	Atom-centered fragments
ARR	156	Atomic ratio	Constitutional descriptors

**Table 3.11.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the dopamine receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
Depressant-50	309	Ghose-Viswanadhan-Wendoloski antidepressant-like index at 50%	Molecular properties
BELm3	299	Lowest eigenvalue n. 3 of Burden matrix	Burden eigenvalues
nNq	259	Number of quaternary N	Functional group counts
nN+	247	Number of positively charged N	Functional group counts
nPyrrolidines	246	Number of pyrrolidines	Functional group counts
C-006	238	CH2RX	Atom-centered fragments
F-084	221	F attached to C1 (sp <sup>2</sup> )	Atom-centered fragments
O-058	206	Corresponds to =O	Atom-centered fragments
nTB	185	Number of triple bonds	Constitutional descriptors
C-008	183	CHR2X	Atom-centered fragments
H-047	182	H attached to C1(sp <sup>3</sup> ) / CO(sp <sup>2</sup> )	Atom-centered fragments
TPSA(NO)	172	Topological polar surface area using N, O polar contributions	Molecular properties
MLOGP2	171	Squared Moriguchi octanol-water partition coefficient	Molecular properties
nArOH	170	Number of ethers (aromatic)	Functional group counts
nO	167	Number of oxygen atoms	Constitutional descriptors
MATS6e	166	Moran autocorrelation - lag 6 / weighted by atomic Sanderson electronegativities	2D autocorrelations
nIsothiazoles	164	Number of isothiazoles	Functional group counts
Hypertens-80	159	Ghose-Viswanadhan-Wendoloski antihypertensive-like index at 80%	Molecular properties

T(O..O)	156	Sum of topological distances between O..O	Topological descriptors
nHDon	147	Number of doner atoms for H-bonds (N and O)	Functional group counts

**Table 3.12.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the histamine receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
nNq	37	Number of quaternary N	Functional group counts
Jhetv	31	Balaban-type index from van der Waals weighted distance matrix	Topological descriptors
JGI2	22	Mean topological charge index of order 2	Topological charge indices
IVDE	22	Mean information content on the vertex degree equality	Information indices
T(N..I)	21	Sum of topological distances between N..I.	Topological descriptors
nPyridines	21	Number of pyridines	Functional group counts
MLOGP2	21	Squared Moriguchi octanol-water partition coefficient	Molecular properties
nTB	20	Number of triple bonds	Constitutional descriptors
nN+	19	Number of positively charged N	Functional group counts
S-107	19	R2S / RS-SR	Atom-centered fragments
TPSA(NO)	19	Topological polar surface area using N, O polar contributions	Molecular properties
Hnar	19	Narumi harmonic topological index	Topological descriptors
Hypertens-80	19	Ghose-Viswanadhan-Wendoloski antihypertensive-like index at 80%	Molecular properties
MATS6e	18	Moran autocorrelation - lag 6 / weighted by atomic Sanderson electronegativities	2D autocorrelations
GATS7m	18	Geary autocorrelation - lag 7/weighted by atomic masses	2D autocorrelations
F-084	17	F attached to C1 (sp <sup>2</sup> )	Atom-centered fragments
T(N..N)	17	Sum of topological distances between N..N.	Topological descriptors
Mv	17	Mean atomic van der Waals volume (scaled on carbon atom)	Constitutional descriptors



nRSR	17	Number of sulfides	Functional group counts
JGI3	17	Topological charge index of order 3	Topological charge indices

**Table 3.13.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the muscarinic receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
nO	111	Number of oxygen atoms	Constitutional descriptors
nArOH	84	Number of ethers (aromatic)	Functional group counts
GATS8v	66	Geary autocorrelation - lag 8/weighted by atomic van der Waals volumes	2D autocorrelations
nNq	58	Number of quaternary N	Functional group counts
C-006	51	CH2RX	Atom-centered fragments
BLTD48	51	Verhaar model of Daphnia baseline toxicity from MLOGP (mmol/l)	Molecular properties
nR11	50	Number of 11-membered rings	Constitutional descriptors
nCrs	49	Number of ring secondary C(sp <sup>3</sup> )	Functional group counts
T(O..O)	47	Sum of topological distances between O..O	Topological descriptors
C-008	45	CHR2X	Atom-centered fragments
Psychotic-80	44	Ghose-Viswanadhan-Wendoloski antipsychotic-like index at 80%	Molecular properties
nN(CO)2	44	Number of imides (-thio)	Functional group counts
MATS1p	43	Moran autocorrelation - lag 1 / weighted by atomic polarizabilities	2D autocorrelations
O-062	42	O- (negatively charged)	Atom-centered fragments
O-060	41	Al-O-Ar / Ar-O-Ar / R..O..R / R-O-C=X	Atom-centered fragments
D/Dr09	40	distance detour ring index of order 9	Topological descriptors
GATS4m	40	Geary autocorrelation - lag 4/weighted by atomic masses	2D autocorrelations
nN+	40	Number of positively charged N	Functional group counts

GATS1v	39	Geary autocorrelation - lag 1/weighted by atomic van der Waals volumes	2D autocorrelations
H-047	39	H attached to C1(sp3) / CO(sp2)	Atom-centered fragments

**Table 3.14.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify binders vs. non-binders against the sigma receptor family.

Descriptor	Frequency	Interpretation	Descriptor Category
Depressant-50	68	Ghose-Viswanadhan-Wendoloski antidepressant-like index at 50%	Molecular properties
nO	35	Number of oxygen atoms	Constitutional descriptors
C-006	28	CH2RX	Atom-centered fragments
TPSA(NO)	26	Topological polar surface area using N, O polar contributions	Molecular properties
RBN	23	Number of rotatable bonds	Constitutional descriptors
nH	21	Number of hydrogen atoms	Constitutional descriptors
C-008	21	CHR2X	Atom-centered fragments
C-020	20	=CX2	Atom-centered fragments
BLI	20	Kier benzene-likeness index	Topological descriptors
nNq	20	Number of quaternary N	Functional group counts
F-084	19	F attached to C1 (sp2)	Atom-centered fragments
MATS7e	16	Moran autocorrelation - lag 7 / weighted by atomic Sanderson electronegativities	2D autocorrelations
CIC1	16	Complementary information content (neighborhood symmetry of 1-order)	Information indices
MSD	15	Mean square distance index (Balaban)	Topological descriptors
EEig13d	15	Eigenvalue 13 from edge adj. matrix weighted by dipole moments	Edge adjacency indices
H-052	15	H attached to CO(sp3) with 1X attached to next C	Atom-centered fragments
C-040	14	R-C(=X)-X / R-C#X / X=C=X	Atom-centered fragments
GATS5m	14	Geary autocorrelation - lag 5/weighted by atomic masses	2D autocorrelations

GATS8m	14	Geary autocorrelation - lag 8/weighted by atomic masses	2D autocorrelations
T(F..F)	14	Sum of topological distances between F..F.	Topological descriptors

We were also able to identify several common MFD across almost all six receptor families. These descriptors included: nN+, the number of positively charged nitrogen atoms (functional group counts); C-006, CH2RX (atom centered fragments); nNq, number of quaternary nitrogens (functional group counts); T(O..O), sum of topological distances between O..O (Topological descriptors); nO, number of oxygen atoms (constitutional descriptors); O-058, corresponds to =O (atom centered fragments); MLOGP, Moriguchi octanol-water partition coefficient (molecular properties). See Tables 3.9-3.14 for further details.

### **Virtual Screening and Experimental Validation**

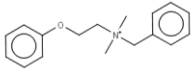
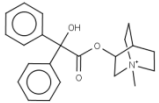
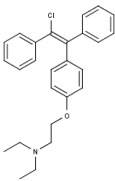
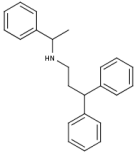
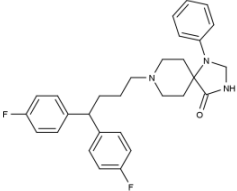
Our primary aim from generating family-based models is to provide a fast, large scale system that allows for virtual activity profiling across receptor families to predict polypharmacology profiles and consequently establish virtual biological networks. Since our models proved to be reasonably accurate based on external validation set statistics, we used the best models to mine a large external database of approved and potential drugs for putative binders to six receptor families. An important condition that assures reliable predictions by the model is the use of AD. Therefore, two types of AD were employed in the virtual screening of compound databases. The first is a local AD which is defined for each of the individual classification models and using a z threshold of 0.5. The second is a global AD that acts as a filter and ensures some level of global similarity between the predicted compounds and the compounds in the modeling set. Herein, we kept the latter AD as an optional filter because we wanted to explore a larger and more diverse set of compound.

In an attempt to identify putative binders across all six receptor families (5-HT, adrenergic alpha, dopamine, histamine, muscarinic and sigma), validated consensus *k*NN and

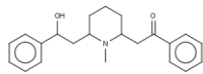
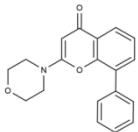
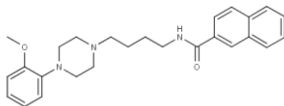
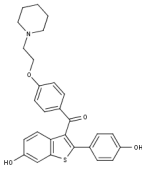
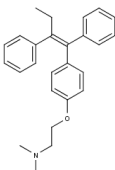
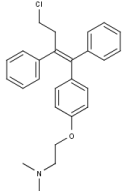
SVM models for six different receptor families were used for virtual screening of 59000 and 4678 molecules within the WDI and DrugBank chemical libraries subsequently. Refer to scheme 1 for steps of VS process.

Eleven structurally diverse hits (**1-11**, see Table 3.15) were selected from the final consensus virtual screening hits of the WDI for further experimental validation taking into account both their commercial availability and cost. We also took into account that these eleven hits were predicted as binders for almost all receptor families. All eleven hits were tested at the PDSP in radioligand binding assays across all receptor subtypes of the six receptor families mentioned above. See Table 3.16 for details.

**Table 3.15.** Eleven VS hits selected from WDI Chemical library.

Cp. ID	Structure/Name	PDSP ID	Therapeutic Class/Use
1		14816	Anthelmintic
2	Bephenium 	14817	An anticholinergic drug.
3	Clidinium 	13499	SERM
4	Clomifene 	14821	Calcium channel blocker
5	Fendiline 	14815	Antipsychotic
	Fluspirilene		



6		14824	VMAT2 ligand, it also inhibits the reuptake of dopamine and serotonin, acts as a mixed agonist-antagonist at nicotinic acetylcholine receptors, and an antagonist at $\mu$ -opioid receptors.[
7	Lobeline 	14825	Morpholino derivative of quercetin . It is a potent inhibitor of phosphoinositide 3-kinase s (PI3Ks)
8	LY-294002 	14814	Drug used in scientific research which acts as a moderately selective dopamine D <sub>3</sub> receptor partial agonist.
9	Prestwick-559 	13505	SERM
10	Raloxifene 	13506 678 10572	SERM
11	Tamoxifen 	16514	SERM
	Toremifene		

**Table 3.16.** Experimental validation results for the 11 computational hits predicted to be ligands to six families of receptors as a result of QSAR-based mining of the WDI chemical screening library (see text for abbreviations).

Compound	5-HT			Adrenergic			Dopamine		
	kNN CS	SVM CS	Exp.	kNN CS	SVM CS	Exp.	kNN CS	SVM CS	Exp.
Bephenium	0.94	1.00	NB	0.93	0.96	B	0.87	0.99	B
Clidinium	0.84	0.88	NB	0.94	0.94	NB	0.71	0.69	NB
Clomiphene	0.97	0.98	B	0.95	1.00	B	0.96	0.94	B
Fendiline	0.86	0.95	B	0.93	0.99	B	0.87	0.98	B
Fluspirilene	0.96	0.98	B	0.99	1.00	B	0.94	0.87	B
Lobeline	0.73	0.98	B	0.88	1.00	B	0.75	0.94	B
LY-294002	0.79	1.00	B	0.69	0.99	NB	0.64	0.91	NB
Prestwick-559	0.97	1.00	B	1.00	1.00	B	0.84	1.00	B
Raloxifene	0.90	1.00	B	0.90	1.00	B	0.87	0.98	B
Tamoxifen	0.99	1.00	B	0.95	1.00	B	0.99	1.00	B
Toremifene	0.99	1.00	B	0.96	1.00	B	0.98	1.00	B
kNN Accuracy1 <sup>a</sup>		82%			82%			82%	
kNN Accuracy2 <sup>b</sup>		86%			88%			100%	
SVM Accuracy1 <sup>c</sup>		82%			82%			82%	
SVM Accuracy2 <sup>d</sup>		90%			82%			91%	
Compound	Histamine			Muscarinic			Sigma		
	kNN CS	SVM CS	Exp.	kNN CS	SVM CS	Exp.	kNN CS	SVM CS	Exp.
Bephenium	0.94	0.99	NB	0.48	0.98	NB	0.84	0.99	B
Clidinium	0.77	0.84	B	0.14	0.82	B	0.72	0.78	B
Clomiphene	0.92	0.97	B	0.81	0.97	B	0.95	0.96	B
Fendiline	0.95	0.97	B	0.69	0.98	B	0.85	0.98	B
Fluspirilene	0.93	0.95	B	0.84	0.94	B	1.00	0.92	B
Lobeline	0.97	0.97	B	0.46	0.97	B	0.71	0.96	B
LY-294002	0.64	0.97	NB	0.34	0.96	NB	0.74	0.95	NB
Prestwick-559	0.97	1.00	B	0.66	1.00	B	1.00	1.00	B
Raloxifene	0.89	0.99	B	0.52	0.99	B	0.86	0.99	B
Tamoxifen	0.94	1.00	B	0.75	1.00	NB	0.90	1.00	B
Toremifene	0.93	1.00	B	0.77	1.00	B	0.95	1.00	B
kNN Accuracy1		82%			72%			91%	
kNN Accuracy2		88%			100%			100%	
SVM Accuracy1		82%			72%			91%	
SVM Accuracy2		82%			72%			91%	

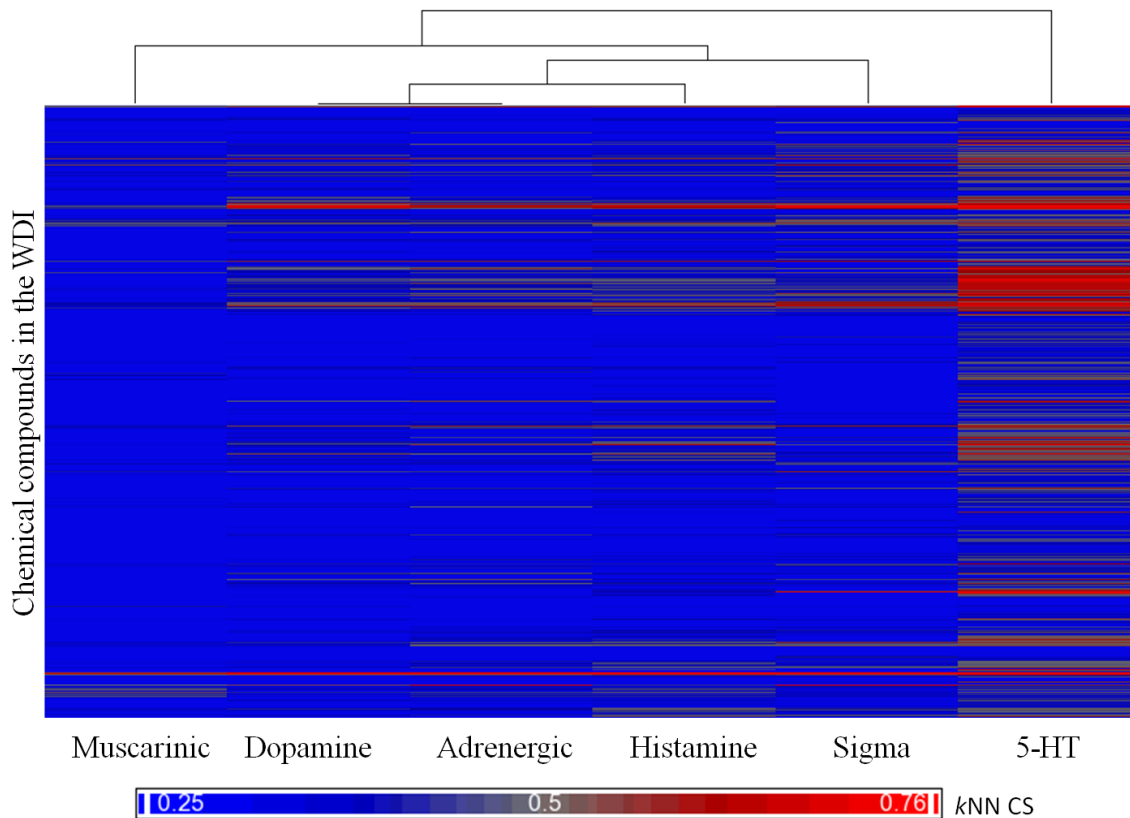
<sup>a</sup>*k*NN Accuracy1, accuracy of consensus *k*NN models at a CPT of 0.51/0.49; <sup>b</sup>*k*NN Accuracy2, accuracy of consensus *k*NN models at a CPT of 0.90/0.10; <sup>c</sup>SVM Accuracy1, accuracy of consensus SVM models at a CPT of 0.51/0.49; <sup>d</sup>SVM Accuracy2, accuracy of consensus SVM models at a CPT of 0.90/0.10.

Experimental validation results indicated that both *k*NN and SVM combined with Dragon descriptors performed very well on almost all compounds and across all receptor family models with predictions accuracies ranging from 82% to 100%. In case of 5-HT family models both *k*NN and SVM achieved overall prediction accuracies of 82% and 86 % applying CPTs of 0.51/0.49 and 0.90/0.10 consecutively. Similarly, the prediction accuracy of adrenergic alpha models was 82% for both *k*NN and SVM with a CPT of 0.50/0.49, while applying a CPT of 0.90/0.10 increased the accuracy of *k*NN predictions up to 88% (SVM prediction accuracy remained unchanged). Generally, applying a strict CPT of 0.90/0.10 improved the prediction accuracy of our models especially in case of *k*NN models. See Table 14 for comparison between model predictions and experimental validation results for all eleven VS hits across the six receptor families.

### **Biological Relevance**

In addition to using the models for virtual screening to prioritize hits for further experimental validation (see section about virtual screening and experimental validation), we predicted a full polypharmacology matrix for 46079 compounds in the WDI database (see Fig. 3.3). Accumulating such matrices using different computational tools developed by all many computational groups and making them publicly available, will shape the future of *in silico* receptoromics; we can compare the binding/activity potential for a compound against a specific receptor or receptor family by comparing all predictions for this compound across a multitude of computational tools. A consensus prediction generated from all these predictors is more likely to hold the accurate answer/prediction.

**Figure 3.5.** The heatmap of  $k$ NN CS for 46079 compounds in the WDI across six receptor families analyzed by hierarchical clustering of the pairwise similarities in descriptor frequencies using Euclidean distances and CSs. The bar-view is a key for coloring according to CS where red color indicates CS greater than 0.50 (i.e., binders/actives) while blue color denotes least CS < 0.50 (i.e., non-binders/inactives).



## Conclusions

QSAR models are becoming increasingly attractive as robust computational tools for virtual screening due to both their computational efficiency and success rates [reviewed in (Tropsha & Golbraikh 2007) as well as in a recent monograph (Varnek & Tropsha 2008)]. In this study, we have developed internally validated and externally predictive QSAR models for the classification of compounds into binders and non-binders across six different receptor families (5-HT, adrenergic alpha, dopaminergic, histamine, muscarinic, and sigma). We have shown that by using the *k*NN and SVM modeling strategies (combined with Dragon descriptors) as well as CPTs, it is possible to develop QSAR models with high external prediction accuracy.

The analysis of most frequent descriptors selected by QSAR models helps interpret the binding affinity to different receptor families in terms of chemical features. For example, we found that some functional group descriptors were frequently used in accepted *k*NN-Dragon models across almost all six receptor families, suggesting they may play a critical role in defining binding affinities of organic compounds to biogenic amine GPCRs and sigma receptors. These descriptors included: the number of positively charged nitrogen atoms, the number of quaternary nitrogens, the sum of topological distances between O..O, the number of oxygen atoms (constitutional descriptors); and octanol-water partition coefficients.

Encouraged by our model validation results, we have applied these models for virtual screening of the 59000 compounds in WDI database and 4678 compounds in DrugBank. Eleven structurally diverse VS hits were prioritized and experimentally tested at PDSP in radioligand binding assays. Nine compounds were found to bind to at least one receptor subtype among the predicted families with binding affinities between 0.6 - 9000 nM. Thus,

these models will be highly valuable to assess the potential of chemicals to bind several families of GPCRs in an effort to predict interesting polypharmacological profiles.

These predictors will be made publicly available at the ChemBench server established in the Laboratory for Molecular Modeling ([chembench.mml.unc.edu](http://chembench.mml.unc.edu)).



**CHAPTER 4**

**THE 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**

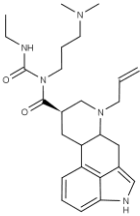
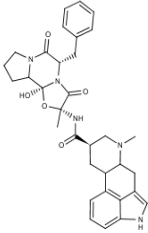
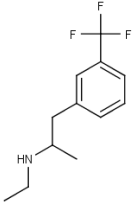
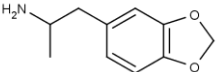
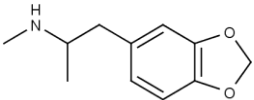
**Introduction**

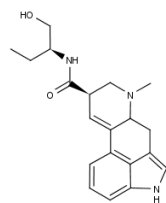
During the last decade, several drugs have been shown to cause cardiac valvulopathy in humans. The initial discovery of drug-induced valvulopathy occurred when the anorectic drug fenfluramine (approved by the FDA in 1973), one of the active ingredients of the anorectic drug combination fen-phen, was found to increase the risk of developing two potentially serious conditions, pulmonary hypertension and valvular heart disease (VHD), in individuals receiving these medications to treat obesity (Connolly et al. 1997). More recently, a group at the Mayo Clinic reported VHD in patients taking the anti-Parkinson drug pergolide (Pritchett et al. 2002). After the initial 2002 report, other cases of VHD associated with pergolide or other dopamine agonists such as cabergoline used as anti-parkinsonian drugs were identified (Peralta et al. 2006, Yamamoto et al. 2006, Yamamoto & Uesugi 2007). In January of 2007, the New England Journal of Medicine published two large European studies that independently verified the association of VHD with pergolide and carbergoline (Schade et al. 2007, Zanettini et al. 2007). Finally, on March 29, 2007, the Food and Drug Administration issued a Public Health Advisory for the voluntary market

withdrawal of pergolide. These stunning withdrawals of drugs from the market stressed the importance of elucidating the mechanism by which these drugs induce valvulopathy and of determining the valvulopathic risk that may be associated with new drug candidates or even existing drugs.

To date, all but two of the VHD-associated drugs are ergoline derivatives (dihydroergotamine, methysergide, pergolide and carbergoline) (see Table 1). The two non-ergoline VHD-associated drugs are fenfluramine (Connolly et al. 1997) and 3,4-methylenedioxymethamphetamine (MDMA, ecstasy) (Droogmans et al. 2007, Setola et al. 2003), both of which are amphetamine analogues (see Table 4.1). Thus, it appears that compounds from both the ergoline and phenylisopropylamine families can produce VHD (Setola & Roth 2008).

**Table 4.1.** Chemical structures of marketed drugs known as 5-HT<sub>2B</sub> receptor agonists and associated with VHD.

Compound Structure	Name PubChem CID	5-HT <sub>2B</sub> Agonist	VHD
	Carbergoline 54746	Yes	Yes
	Dihydroergotamine 10531	Yes	Yes
	Fenfluramine 3337	Yes	Yes
	MDA 1614	Yes	?? <sup>a</sup>
	MDMA 1615	Yes	Yes

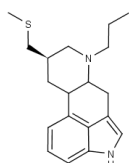


Methylergonovine

Yes

Yes

8226



Pergolide

Yes

Yes

47811

---

<sup>a</sup>Unknown.

There is increasing evidence that activation of serotonin 2B receptors (5-HT<sub>2B</sub>) may play a significant role in the pathogenesis of drug-induced valvulopathy (Rothman et al. 2000, Roth 2007, Berger et al. 2009). For instance, VHD-associated drugs such as fenfluramine (Setola & Roth 2005), ergotamine (Setola & Roth 2005), pergolide (Newman-Tancredi et al. 2002, Setola et al. 2003) and cabergoline, and/or selected active metabolites (such as norfenfluramine and methylergonovine) (Setola & Roth 2005), all potently activate 5-HT<sub>2B</sub> receptors. Chemically similar medications that do not activate 5-HT<sub>2B</sub> receptors (e.g., lisuride) seemingly do not cause valvular heart disease, further implicating the 5-HT<sub>2B</sub> receptor—but not other receptors that bind ergopeptines/ergolines and phenylisoproylamines with high affinity—in the pathogenesis of heart-valve disease (Roth 2007).

Additionally, valvulopathy-associated drugs have been shown to induce DNA synthesis in cultured interstitial cells from human cardiac valves via 5-HT<sub>2B</sub> receptor activation (Setola et al. 2003). It has been suggested that the valvulopathy induced by 5-HT<sub>2B</sub> receptor agonists is caused by the inappropriate mitogenic stimulation of normally quiescent valve cells, resulting in an overgrowth valvulopathy (Roth 2007, Setola et al. 2003). Although the precise signaling pathways underlying drug-induced valvulopathy remain elusive, 5-HT<sub>2B</sub> receptors are known to activate mitogenic pathways through the phosphorylation of Src kinase and extracellular regulated kinases (ERK), as well as through receptor tyrosine kinase transactivation (Nebigil et al. 2000a, Nebigil et al. 2000b), consistent with a role in regulating heart valve interstitial cell proliferation.

The discoveries that 5-HT<sub>2B</sub> receptors were (1) abundantly expressed in heart valves (Fitzgerald et al. 2000), (2) activated by fenfluramine and its metabolite, norfenfluramine (Fitzgerald et al. 2000, Rothman et al. 2000), and (3) activated by other valvulopathy-

inducing drugs (Rothman et al. 2000, Setola et al. 2003) suggested that 5-HT<sub>2B</sub> receptors were involved in the etiology of valvulopathy (Fitzgerald et al. 2000, Rothman et al. 2000). Subsequently, several other 5-HT<sub>2B</sub> agonists were also found to be valvulopathogenic (Setola et al. 2003). Since 5-HT<sub>2B</sub> agonists have the potential of causing valvulopathic side-effects, it has been suggested that all pharmaceuticals should be screened for activity at 5-HT<sub>2B</sub> receptors prior to further commercial development (Levy 2006, Roth 2007).

Similar to experimental high throughput screening (HTS), virtual screening (VS) is typically employed as a ‘hit’ identification tool (Stahura & Bajorath 2004). The experimental screening of all molecules against all biological targets is generally cost- and time-prohibitive. Therefore, pre-selection of compounds by VS that have a reasonable probability to act against a given biological target is highly attractive. Typically, VS approaches imply the use of structure based methodologies; nevertheless, we have repeatedly advocated for the use of ligand based cheminformatics approaches such as QSAR models in virtual screening (reviewed in a recent monograph (Varnek & Tropsha 2008)).

Herein, we report on the development of *in silico* screening tools for identifying compounds with potentially serious valvulopathic side effects. These tools can be employed as filters to flag and de-select the potentially harmful compounds at the preclinical stage of drug development, thereby potentially avoiding significant economic and human health consequences incurred at later stages of drug discovery. To achieve this goal, validated and externally predictive, binary QSAR models were generated for 5-HT<sub>2B</sub> active *vs.* inactive compounds as defined in 5-HT<sub>2B</sub> functional assays. Similar studies to develop QSAR models for 5-HT<sub>2B</sub> actives *vs.* inactives were reported recently by Chekmarev *et al* (Chekmarev et al. 2008). However, in our investigations we considered a larger dataset that contained the most

complete set of all known valvulopathogens reported by Huang *et al* (Huang et al. 2009) and we validated our predictions experimentally in binding assays.

To obtain the most statistically robust and predictive models, we have employed the combinatorial QSAR strategy (de Cerqueira et al. 2006, Kovatcheva et al. 2004) implemented as part of our predictive QSAR modeling workflow (reviewed in Tropsha and Golbraikh (Tropsha & Golbraikh 2007)). All models were subjected to rigorous internal and external validation. The results confirmed the high external prediction accuracy of our computational models, which led us to conclude that these models can be used reliably to screen chemical databases to identify putative 5-HT<sub>2B</sub> actives. Screening the WDI database using these models led to the identification of 122 possible 5-HT<sub>2B</sub> actives; 10 of these computational hit compounds were experimentally tested in 5-HT<sub>2B</sub> radioligand binding assays at the NIMH Psychoactive Drug Screening Program (PDSP), UNC Chapel Hill (<http://pdsp.med.unc.edu/>). Experiments confirmed that 9 out of 10 compounds were true actives implying a hit rate of 90%. These results indicate the reliability of our computational models as efficient predictors of compounds' affinity towards 5-HT<sub>2B</sub> receptors. We suggest that the computational models developed in this study could be used as drug liability predictors similar to commonly used predictors (Mohan et al. 2007, Simon-Hettich et al. 2006) of other undesired side effects such as carcinogenicity (Benfenati et al. 2009, Ruiz et al. 2008, Venkatapathy et al. 2009), mutagenicity (Benfenati et al. 2009, Papa et al. 2008, Zhang et al. 2008b), PGP binding (de Cerqueira et al. 2006), or hERG binding (Ekins et al. 2002, Garg et al. 2008, Seierstad & Agrafiotis 2006, Yoshida & Niwa 2006). Our models can be used to flag compounds that are expected to bind to 5-HT<sub>2B</sub> receptors but they cannot distinguish agonists from antagonists. Nevertheless, as demonstrated in this study, these

putative 5-HT<sub>2B</sub> binders can be tested in functional assays for their potential to activate 5-HT<sub>2B</sub> receptors to further assess their valvulopathic potential.

## **Materials and Methods**

### **Dataset**

The PDSP recently screened roughly 2200 FDA-approved drugs and investigational, drug-like molecules against 5-HT<sub>2B</sub> receptors (Huang et al. 2009). However, this modeling study was initiated prior to the completion of the screening of the entire compound library. At the time this study began, screening against 5-HT<sub>2B</sub> receptors had been completed for 800 compounds. This set became the basis for our model development. After preprocessing of the 800-compound dataset and deleting duplicates, the final dataset consisted of a class of 146 ‘actives’, and another class of 608 ‘inactives’. Detailed PDSP protocols are available online (<http://pdsp.med.unc.edu/>) and in Huang *et al* (Huang et al. 2009). All chemical structures were obtained from PubChem (PubChem 2009) as SDF files. By the time our modeling studies were completed, functional data for the remainder of the 2200 compounds (1400 compounds) had become available. These ‘new’ data became a source for additional, independent validation sets.

### **Preprocessing of the Dataset**

For the purposes of this work, the data were curated as follows: First, all molecules were “washed” using the Wash Molecules tool in MOE (MOE 2008) (v.2007.09). Using this tool, we processed 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 are found in the MOE manual (MOE 2008)). Second, we used ChemAxon Standardizer (ChemAxon JChem 2010) to harmonize the representation of aromatic rings. Finally, the analysis of the normalized molecular structures resulted in detection of 46 duplicate compounds (i.e., different salts or isomeric states). The functional data for duplicated compounds were found to be identical, so in each case a single example was removed. The curated subset of the original 5-HT<sub>2B</sub> dataset used in this work contains 754 unique organic compounds (146 actives and 608 inactives). All details about the dataset are available in Supporting Information.

### **Dataset Division for Model Building and Validation**

All QSAR models generated in this study to classify actives *vs.* inactives were validated by predicting two external validation sets. Each dataset employed in QSAR studies was first randomly divided into a modeling and a validation sets. Additionally, as described above, an independent validation set became available after we completed our modeling studies. Details about this external set are available in Supporting Information, and in Huang *et al* (Huang *et al.* 2009).

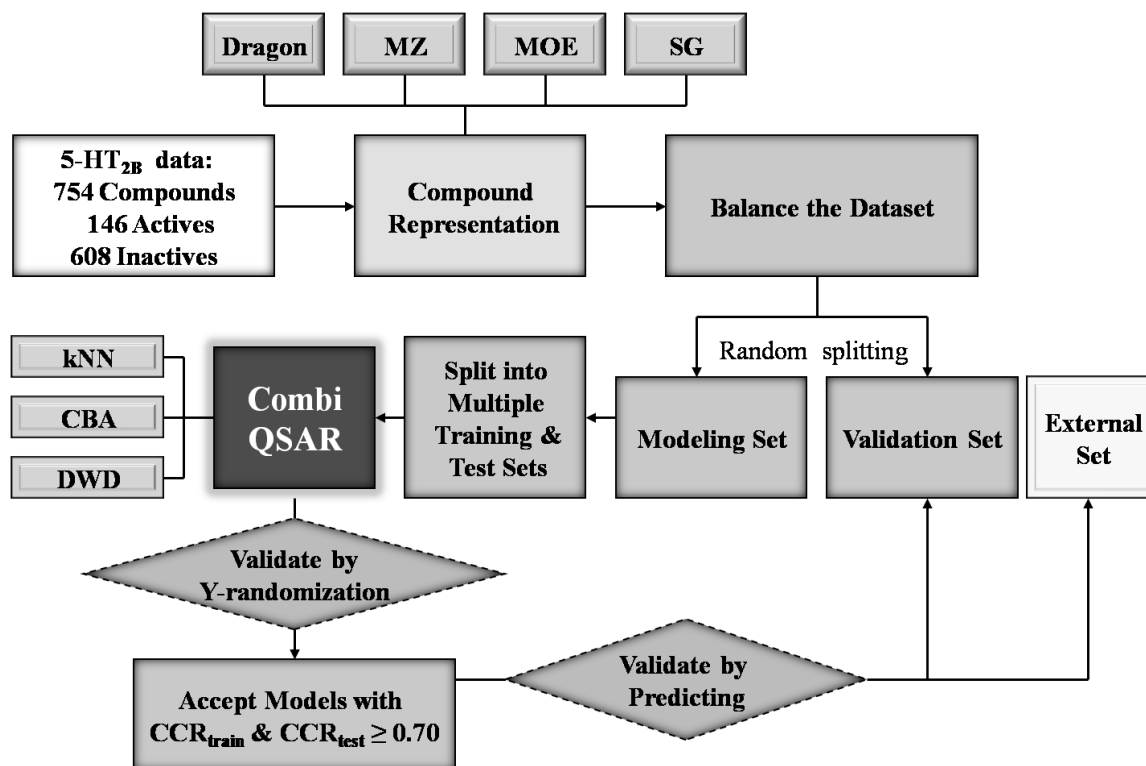
Another level of internal validation was achieved by comparing model performance for training and test sets. This approach is always employed as a part of our predictive QSAR modeling workflow (Tropsha 2010) to emphasize the fact that training-set-only modeling is not sufficient to obtain reliable models that are externally predictive (Golbraikh & Tropsha 2002a). Thus, for each collection of descriptors, the modeling sets were further partitioned into multiple chemically diverse training and test sets of different sizes using the Sphere

Exclusion method implemented in our laboratory (Golbraikh & Tropsha 2002b). Only models that were highly predictive on the test sets were retained for the consensus prediction of the external validation sets. Finally, only those models that were shown to be highly predictive on both external sets were used in consensus fashion for virtual screening of external compound libraries.

### **Computational Methods**

A combinatorial QSAR approach (Combi-QSAR) (de Cerqueira et al. 2006, Kovatcheva et al. 2004) was used to generate classification models for actives *vs.* inactives (Fig. 4.1). In this study, four types of descriptors were applied in combination with three types of statistical methods.

**Figure 4.1.** The workflow for QSAR model building and validation as applied to the 5-HT<sub>2B</sub> dataset (see text for abbreviations).



## Molecular Descriptors

Four sets of molecular descriptors were considered in our modeling studies: Dragon (Dragon 2007), MZ (MolconnZ 2006), MOE (MOE 2008), and SG descriptors (Khashan et al. 2005) developed in this laboratory. Each type of these descriptors was used separately with each of the classification methods in the context of our Combi-QSAR strategy. All details about descriptors are mentioned in Chapter 2.

## Balancing the Dataset Using Similarity Searching

The dataset used for model building was imbalanced, consisting of 146 actives vs. 608 inactives. Therefore, only a subset of the larger class of inactives of approximately the same size as the actives was used in model building unless otherwise indicated. This subset was selected to include inactives that were most similar to the actives. Given the vast array of available chemical descriptors and the large number of similarity measures, it is always difficult to decide *a priori* which combination of descriptors/similarity metrics to use. This problem has been highlighted in several recent publications (Holliday et al. 2002, Sheridan & Kearsley 2002). Therefore, similarity searching studies were performed using three types of molecular descriptors: fingerprints (FP), Dragon, and MZ, and applying two similarity metrics, i.e., Euclidean distance and Tanimoto coefficient (Tc). The similarity cutoff was chosen to obtain the most balanced (with roughly equal number of compounds from each class) subset of compounds.

**Fingerprints.** 166 MACCS (MDL Ltd 1992) structural keys implemented in MOE 2007.09 software were calculated for all compounds. The similarity searching was performed using an in-house written script applying Tanimoto coefficients for similarity measures.

**Dragon Descriptors.** Normalized Dragon descriptors of the original dataset were employed to calculate similarities between all compounds in the dataset using Euclidean distance as similarity metric; variable similarity thresholds were used to down-sample the larger class (inactives). Although many schemes could be considered for down-sampling the larger classes, we used the similarity threshold based approach since it restricts the larger class to compounds most similar to the smaller class molecules. This approach makes it more challenging to develop statistically significant models capable of discriminating smaller class compounds from most chemically similar molecules in the larger class. Therefore, it increases the robustness of the binary QSAR models.

**MolConnZ Descriptors.** Similar procedures to those described above for Dragon descriptors were used.

## **QSAR Methods**

We used the *k*NN classification method (Zheng & Tropsha 2000), CBA (Liu et al. 1998, Liu et al. 1999), and DWD (Marron et al. 2007) . All details about these methods are discussed in Chapter 2.

## **Robustness of QSAR Models**

Y-randomization test is a widely used validation technique to ensure the robustness of a QSAR model (Wold & Eriksson 1995). This test includes (i) randomly shuffling the dependent-variable vector, Y-vector of training sets (class labels in this study) and (ii) rebuilding models with the randomized activities (class labels) of the training set. All calculations are repeated several times using the original independent-variable matrix. It is expected that the resulting QSAR classification models, built with randomized activities for

the training set, should generally have low CCRs for training, test, and external validation sets. It is likely that sometimes, though infrequently, high CCR values may be obtained due to a chance correlation or structural redundancy of the training set. However, if some QSAR classification models obtained in the Y-randomization test have relatively high CCR it implies that an acceptable QSAR classification model cannot be obtained for the given dataset by the particular modeling method. Y-randomization test was applied to all datasets considered in this study, and the test was repeated five times in each case.

### **Applicability Domains of *k*NN QSAR Models**

Formally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, since the training set models are developed in *k*NN QSAR modeling by interpolating activities of the nearest neighbor compounds, a special applicability domain (i.e., similarity threshold) should be introduced to avoid making predictions for compounds that differ substantially from the training set molecules (Shen et al. 2004).

The similarity was estimated using Euclidean distances in high-dimensional descriptors space. Compounds with the smallest distance between them have the highest similarity. The distribution of distances (pairwise similarities) of compounds in our training set is computed to produce an applicability domain threshold,  $D_T$ , calculated as in Equation 2.9.

In this study two types of applicability domains were employed. First, a global applicability domain that ensures some level of global similarity (using all descriptors for similarity calculations) between the predicted compounds and the compounds in the

modeling set. The second is a local domain which is the applicability domain of each of the individual models using only descriptors used for the model building.

### **Consensus Prediction**

Our experience suggests that consensus prediction of the target property for external compounds, i.e., when the compound activity is calculated by averaging values predicted by all individual models that satisfy our acceptability criteria always provides the most stable and accurate solution (Zhu et al. 2008a). In general, consensus prediction implies averaging the predictions for each compound by majority voting for classification QSAR models, using all models passing the validation criteria (e.g.,  $CCR_{\text{train}} \geq 0.70$  and  $CCR_{\text{test}} \geq 0.70$ ). In order to determine the confidence in the obtained predictions we need to define a consensus score. The consensus scores employed in this study take into account the total number of models used to predict the compound's activity, and the number of models that predicted the compound to belong to a specific class. Since we define two classes of compounds, i.e., class 1 (actives) and class 0 (inactives), some models may predict a compound to belong to class 0 and others may predict it to belong to class 1. As a result, a consensus score between 0 and 1 will be obtained for each of the predicted compounds. As an additional measure of confidence (and an additional applicability domain criterion) we only accepted those predictions that had an average predicted value (consensus score) above 0.70 (for actives) or below 0.30 (for inactives).

### **Virtual Screening and Compound Selection for Experimental Validation**

To identify putative actives, validated consensus models generated for 5-HT<sub>2B</sub> ligands were used for virtual screening of ca. 59,000 molecules within the WDI (Daylight 2004)

chemical library; the selection of hits was limited by the applicability domains of each models. 122 compounds were identified as VS hits (by consensus agreement between all accepted models, see Table S1 of Supporting Information for details) and 10 structurally diverse and commercially available hits were purchased from different suppliers and tested at PDSP in 5-HT<sub>2B</sub> radioligand binding assays.

## **Results and Discussion**

### **Combinatorial QSAR Modeling of 5-HT<sub>2B</sub> Actives vs. Inactives**

**Balancing the Dataset.** The original dataset of 146 actives and 608 inactives was first balanced by downsizing the class of inactives. Similarity searching between active and inactive compounds using Tc cutoff of 0.7 resulted in 195 inactives (that were similar to at least one active compound with Tc above 0.7), which were combined with the 146 actives to form the modeling set of 342 compounds. Dragon and MZ descriptors were generated for this 342-compound modeling set to be used separately with *k*NN. However, similarity searching using Dragon and MZ descriptors and applying Euclidean distance-based threshold resulted in a 304- (146 actives and 158 inactives) and 325-compound (146 actives and 179 inactives) modeling sets respectively. Thus, slightly different modeling sets were used depending on the type of descriptors.

***k*NN Classification.** *k*NN method was used with each of the following descriptor types: DRAGON, MZ, MOE, and SG descriptors. Models were built for the three datasets resulting from the down-sampling of the original dataset. First, a validation set (15-20% of the dataset) was excluded from each of the resulting datasets randomly. The compounds in the remaining modeling set (85-80% of the original dataset) were divided into multiple



training and test sets (28-40 divisions). Multiple QSAR models were generated independently for all training sets and applied to the test sets. Generally, we accepted models with CCR values for both the training and test set greater than 0.70. *k*NN combined with subgraphs and Dragon descriptors were the two best performing methods based on validation set statistics (Table 4.2). *k*NN-subgraphs (*k*NN-SG) had a  $CCR_{\text{evs}} = 0.80$ , while *k*NN-Dragon gave a  $CCR_{\text{evs}} = 0.72$ .

Results of the Y-randomization test confirmed that *k*NN classification models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  values  $\geq 0.70$  were robust. None of the models with randomized class labels of the training set compounds had  $CCR_{\text{rand}} > 0.54$  for any dataset.

**Table 4.2.** Performance of  $k$ NN classification methods to classify actives vs. inactives based on external validation set statistics.

Model	Num.	Confusion Matrix						Statistics for the Models				
	Mod <sup>a</sup>	N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>
A <sup>f</sup>	908	26	34	20	23	11	6	0.77	0.68	1.41	1.49	0.72
B <sup>g</sup>	235	38	36	22	20	16	16	0.58	0.56	1.13	1.14	0.57
C <sup>h</sup>	619	32	38	17	29	9	15	0.53	0.76	1.38	1.24	0.65
D <sup>i</sup>	387	30	40	16	29	11	14	0.04	0.73	0.26	1.90	0.63
E <sup>j</sup>	123	30	40	20	26	14	10	0.67	0.65	1.31	1.32	0.66
F <sup>k</sup>	93	30	40	23	33	7	7	0.77	0.83	1.63	1.56	0.80

<sup>a</sup>Num. Mod, number of models with  $CCR_{train}$  and  $CCR_{test} \geq 0.70$ ; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set; <sup>e</sup>CCR<sub>rand</sub>, correct classification rate of the random models using the external validation set; <sup>f</sup>A,  $k$ NN-Dragon; <sup>g</sup>B,  $k$ NN-MZ; <sup>h</sup>C,  $k$ NN-Dragon-FP; <sup>i</sup>D,  $k$ NN-MZ-FP; <sup>j</sup>E,  $k$ NN-MOE-FP; <sup>k</sup>F,  $k$ NN-SG.

**Classification Based on Association.** The CBA method was applied to classify the dataset using SG descriptors. A dataset of 342 compounds (146 actives and 196 inactives), resulting from the downsizing process with FP and Tanimoto distances, was used. The dataset was split randomly into training (267 compounds) and validation sets (75 compounds). A total of 1371 closed frequent subgraphs were generated with FFSM (see Methods in Chapter 2) from the training set using a support value of 12% and a maximum size limit of the fragments of 7. The training set consisting of 267 compounds (111 actives and 156 inactives) was then used to build the classifier in CBA. The classifier gave a  $CCR_{\text{train}}$  of 0.79. Then the validation set consisting of 75 compounds (35 actives and 40 inactives) was used to assess the robustness of the classifier. The  $CCR_{\text{evs}}$  was 0.65 which is not as high as the CCR value for the training set.

**DWD Modeling.** The DWD method was applied to classify the dataset using Dragon descriptors. A dataset of 304 compounds (146 actives and 158 inactives), resulting from the downsizing process with Dragon descriptors and Euclidean distances, was used. The dataset was split randomly into training (244 compounds) and validation sets (60 compounds). A total of 387 Dragon descriptors were generated for the training set. The training set consisting of 244 compounds (120 actives and 124 inactives) was then used to build the DWD model. The DWD model was able to group compounds in this dataset based on their biological classes with a  $CCR_{\text{evs}} = 0.70$  (TP=18, TN=24, FP=10, FN=8), setting the cutoff at “0.15”. DWD was further used to rank Dragon descriptors according to their importance for discriminating the two classes of compounds (actives vs. inactives). DWD uses class label information where positive (for actives) and negative (for inactives) signs are assigned to each descriptor value to indicate its importance to the corresponding class. The top 20 highly

weighted descriptors (based only on weights' values and ignoring the signs) are presented in Table 4.3.

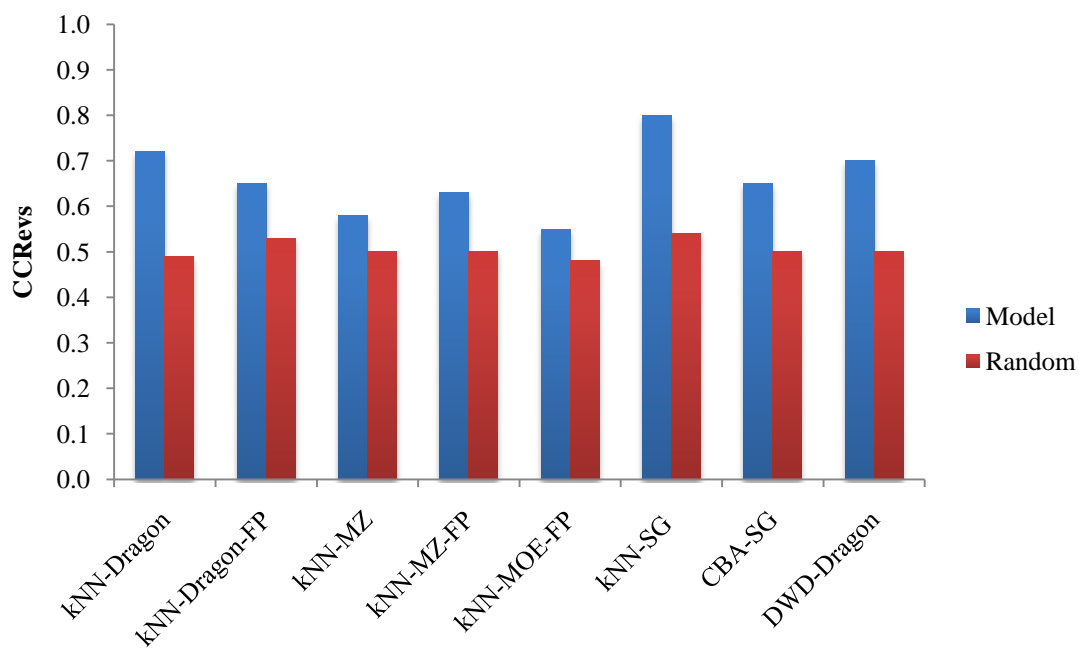
**Table 4.3.** Top twenty highly weighted Dragon descriptors by DWD for 5-HT<sub>2B</sub> actives vs. inactives.

Descriptor	DWD Weight	Interpretation
JGI9	2.07E-01	Mean topological charge index of order 9
nRNR2	1.99E-01	Number of tertiary amines (aliphatic)
C-006	1.79E-01	CH2RX
C-024	1.66E-01	R--CH--R
T(N..F)	1.52E-01	Sum of topological distances between N..F
H-047	1.51E-01	H attached to C1 (sp <sup>3</sup> )/CO (sp <sup>2</sup> )
H-049	-1.42E-01	H attached to C3(sp <sup>3</sup> )/C2(sp <sup>2</sup> )/C3(sp <sup>2</sup> )/C3(sp)
C-027	-1.38E-01	R--CH--X
JGI10	-1.38E-01	Mean topological charge index of order 10
GATS7e	1.38E-01	Geary autocorrelation - lag 7 / weighted by atomic Sanderson electronegativities
PCR	1.35E-01	Ratio of multiple path count over path count
nBnz	1.32E-01	Number of benzene-like rings
H-051	-1.30E-01	H attached to alpha-C
GATS8m	1.21E-01	Geary autocorrelation - lag 8/weighted by atomic masses
C-013	1.16E-01	CRX3
C-034	-1.14E-01	R--CR..X
BELe8	1.14E-01	Lowest eigenvalue n. 8 of Burden matrix / weighted by atomic Sanderson electronegativities
JGI5	-1.11E-01	Mean topological charge index of order 5
nN+	-1.10E-01	Number of positively charged nitrogen
GATS3p	1.08E-01	Geary autocorrelation - lag 3 / weighted by atomic polarizabilities

### Comparison of Binary QSAR Approaches for Classifying 5-HT<sub>2B</sub> Actives vs. Inactives

The performance of different binary QSAR approaches employed as part of combinatorial QSAR strategy for 5-HT<sub>2B</sub>, and based on validation set statistics, is summarized in Figure 4.2. *k*NN-SG, and *k*NN-Dragon were the best performing methods for classifying 5-HT<sub>2B</sub> actives vs. inactives based on validation set statistics (Table 4.2), yielding the highest CCR<sub>evs</sub> of 0.80 in case of *k*NN-SG. On the contrary, *k*NN-MZ was the worst performing method with a CCR<sub>evs</sub> of 0.57 which was very close to random. It was also interesting to see that *k*NN-SG performed much better than CBA-SG with CCR<sub>evs</sub> = 0.80 in the former case and 0.65 in the latter. These results confirm the importance of employing the combinatorial QSAR approach to find the most predictive QSAR method/descriptor combination for each specific dataset.

**Figure 4.2.** Comparison of CCR values for the external validation set ( $CCR_{evs}$ ) for different QSAR models developed to classify actives *vs.* inactives.  $CCR_{evs}$  values for models built with both real (blue) and randomized (red) activities of the training sets are shown (see text for abbreviations).



Our models also indicated that the nature of the descriptors used has a dramatic effect on the performance of the modeling methods. It was clear that MOE and MZ descriptors did not perform very well in all tested cases irrespective of the applied modeling techniques. On the contrary, Dragon descriptors afforded most significant models with all methods and in all tests, for both validation and external sets.

### **Model Validation by Predicting Drugs Known to be 5-HT<sub>2B</sub> Actives and Valulopathogens**

Both fenfluramine and dexfenfluramine (known to be 5-HT<sub>2B</sub> actives and agonists, which were not included in our modeling sets) were predicted as 5-HT<sub>2B</sub> actives using consensus models to classify actives vs. inactives. The consensus scores using *k*NN-Dragon were 0.79 for both compounds. Our previous studies suggest that consensus prediction that is based on the results obtained by all validated predictive models always provides the most stable solution (Zhu et al. 2008a). A 5-HT<sub>2B</sub> active compound can have consensus scores in the interval [0.5-1.0]. The closer value to 1.0 the greater is the confidence in the prediction. Therefore, we can claim that both compounds were predicted as actives with statistically significant consensus scores.

These results highlight the predictive power of our validated models that could have predicted the possible dangerous side effects of these two drugs by suggesting that they may be 5-HT<sub>2B</sub> actives. This prediction would have suggested that these compounds should be tested experimentally in 5-HT<sub>2B</sub> functional assays and prevented from further development as potentially unsafe medicines. This example illustrates the potential use of models developed in this study as computational drug safety alerts.



## Model Validation by Predicting an External Set

An additional 16-compound set was obtained from PDSP after we finished our modeling studies. This external set was used to further assess the robustness and the predictive power of our models. All 16 compounds were 5-HT<sub>2B</sub> actives including 4 agonists and 12 antagonists.

The 16 external compounds were predicted using all consensus models built to classify actives vs. inactives. *k*NN-Dragon was the best performing method on this external set with a CCR<sub>ex</sub> of 0.81. Predictions were made by applying local model applicability domains with  $Z = 0.5$  (see Applicability Domain of *k*NN QSAR Models). It was interesting to find that *k*NN-Dragon had  $CCR \geq 0.72$  with both the validation (CCR<sub>evs</sub> = 0.72) and the external (CCR<sub>ex</sub> = 0.81) sets. However, *k*NN-SG (the best performing method on validation sets) was not as good with the external set (CCR<sub>ex</sub> = 0.65) as it was with the validation set (CCR<sub>evs</sub> = 0.80). CBA-SG gave a CCR<sub>ex</sub> = 0.65, which was consistent with its performance with the validation set (CCR<sub>evs</sub> = 0.65) but less than CCR<sub>train</sub> (0.79). The latter results using SG descriptors with *k*NN and CBA might be due to the limitation that frequent subgraphs are derived from the training set compounds; therefore, it is possible that fragments that are frequent in the external set are not represented in the frequent subgraphs used for prediction. Our current applicability domain filter, which is calculated using the fragments in the training set, does not account for this possibility. It is clear that a more stringent applicability domain filter could be applied in this case, which uses the distribution of subgraphs counts between the training and test set, but this has not been implemented yet within our current method.

## The Importance of Variable Selection

Since *k*NN-Dragon was the best performing method to classify actives *vs.* inactives based on the results for all validation sets, we thought it would be interesting to check the performance of *k*NN using all 387 Dragon descriptors, generated for the actives *vs.* inactives modeling set, without variable selection. The results of this test are shown in Table 4.4. Comparison of modeling results for *k*NN-variable-selection ( $CCR_{\text{evs}} = 0.72$ ) *vs.* *k*NN-without-variable-selection ( $CCR_{\text{evs}} = 0.52$ ) clearly indicates that variable selection is a vital part of modeling. Furthermore, the top 20 most frequent descriptors (MFD) selected by *k*NN models (Table 4.5) and top 20 highly weighted descriptors by DWD based only on weights and ignoring the sign (Table 4.3) were used independently with the *k*NN method (with no variable selection) to predict actives *vs.* inactives (Table 4.3). Models built with either the top 20 DWD-selected Dragon descriptors or MFD from Dragon-*k*NN and using 1-5 nearest neighbors gave  $CCR_{\text{evs}} \sim 0.5$  (Table 4.3). These results illustrated again that SA-based variable selection procedures implemented in our *k*NN QSAR method (Zheng & Tropsha 2000) lead to models with the highest external predictive power as compared to any other approach not relying on variable selection for model optimization.

**Table 4.4.** Comparison between different *k*NN-Dragon QSAR models generated with or without variable selection.

Mod.	Num. Mod <sup>a</sup>	Confusion Matrix						Statistics for the Models					
		N1 <sup>b</sup>	N0 <sup>c</sup>	TP	TN	FP	FN	SE	SP	En1	En0	CCR <sub>evs</sub> <sup>d</sup>	Cover-age <sup>e</sup>
A <sup>f</sup>	908	26	34	20	23	11	6	0.77	0.68	1.41	1.49	0.72	100%
B <sup>g</sup>	1	26	34	10	22	10	8	0.38	0.65	1.13	1.36	0.52	83%
C <sup>h</sup>	1	26	34	14	15	19	9	0.54	0.44	0.98	1.12	0.49	95%
D <sup>i</sup>	1	26	34	14	15	19	9	0.54	0.44	0.98	1.12	0.49	95%

<sup>a</sup>Num. mod, number of models with CCR<sub>train</sub> and CCR<sub>test</sub> ≥ 0.70; <sup>b</sup>N1, number of actives; <sup>c</sup>N0, number of inactives; <sup>d</sup>CCR<sub>evs</sub>, correct classification rate of the consensus models using the external validation set; <sup>e</sup>Coverage, percentage of predicted external compounds; <sup>f</sup>A, *k*NN-Dragon; <sup>g</sup>B, *k*NN-Dragon-NVS where *k*NN model was generated using all 387 Dragon descriptors with no variable selection and 1 nearest neighbor (NN); <sup>h</sup>C, *k*NN-Dragon-MFD where the *k*NN model was generated with top twenty most frequent Dragon descriptors and 1NN; <sup>i</sup>D, *k*NN-Dragon-DWD where the *k*NN model was generated with top twenty highly weighted Dragon descriptors by DWD and one NN.

**Table 4.5.** Top twenty most frequently used Dragon descriptors in validated *k*NN-Dragon models to classify 5-HT<sub>2B</sub> actives vs. inactives.

Descriptor	Frequency	Interpretation
nN=C-N<	157	Number of amidine derivatives
C-027	152	R--CH--X
MATS5v	151	Moran autocorrelation - lag 5 / weighted by atomic van der Waals volumes
GATS5v	148	Geary autocorrelation - lag 5 / weighted by atomic van der Waals volumes
C-033	146	R--CH..X
MATS4v	135	Moran autocorrelation - lag 4 / weighted by atomic van der Waals volumes
MLOGP2	135	Squared Moriguchi octanol-water partition coeff. (logP <sup>2</sup> )
N-074	131	R#N / R=N-
nR=Cp	130	Number of terminal primary C(sp <sup>2</sup> )
MATS4m	128	Moran autocorrelation - lag 4 / weighted by atomic masses
nFuranes	128	Number of furanes
C-035	125	R--CX..X
nArNR2	119	Number of tertiary amines (aromatic)
nPyrroles	119	Number of pyrroles
nPyridines	118	Number of Pyridines
nArCO	118	Number of ketones
nROCON	117	Number of (thio-) carbamates (aliphatic)
nBeta-Lactams	117	Number of Beta-Lactams
H-053	114	H attached to CO(sp <sup>3</sup> ) with 2X attached to next C
C-008	113	CHR2X

Mechanistic interpretability is frequently regarded as very important feature of QSAR models. We generally argue that only models that have been extensively validated on external datasets and identified experimentally-confirmed hits should be subjected to interpretation. Furthermore, very few classes of models, specifically, those based on (multiple) linear regression and small number of descriptors can afford a relatively straightforward interpretation. The interpretation of multi-parametric statistical models developed with non-linear optimization algorithms (as in this study) should be attempted with great care because of strong and often poorly understood interplay between descriptors. Furthermore, although we could foresee that in some cases medicinal chemists may want to modify their candidate compounds to prevent 5HT<sub>2B</sub> binding, the tools developed in this study are predominantly intended for virtual screening of libraries of drug candidates to flag and possibly eliminate compounds that are likely to bind 5HT<sub>2B</sub> receptor, not to design new compounds; and any compound designed by chemists could be passed through our models. Therefore, we only restricted the discussion in this paper to the most frequent descriptors found by all acceptable *k*NN models and the most highly weighted descriptors selected by DWD to stress that the process of variable selection employed as part of model optimization has indeed converged on a small number of descriptors.

### **Virtual Screening of the World Drug Index Database to Identify Putative 5-HT<sub>2B</sub> Ligands**

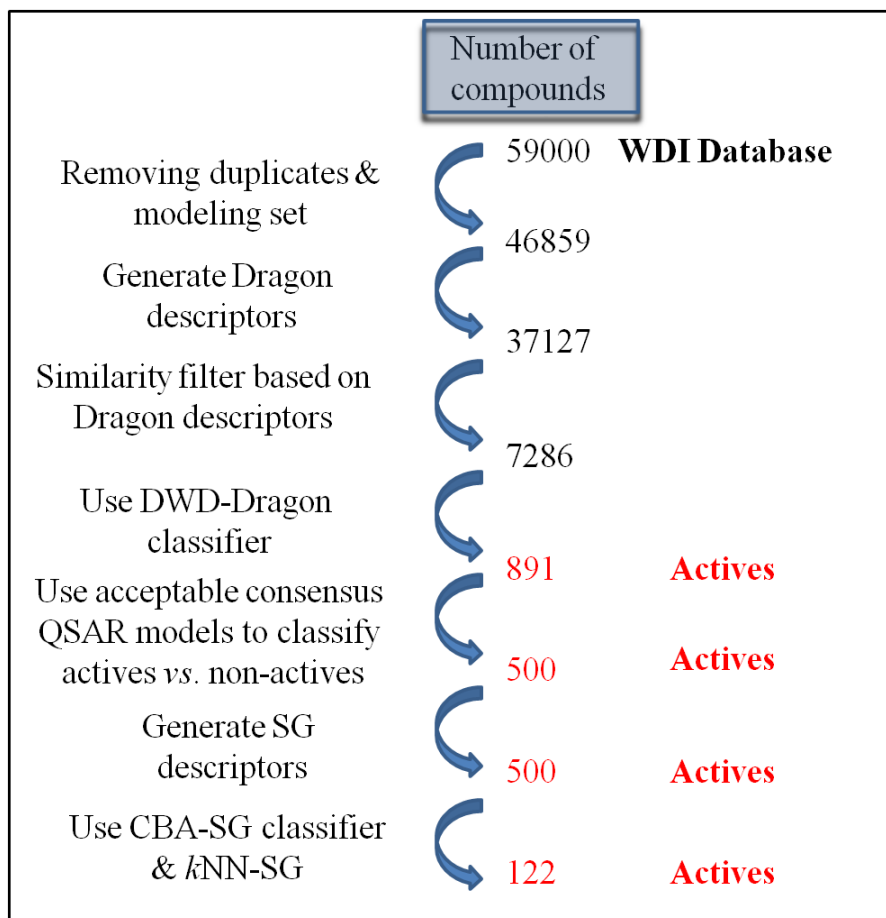
Since our models proved to be reasonably accurate based on two external validation sets, we used the best models to mine a large external database of approved and potential drugs for putative 5-HT<sub>2B</sub> actives. An important condition that assures reliable predictions by the model is the use of AD. Therefore, two types of AD were employed in the virtual

screening of compound databases. The first is a global AD that acts as a filter and ensures some level of global similarity between the predicted compounds and the compounds in the modeling set. The second is a local AD which is defined for each of the individual classification models.

The WDI database of ca. 59000 compounds (approved or investigational drugs) was used for virtual screening (Fig. 4.3). This original collection had many duplicates (i.e., many salt forms for the same chemical entity). The duplicates were removed using MOE: keeping unique structures and deleting duplicates. We also removed all compounds included in our modeling and external validation sets. Dragon descriptors were generated for the remaining 46859 unique compounds in the database; 926 compounds were excluded because Dragon was unable to calculate at least one of the descriptors generated for the modeling set. The remaining 45933 compounds were then subjected to a global AD filter for the actives *vs.* inactives modeling set using a strict Z cutoff of 0.5 (which formally places the allowed pairwise distance threshold at the mean of all pairwise distance distribution for the training set plus one-half of the standard deviation). Obviously, increasing the AD would increase the number of computational hits identified by virtual screening. However, our experience suggests that such increase is typically accompanied by the decrease in prediction accuracy. Additionally, we required that the nearest neighbor in the modeling set of a compound from the virtual library be an active. The resulting 7286 compounds were then classified into actives *vs.* inactives using DWD-Dragon classifier resulting in 891 actives. Next, all *k*NN-Dragon models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}} \geq 0.70$  were employed in consensus fashion to predict these 891 compounds resulting in a selection of the 500 active hits. At this point, SG descriptors were generated for these 500 molecules. CBA-SG classifier followed by *k*NN-SG

consensus models were used as final filters for the determination of 122 compounds regarded as putative 5-HT<sub>2B</sub> actives.

**Figure 4.3.** Steps of the virtual screening of the WDI database to identify putative 5-HT<sub>2B</sub> ligands (see text for the abbreviations).

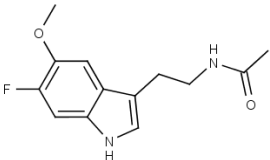
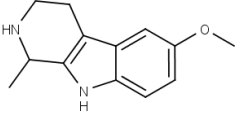
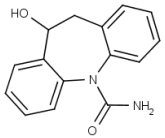
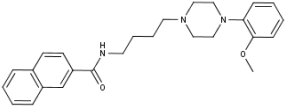
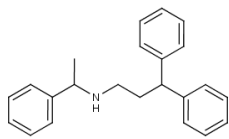


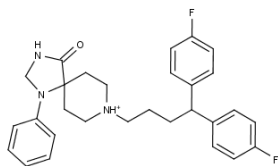


## Experimental Validation

Ten structurally diverse hits (**1-10**, see Table 4.6) were selected from the final consensus virtual screening hits for further experimental validation taking into account both their commercial availability and cost (see Table 4.6). To our satisfaction, nine compounds were confirmed to inhibit 5-HT<sub>2B</sub> radioligand binding, which implies a hit rate of 90 %.  $K_i$  values were in the range 0.8 – 3,127 nM, with 4 compounds having  $K_i$  values < 100 nM. The four highest affinity compounds were: **4** ( $K_i=33$  nM, see Fig. 4 (A)), **7** ( $K_i=0.8$  nM, see Setola *et al*, 2003 (Setola *et al.* 2003)), **9** ( $K_i=70$  nM, see Fig. 4 (B)), and **10** ( $K_i=69$  nM, see Fig. 4 (C)). It should be noted that **7**, though not included initially in our dataset, was known to be a valvulopathic compound and had been tested against 5-HT<sub>2B</sub> receptors in both binding ( $K_i=0.8$  nM) (Setola *et al.* 2003) and functional assays (pEC<sub>50</sub> for 5-HT<sub>2B</sub>-Mediated calcium flux = 7.67) (Huang *et al.* 2009). In order to determine the activity of the remaining eight 5-HT<sub>2B</sub> ligands, all compounds were tested at the PDSP in 5-HT<sub>2B</sub> functional assays. Results indicated that methylergometrine was the only compound among the nine 5-HT<sub>2B</sub> ligands that possessed strong agonist activity.

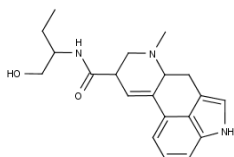
**Table 4.6.** Experimental validation results for the 10 computational hits predicted as 5-HT<sub>2B</sub> ligands as a result of QSAR-based mining of the WDI chemical screening library.

Chemical Structure/ Name	PubChem CID	PDSP ID	Predicted 5-HT <sub>2B</sub> Activity	Experimental K <sub>i</sub> (nM)
 <p>(1) 6-Fluoromelatonin</p>	43922	14809	Active	2,495.0
 <p>(2) Adrenoglomerulotropin</p>	71028	14807	Active	491.0
 <p>(3) CGP-13698</p>	114709	14806	Active	>10,000
 <p>(4) DO-897</p>	3038495	14814	Active	33.1
 <p>(5) Fendiline</p>	3336	14821	Active	3,217.0



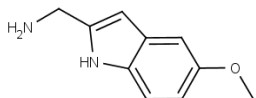
1715104 14815 Active 151.4

**(6) Fluspirilene**



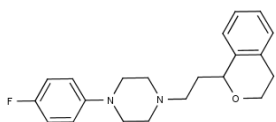
4140 27769 Active 0.8

**(7) Methylergometrine**



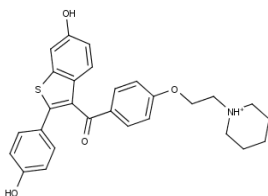
195658 14805 Active 1,617.0

**(8) PIM-35**



9909648 13513 Active 69.6

**(9) PNU-96415E**

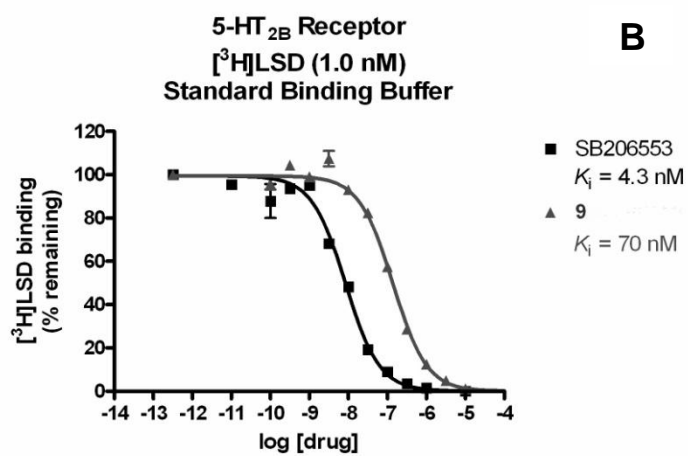
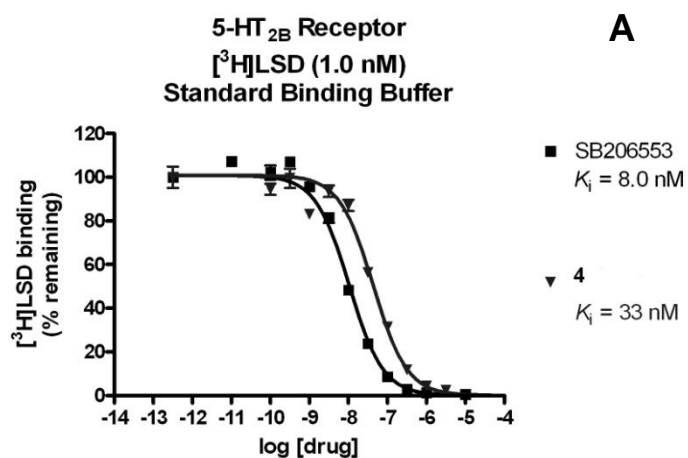


15940170 13505 Active 69.0

**(10) Raloxifene**

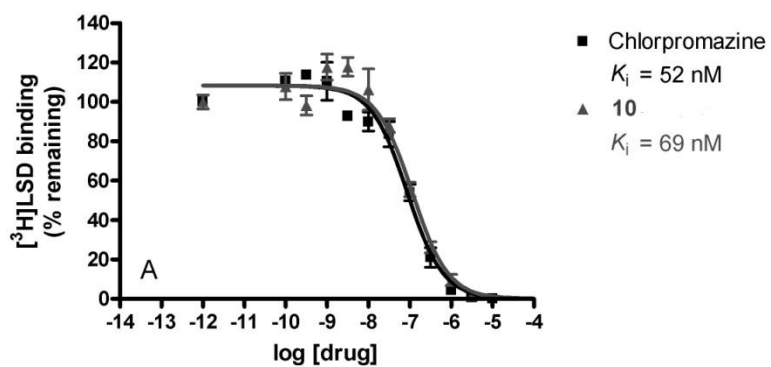
---

**Figure 4.4.** Competition binding at 5-HT<sub>2B</sub> receptors for (A) **4** (triangle) and SB206553 (square), (B) **9** (triangle) and SB206553 (square), and (C) **10** (triangle) and chlorpromazine (square), versus [<sup>3</sup>H]LSD.



5-HT<sub>2B</sub> Receptor  
[<sup>3</sup>H]LSD (1.5 nM)  
Standard Binding Buffer

C



This low hit rate of 11.1% for identifying validated agonists is in fact not surprising in light of Huang *et al* (Huang et al. 2009) major finding that potent 5-HT<sub>2B</sub> receptor agonism is a relatively rare occurrence among drugs and drug-like compounds. However, to arrive at such conclusions, Huang *et al* screened a composite library containing three publicly available collections of FDA-approved and investigational medications and one internally compiled library. Of the approximately 2200 compounds screened, 27 5-HT<sub>2B</sub> receptor agonists were identified; thus, the validated hit rate was 1.2%.

These results illustrate that the validated QSAR workflow, as employed in this paper, could be used as a general tool for identifying 5-HT<sub>2B</sub> ligands by the means of virtual screening of chemical libraries using rigorously built QSAR models. As we demonstrated in this study, our models identify a relatively small number of VS hits making it feasible to employ experimental tools to validate predictions in 5-HT<sub>2B</sub> binding and functional assays. Ten compounds selected from a large external library have been tested experimentally in this proof-of-concept study resulting in very high experimentally confirmed hit rate. The list of all compounds predicted to be 5-HT<sub>2B</sub> actives is available in Appendix I.

To verify the diversity of the experimentally validated hits, we have compared the results of QSAR-based virtual screening with simple similarity searches. Similarity calculations were done using two different descriptor-metric combinations: (1) MACCS structural keys and Tanimoto coefficients (as a standard similarity searching approach, see Table 4.7 and Figure 4.5) and (2) Dragon descriptors and Euclidean distances (to compare directly with our best performing QSAR models of *k*NN-Dragon, see Table 4.8 and Figure 4.6. The nearest neighbor compounds (based on Tanimoto similarities and MACCS keys) from the active compounds in the dataset and the 10 experimentally validated VS hits are

reported in Table 4.9. Results of similarity analyses indicated that neither technique would be able to efficiently identify the diverse hits obtained with our methods. Hence, our studies illustrated the power of combi-QSAR-based VS in prioritizing compounds (which are not just close analogs of the modeling set compounds) from screening libraries to achieve high success rates when experimentally validated.

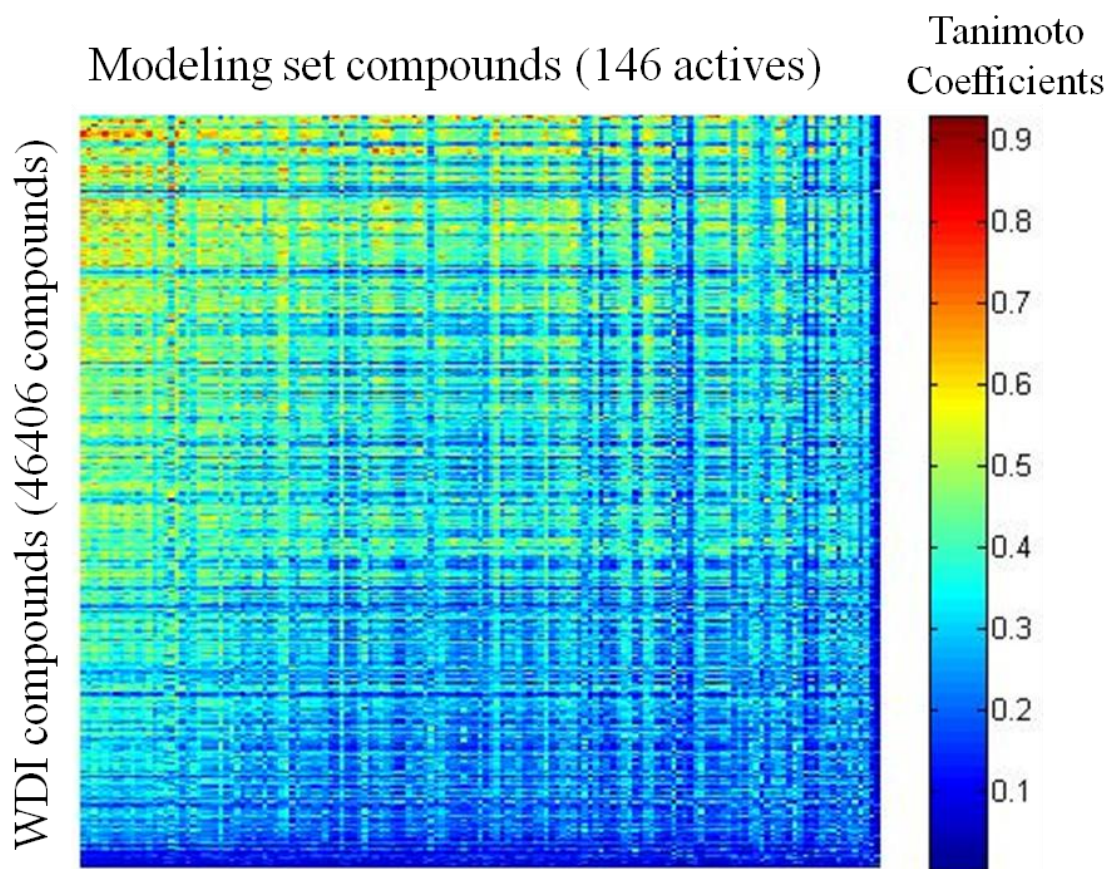
**Table 4.7.** Virtual screening recovery results using Tanimoto similarities and 166 MACCS

keys.

<b>Tanimoto Coefficient</b>	<b>46406 WDI Compounds</b>	<b>122 VS Hits</b>	<b>10 Tested Hits</b>
$\geq 0.9$	286	2	2
$\geq 0.8$	1341	4	3
$\geq 0.7$	7048	13	8
$\geq 0.6$	21431	38	9
$\geq 0.5$	36719	81	9
$\geq 0.4$	44208	115	10
$\geq 0.3$	45860	122	10
$\geq 0.2$	46220	122	10
$\geq 0.1$	46301	122	10
$\geq 0.0$	46406	122	10



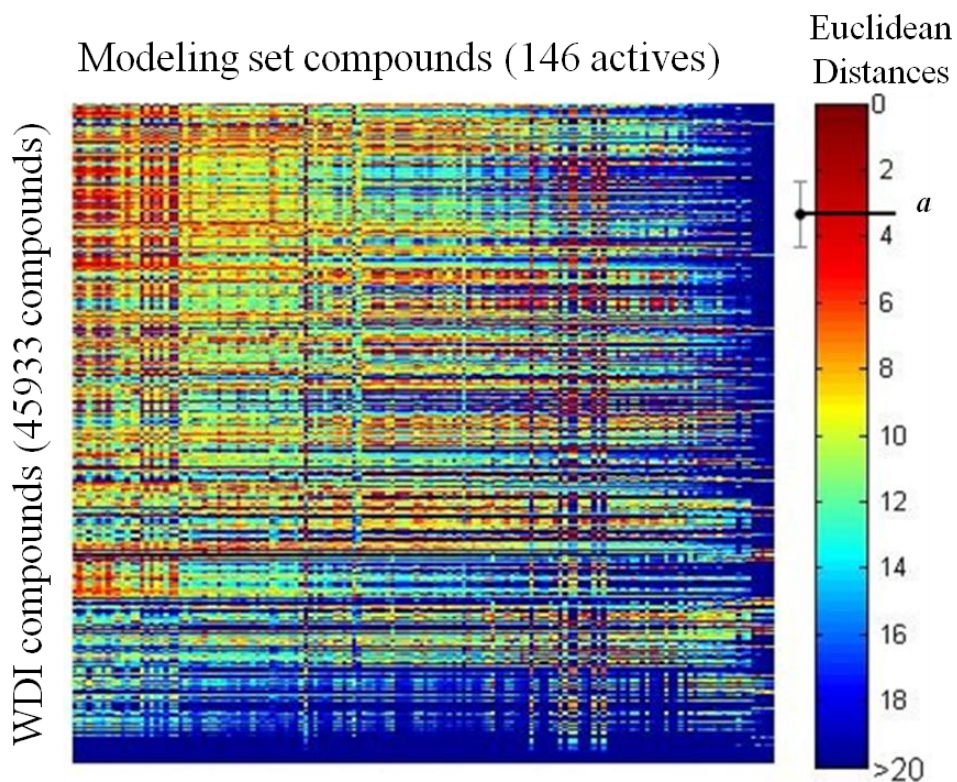
**Figure 4.5.** The heatmap of the similarity between the 146 5-HT<sub>2B</sub> actives from the modeling set and 46406 WDI compounds. Virtual screening compounds included in WDI were analyzed by estimating pairwise structural similarities with modeling set actives using Tanimoto coefficients and 166 MACCS structural keys. The bar-view is a key for coloring according to similarity/dissimilarity based on Tanimoto coefficients where red color indicates most similar compounds while blue color denotes least similar compounds.



**Table 4.8.** Virtual screening recovery results using Euclidean distances and 903 Dragon descriptors.

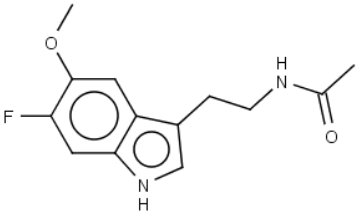
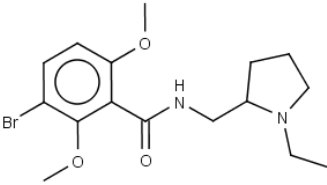
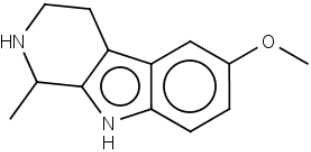
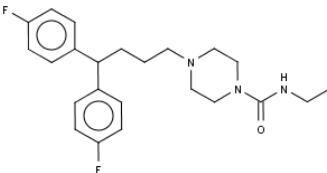
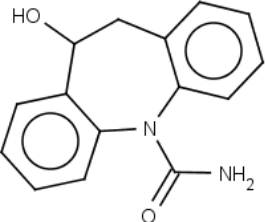
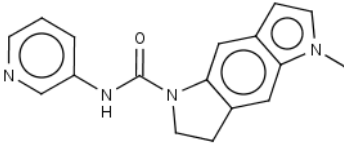
<b>Euclidean Distance</b>	<b>54933WDI Compounds</b>	<b>122 VS Hits</b>	<b>10 Tested Hits</b>
≤ 0.5	157	0	0
≤ 1.0	948	3	3
≤ 1.5	4191	10	6
≤ 2.0	9835	30	7
≤ 2.5	16411	54	9
≤ 3.0	22419	73	10
≤ 3.5	27200	86	10
≤ 4.0	31035	96	10
≤ 4.5	34008	102	10
≤ 5.0	36340	116	10
≤ 5.5	38106	119	10
≤ 6.0	39528	120	10
≤ 6.5	40714	121	10
≤ 7.0	41550	122	10
≤ 10.0	43738	122	10
≤ 50.0	45692	122	10
≤ 100	45880	122	10
≤ 120	45932	122	10
≤ 130	45933	122	10

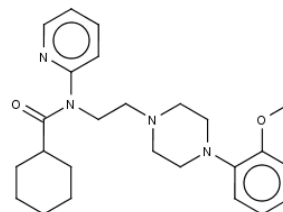
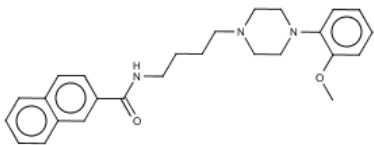
**Figure 4.6.** The heatmap of distances between the 146 5-HT<sub>2B</sub> actives from the modeling set and 45933 WDI compounds. Virtual screening compounds included in WDI were analyzed by estimating pairwise structural similarities with modeling set actives using Euclidean Distances and 903 Dragon descriptors. The bar-view is a key for coloring according to similarity/dissimilarity based on Euclidean distances where red color indicates most similar compounds while blue color denotes least similar compounds. Additionally, in this figure we colored all instances with distances above 20 with blue.



<sup>a</sup>Mean and SD of nearest neighbor distances for modeling set actives (146 compounds).

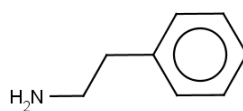
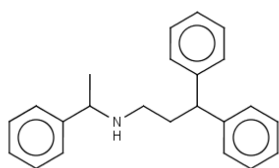
**Table 4.9.** Nearest neighbor compounds from the active compounds in the dataset and the ten experimentally validated VS hits.

Virtual screening hits	<sup>a</sup> Nearest neighbor from the modeling set
 <p><b>1</b></p>	 <p>PubChem CID 54940 (66 % similarity)</p>
 <p><b>2</b></p>	 <p>PubChem CID 108029 (70 % similarity)</p>
 <p><b>3</b></p>	 <p>PubChem CID 5163 (49 % similarity)</p>



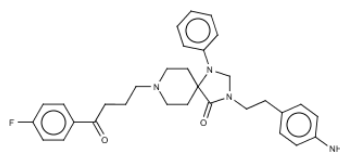
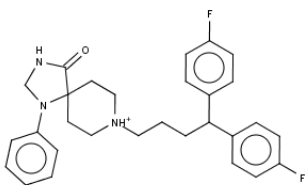
4

PubChem CID 5684 (88 % similarity)



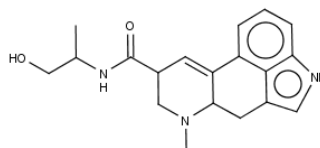
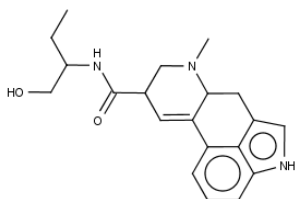
5

PubChem CID 1001 (71 % similarity)



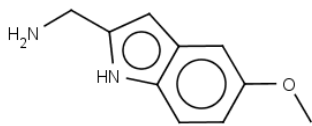
6

PubChem CID 125085 (90 % similarity)

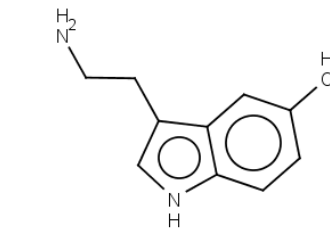


7

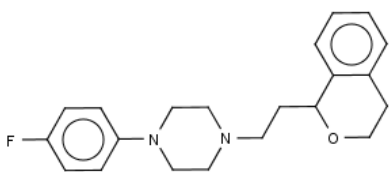
PubChem CID 3250 (98 % similarity)



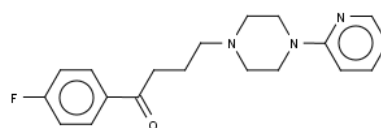
**8**



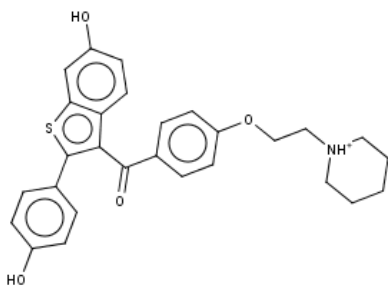
PubChem CID 5202 (76 % similarity)



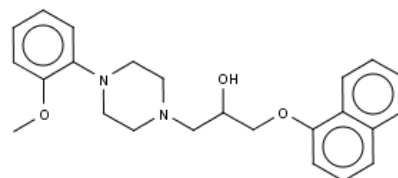
**9**



PubChem CID 15443 (79 % similarity)



**10**



PubChem CID 4418 (74 % similarity)

---

<sup>a</sup>Nearest neighbor from the modeling set compounds based on MACCS structural keys and Tanimoto distances.

We also think that agonist *vs.* antagonist models will be highly useful as more data about agonist compounds become available. The small number of known 5-HT<sub>2B</sub> agonists made it impossible at this stage to develop statistically significant models that could distinguish agonists from antagonists. Thus, the current study was limited to building binder *vs.* non-binder models. We will continue with our efforts to develop quantitative 5-HT<sub>2B</sub> agonist predictors as we accumulate more experimental data.

## **Conclusions**

QSAR models are becoming increasingly attractive as robust computational tools for virtual screening due to both their computational efficiency and success rates [reviewed in (Tropsha & Golbraikh 2007) as well as in a recent monograph (Varnek & Tropsha 2008)]. In this study, we have applied a combinatorial QSAR approach to a dataset of 800 compounds experimentally annotated as 5-HT<sub>2B</sub> receptor agonists, antagonists and inactives resulting in statistically validated and externally predictive models. Specifically, we have applied a combi-QSAR approach utilizing three different classification methods (*k*NN, CBA and DWD) and four different descriptor types (Dragon, MZ, MOE and SGs) to generate classification QSAR models to discriminate between 5-HT<sub>2B</sub> actives (agonists and antagonists) from inactives. Predictive models with classification accuracies as high as 0.80 for actives *vs.* inactives, as estimated on external validation sets, were obtained.

Classification models for actives *vs.* inactives were further validated by predicting an external validation set obtained after we completed the modeling studies. The high accuracy of prediction for the second external validation set proved that our models were indeed rigorous. Therefore, we posited that our studies afforded a robust computational tool to

predict potential 5-HT<sub>2B</sub> activity and consequently prioritize hits for testing in functional 5-HT<sub>2B</sub> assays to predict valvulopathic side effects of drugs and drug candidates that act as 5-HT<sub>2B</sub> agonists. We suggested that this computational predictor could be used to eliminate high risk compounds at the early stages of the drug development process. To illustrate this point, we have used this predictor retrospectively to evaluate the valvulopathic potential of two drugs withdrawn from the U.S. market for this reason, i.e., fenfluramine and dexfenfluramine. Both drugs were not included in our modeling set and both were indeed predicted with high confidence as actives for binding to 5-HT<sub>2B</sub> receptors.

Encouraged by our model validation results, we have applied these models for virtual screening of the 59000 compounds in WDI database. Our classification strategies identified 122 potential 5-HT<sub>2B</sub> ligands. Ten structurally diverse VS hits were experimentally tested at PDSP. Nine compounds were experimentally confirmed as 5-HT<sub>2B</sub> ligands thereby demonstrating a very high success rate of 90%.

The predictor developed in this report is similar in its potential use to other predictors of drug liability such as carcinogenicity and mutagenicity that are widely used in pharmaceutical industry. For instance, the TOPKAT program available in the Discovery Studio (Discovery Studio 2008), is a QSAR-based system that generates and validates accurate, rapid assessments of various types of chemical toxicity solely from a chemical's molecular structure. In contrast, our predictor is a unique specialized tool for the prediction of 5-HT<sub>2B</sub> activity and therefore prioritizing compounds for functional testing against 5-HT<sub>2B</sub> receptors to assess their valvulopathic potential. Therefore, this predictor can be used, along with other computational chemical health risk assessment tools, to evaluate compounds' safety at early stages of the drug development. It can be used as well to verify that all drugs



available on the market are free from possibly fatal valvulopathic risk. This predictor will be made publicly available at the ChemBench server established in the Laboratory for Molecular Modeling ([chembench.mml.unc.edu](http://chembench.mml.unc.edu)). We will also gladly apply this predictor to any compound library that may be of interest to any researcher.

**CHAPTER 5**

**AN INTEGRATIVE CHEMOCENTRIC INFORMATICS APPROACH TO  
DRUG DISCOVERY BASED ON STRUCTURAL HYPOTHESIS FUSION:  
IDENTIFICATION AND EXPERIMENTAL VALIDATION OF SELECTIVE  
ESTROGEN RECEPTOR MODULATORS AS LIGANDS OF 5-  
HYDROXYTRYPTAMINE-6 RECEPTORS**

**Introduction**

Target-oriented drug discovery has become one of the most popular modern drug discovery approaches (Connor et al. 2010, Nicholson et al. 2004, Petak et al. 2010, Raamsdonk et al. 2001, Yang et al. 2010). Target-oriented approaches rely on established functional associations between activation or inhibition of a molecular target and a disease. Modern genomics approaches including gene expression profiling, genotyping, genome-wide association, and mutagenesis studies continue to serve as useful sources of novel hypotheses linking genes (proteins) and diseases and providing novel putative targets for drug discovery.

Recently, functional genomics approaches have been increasingly augmented by chemical genomics (Brenner 2004, Darvas et al. 2004, Nislow & Giaever 2003, Salemme 2003, Zheng & Chan 2002b, Zheng & Chan 2002a), i.e., large scale screening of chemical compound libraries in multiple biological assays (Campbell et al. 2010, Hamadeh et al. 2010, Kiessling & Splain 2010, Ogorevc et al. 2010, Wagner & Clemons 2009). Chemical genomics studies yield data indicating that both physical and functional interactions exist

between chemicals and their biological targets. Such data (either obtained in chemical genomics centers or collected and curated from published literature) is deposited in many public and private databases such as the NIMH Psychoactive Drug Screening Program (PDSP) (PDSP 2009), PubChem (PubChem 2009), ChEMBL (ChEMBL 2010), WOMBAT (Olah et al. 2007) and others (see Oprea and Tropsha (Oprea & Tropsha 2006) for a recent review).

Various *in silico* techniques have been exploited for analyzing target-specific biological assay data. A recent publication by Kortagere and Ekins (Kortagere & Ekins 2010) could serve as a good summary of most common target-oriented computational drug discovery approaches including: (1) structure based virtual screening (docking and scoring) using either experimentally characterized (with X-ray or NMR) or predicted by homology modeling structure of the target protein, (2) chemical similarity searching using known active compounds as queries, (3) pharmacophore based modeling and virtual screening, (4) quantitative structure activity relationship (QSAR) modeling, and (5) network or pathway analysis.

Data resulting from large-scale gene or protein expression or metabolite profiling (often collectively referred to as 'omics' approaches (Burgun & Bodenreider 2008, Kandpal et al. 2009, Polychronakos 2008, Vangala & Tonelli 2007) can be explored not only for specific target identification but also in the context of systems pharmacology to identify networks of genes (or proteins) that may collectively define a disease phenotype. For example, 'omics' data can be used to ask what genes or proteins, or post-translationally modified states of proteins are over- (or under-) expressed in patients suffering from a particular disease. These types of data can be found in a number of public repositories such as the Gene Expression

Omnibus (GEO) (Edgar et al. 2002, Barrett & Edgar 2006), GEOmetadb (Zhu et al. 2008b), the Human Metabolome Database (HMDB) (Wishart 2007, Wishart et al. 2009), Kinase SARfari (Kinase SARfari 2010), the Connectivity Map (cmap) (Lamb et al. 2006), the Comparative Toxicogenomics Database (CTD) (Davis et al. 2009), STITCH (Kuhn et al. 2009, Kuhn et al. 2008), GenBank (Burks et al. 1991, Burks et al. 1990), and others.

Insights into disease pathology and underlying mechanisms can be revealed by the disease ‘gene signature’, i.e., those genes whose expression varies consistently between patients and healthy individuals (controls) (Palfreyman et al. 2002). Gene-expression profiling has been often applied to elucidate the mechanisms underlying the roles of biological pathway in a disease (DeRisi et al. 1997, Lamb et al. 2003), reveal arcane subtypes of a disease (Golub et al. 1999, Perou et al. 2000), and predict cancer prognosis (Pomeroy et al. 2002, van, V et al. 2002). At the same time, the treatment of cultured human cells with chemical compounds that target a disease can produce a drug related ‘gene signature’, i.e., differential expression profile of genes in response to the chemical (Altar et al. 2009, Ogden et al. 2004, Palfreyman et al. 2002, Le-Niculescu et al. 2007). Recently, a group of scientists at the Broad Institute have established the Connectivity Map (cmap) database to catalog the biological responses of a large number of diverse chemicals in terms of their gene expression profiles (Lamb et al. 2006). It has been shown that examining the correlations between gene expression profiles characteristic of a disease and those modulated by drugs may lead to novel hypotheses linking chemicals to either etiology or treatments for a disease (Garman et al. 2008, Golub et al. 1999, Hassane et al. 2008, Hieronymus et al. 2006, Lamb et al. 2006, Riedel et al. 2008, Setlur et al. 2008, Zimmer et al. 2008, Zimmer et al. 2010).

The cmap database provides an unusual but intriguing example of what we shall call a *chemocentric* ‘omics’ database and methodology for generating independent and novel drug discovery hypotheses. Indeed, there exists a wealth of information buried in the biological literature and numerous specialized chemical databases (ChEMBL 2010, Daylight 2004, Olah et al. 2007, PDSP 2009, PubChem 2009) linking chemical compounds and biological data (such as targets, genes, experimental biological screening results; cf. Baker and Hemminger (Baker & Hemminger 2010)). The chemocentric exploration of these sources, either individually or in parallel opens up vast possibilities for formulating novel drug discovery hypotheses concerning the predicted biological or pharmacological activity of investigational chemical compounds or known drugs. The integration and cross-validation of such independent structural hypotheses can increase their level of confidence and can be referred to as structural hypothesis fusion.

Herein, we describe a novel integrative chemocentric informatics approach to drug discovery that combines structural hypotheses generated from independent analysis of both traditional target-specific assay data and those resulting from large scale genomics and chemical genomics studies. As a proof of concept, we have focused on the Alzheimer’s disease as one of the most debilitating neurodegenerative diseases with complex etiology and polypharmacology. We have considered and cross-examined two independent but complimentary approaches to the discovery of novel putative anti-Alzheimer’s drugs. First, we have employed a traditional target-oriented cheminformatics approach to discovering anti-Alzheimer’s agents. We have built QSAR models of ligands binding to 5-hydroxytryptamine-6 receptor (5-HT<sub>6</sub>R), a potential target for the cognitive enhancement in Alzheimer’s disease (Geldenhuis & Van der Schyf 2009); it has been shown that 5-HT<sub>6</sub>R

antagonists can improve memory and cognition in animal models of impaired cognition (Holenz et al. 2006). We have then used models developed with the rigorous predictive QSAR modeling workflow established and implemented in our laboratory (Tropsha 2010) for virtual screening (VS) of the WDI (Daylight 2004) and DrugBank (Wishart et al. 2006, Wishart et al. 2008) to identify putative cognition enhancing agents with potential utility as anti-Alzheimer's agents as compounds predicted to interact with 5-HT<sub>6</sub>R. Second, we have explored (chemo)genomic data available from the cmap project (Lamb et al. 2006) to link chemical compounds and the Alzheimer's disease without making explicit hypotheses about target-specific mechanisms of action, i.e., treating Alzheimer's disease as a complex polypharmacological disease.

We then cross-examined and combined common hits regarded as structural hypotheses resulting from both approaches (i.e., hypothesis fusion) towards common integrated higher-confidence hypotheses supported by two independent lines of computationally-based evidence. Thirteen common hits were tested in 5-HT<sub>6</sub>R binding assays at the PDSP and ten were confirmed experimentally as having activity. Unexpectedly, we found that the confirmed actives included several selective estrogen receptor modulators (SERMs) suggesting that they may be potential anti-Alzheimer's drugs as well. Indeed, we have identified clinical evidence in biomedical literature in support of this hypothesis. We believe that approaches discussed in this study can be applied to a large variety of systems to identify novel drug-target-disease associations.

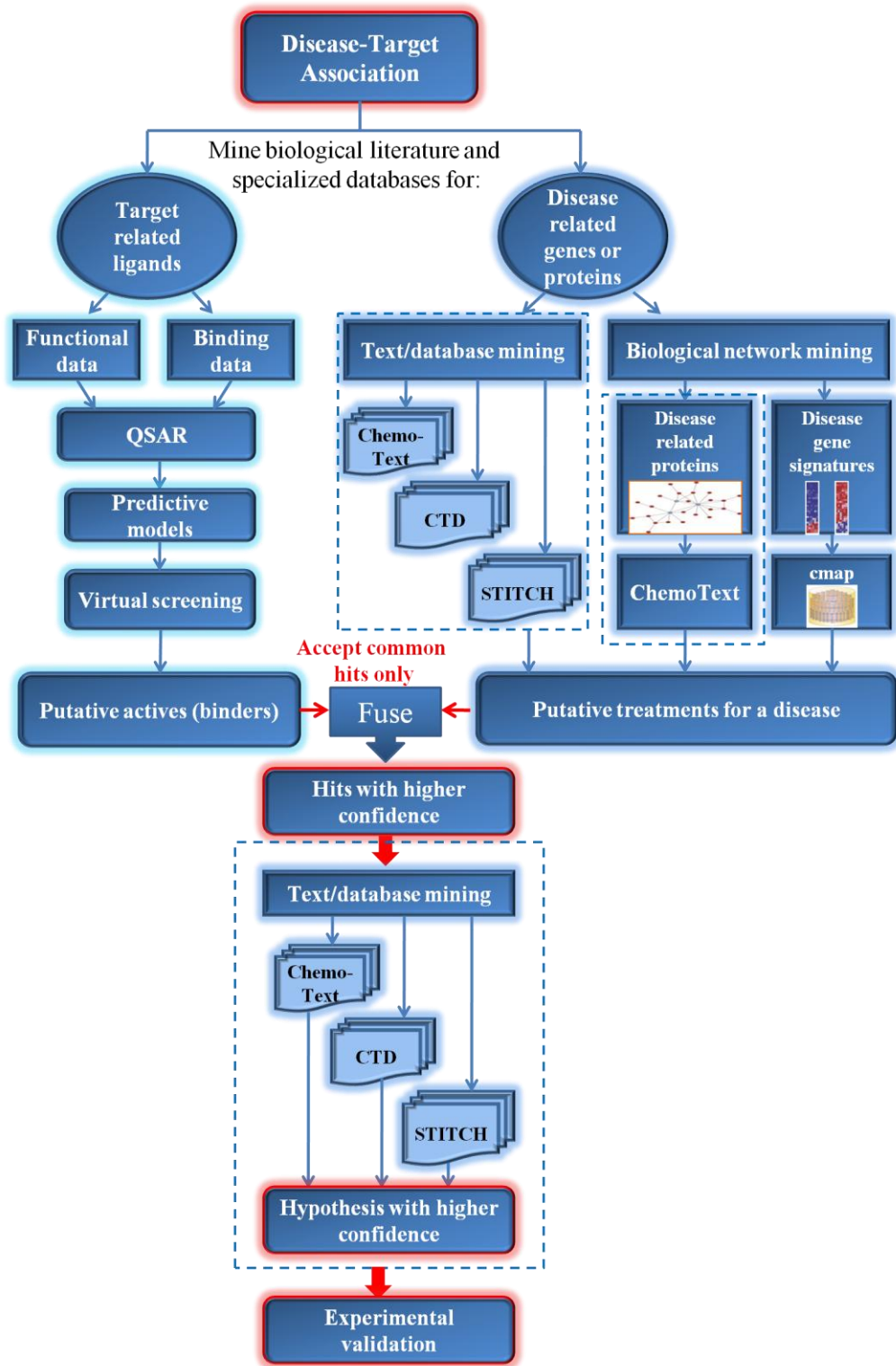
## **Materials and Methods**

### **Integrative Chemocentric Informatics Approach**

We have devised an integrative workflow focused on the discovery of new drug candidates and finding new uses for existing drugs by fusing predictions generated from different data types and methods. Currently, the workflow (Fig. 5.1) incorporates three major components: (1) a module for QSAR-based VS of chemical libraries to identify new ligands for target proteins, (2) a network-mining module to identify small molecule therapeutics for specific diseases without necessarily knowing the underlying target-specific mechanism; this module explicitly relies on cmap,<sup>3</sup> an external online database ([www.broadinstitute.org/cmap/](http://www.broadinstitute.org/cmap/)) that links the effects of different drugs and diseases using gene expression profiles, and (3) ChemoText (Baker & Hemminger 2010), an in-house repository of relationships between chemicals, diseases, proteins, and biological processes. The first two modules have been employed extensively for studies reported herein.

**Figure 5.1.** Study design for the integrative informatics approach for drug discovery integrating network mining, text mining of biological literature, the analysis of disease gene signatures and efficient cheminformatics techniques, to discover novel drugs with desired polypharmacology.





We start our study with identifying established disease-target associations (e.g., 5-HT<sub>6</sub>R is implicated in Alzheimer's disease). Then we mine the biological literature and specialized databases to extract structure activity datasets for ligands known to interact with the biological target of interest. Activity data could be either binding affinities ( $K_i$  values) or functional data (IC<sub>50</sub> values for agonists and antagonists). Binding and functional data could be either continuous (e.g.,  $K_i$  and IC<sub>50</sub> values) or categorical (e.g., binder *vs.* non-binder or agonist *vs.* antagonist) in nature. At this stage we use our QSAR-based VS module (see Fig. 2; predictive QSAR workflow) to generate robust predictive QSAR models for experimental structure activity data that can be employed for VS of chemical libraries to derive new hypotheses about putative actives (binders, or agonists, or antagonists).

Simultaneously, we mine the biological literature for gene signatures associated with the disease and/or for all related protein targets implicated in the disease state. We use these disease related genes and proteins to query specialized databases to extract information about disease-protein (gene)-chemical connections. For example, we use disease gene signatures to query the cmap for putative treatments, and we use related proteins to query ChemoText for related chemicals to establish new disease-protein (gene)-chemical connections. After a thorough analysis of all data, we select hit compounds that are expected to be novel treatments for the disease (cf. Fig. 5.1).

Finally, we fuse hypotheses derived from the QSAR-based VS approach with those derived from text/network mining. Hypothesis fusion is based on structural identity between independently identified hits. The common structural hits are considered for further experimental validation. We assume that the level of confidence in structural hypotheses resulting from independent approaches to knowledge mining in chemocentric databases is

intrinsically higher than that in any computational hit generated in respective independent studies.

## Databases and Datasets

The experimental data for Alzheimer's disease related target 5-HT<sub>6</sub>R were extracted from the PDSP  $K_i$ -DB available in the public domain. The complete 5-HT<sub>6</sub>R dataset included binding affinity data for 250 compounds. We used PubChem (PubChem 2009) to obtain all chemical structures for our datasets in SDF file format. After generating models we used the successful models for virtual screening of WDI and DrugBank.

We also queried cmap database with disease Alzheimer's disease gene signature. Disease gene signatures were populated with Affymetrix U133A probe sets using NetAffx. All details about the databases and tools mentioned herein can be found in Chapter 2.

## Computational Methods

### (1) QSAR Modeling and QSAR-based Virtual Screening

**Preprocessing of the Dataset.** We used a workflow for chemical data curation that was developed in our lab and published recently (Fourches et al. 2010) and discussed in details in Chapter 2. Our analysis resulted in the detection and removal of 56 duplicate chemical entries leaving 194 unique normalized molecular structures. These 194 unique, organic compounds, including 102 binders and 94 non-binders (see Table S1 of Supporting Information) were used for binary QSAR studies. We assigned the 'activity' class for each compound based on its  $K_i$  value(s) obtained from the PDSP and according to PDSP specifications as reported at the PDSP website (<http://pdsp.med.unc.edu/>). Compounds with  $K_i$  values more than or equal to 10  $\mu$ M were considered non-binders and assigned to class 0,

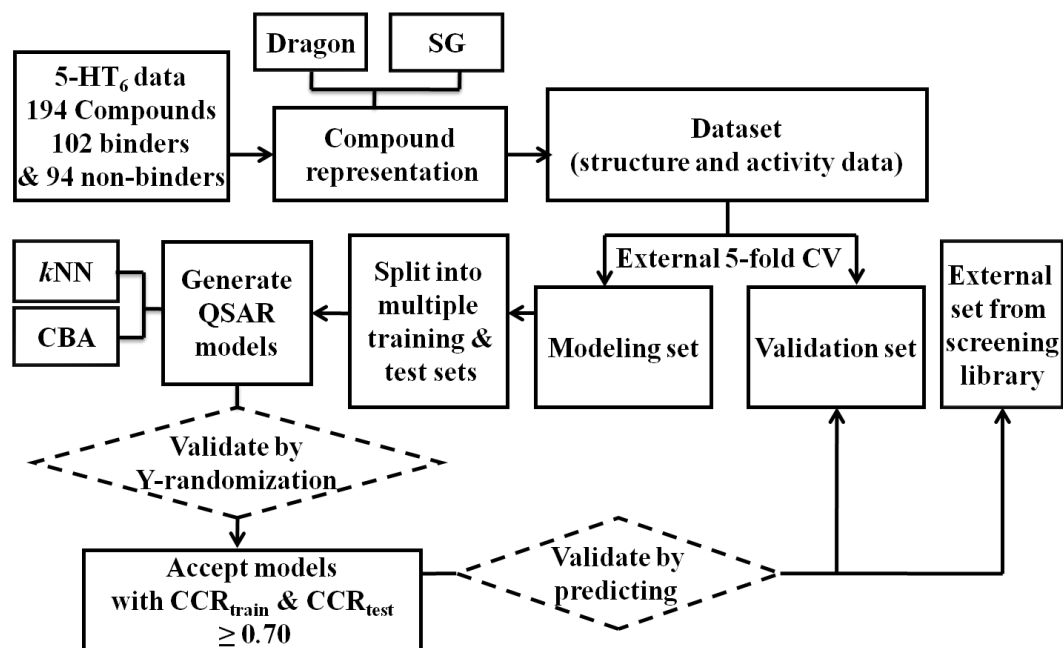
whereas compounds with  $K_i$  values less than 10  $\mu\text{M}$  were considered binders and assigned to class 1.

**Dataset Division for Model Building and Validation.** Following our predictive QSAR modeling workflow (Tropsha 2010), all QSAR models generated to classify 5-HT<sub>6</sub>R binders vs. non-binders were validated by predicting both test and external validation sets. The original dataset of 194 compounds (102 binders and 92 non-binders) was randomly split into 5 different subset of nearly equal size to allow for external 5-fold cross validation (CV) (Hawkins et al. 2003, Kohavi 1995). In this protocol, each subset including 20% of the original dataset was systematically employed as the external validations set while the remaining 80% of the compounds constituted the modeling set.

Another level of internal validation was achieved by comparing model performance for training and test sets. This approach is always employed as a part of our predictive QSAR modeling workflow (Tropsha 2010, Tropsha & Golbraikh 2007) to emphasize the fact that training-set-only modeling is not sufficient to obtain reliable models that are externally predictive (Golbraikh & Tropsha 2002a). Thus, for each collection of descriptors, the modeling sets (each including 80% of the original dataset) were further partitioned into multiple chemically diverse training and test sets of different sizes using the Sphere Exclusion method implemented in our laboratory (Golbraikh & Tropsha 2002b). Only models that were highly predictive on the test sets were retained for the consensus prediction of the external validation sets. Finally, highly predictive models on both external sets were used in consensus fashion for virtual screening of external compound libraries. The model building and validation approach is illustrated schematically in Figure 2.

**QSAR Modeling.** Two QSAR modeling approaches of different nature were used concurrently to generate classification models for 5-HT<sub>6</sub>R binders vs. non-binders (Fig. 5.2). The first approach relied on *k*-nearest neighbor (*k*NN) model optimization method combined with Dragon descriptors, and the second employed classification based on association (CBA) and subgraphs (SG) descriptors.

**Figure 5.2.** The workflow for QSAR model building, validation and virtual screening as applied to 5-HT<sub>6</sub>R dataset.



**Molecular Descriptors.** Both Dragon and SG descriptors were generated for dataset compounds. The final set of Dragon descriptors used in this QSAR study included 331 descriptors. These descriptors were range-scaled so their values ranged from 0 to 1. For generating SG descriptors we used a support value of 15 %, and the defaults for the lower and upper size limits of the generated subgraphs were 2 and 1000 atoms consecutively. The average number of generated SG descriptors was about 400. These descriptors were then used for modeling 5-HT<sub>6</sub>R dataset with Classification Based on Association (CBA) method (Liu et al. 2001). All details about molecular descriptors can be found in Chapter 2.

**Machine Learning Methods.** In this study, we used variable selection classification *k*NN method with Dragon descriptors and the software implemented in our lab (Zheng & Tropsha 2000) to develop QSAR models for 5-HT<sub>6</sub>R binders vs. non-binders. We also used CBA with SG descriptors to build a classifier for 5-HT<sub>6</sub>R binders vs. non-binders. All computational details about these methods are discussed in Chapter 2.

### **Selection and Validation of QSAR Models**

As mentioned earlier, model validation is crucial for QSAR modeling. To evaluate the predictive power of a model, CCR (Eq. 1) values for the training, test, and external validation sets were calculated. We used sensitivity (SE) and specificity (SP) (refer to supplementary material) as well. SE and SP reflect the accuracy of predicting the compounds of binder (class 1) and non-binder (class 0) classes, respectively. We considered a QSAR model to have an acceptable predictive power, if both of the following conditions were satisfied:

(i) CCR for the LOO cross-validation of the training and test sets (i.e., ) were at least 70%

(ii) SE and SP for both training and test sets (i.e., , , , ) were at least 70%.

**Applicability Domains.** We applied AD in this study to avoid unreliable predictions. We defined the AD as a distance threshold  $D_T$  between a compound under prediction and its nearest neighbors of the training set according to Equation 2.9. We set the default of  $Z$  at 0.5. We also defined a global AD in the entire descriptor space. In this case, the same formula (Eq. 2.9) was used,  $Z=0.5$ ,  $k=1$  and Euclidean distances were calculated using all descriptors. Thus, if the distances of the external compound from its  $k$  nearest neighbors (see above) in the training set within either the entire descriptor space or the selected descriptor space exceeded these thresholds, no prediction was made. In this study, applicability domain calculations were carried out using Dragon descriptors and  $k$ NN.

**Robustness of QSAR Models.** Y-randomization (randomization of response) is a widely used approach to validate the robustness of QSAR models (Wold & Eriksson 1995). It consists of rebuilding the models using randomized activities of the training set and subsequent assessment of model statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower values of CCR for the training or the test set than the models built using training set with real activities, or at least these models should not satisfy some of the validation criteria mentioned above. If this condition is not satisfied, models built for this training set with real activities are not reliable and should be discarded.

**Consensus Prediction.** Consensus prediction implies averaging the predictions for each compound by majority voting for classification QSAR models, using all models passing the validation criteria (e.g.,  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  above or equal to 0.70). Our experience suggests that consensus prediction provides the most stable and accurate solutions (Zhu et al. 2008a). In general, in order to determine the confidence in the obtained predictions we need



to define a consensus score. The consensus scores employed in this study take into account the total number of models used to predict a compound's activity, and the number of models that predicted the compound to belong to a specific class. In case of predicted binders (assigned to class 1), we accept predictions made with no less than half of the total acceptable models. Because we define two classes of compounds, i.e., class 1 (binders) and class 0 (non-binders), some models may predict a compound to belong to class 0 and others may predict it to belong to class 1. As a result, consensus scores (CS) between 0 and 1 will be obtained for each of the predicted compounds. As an additional measure of confidence (and an additional applicability domain criterion) we only accepted those predictions that had an average predicted value (consensus score) above 0.70 (for binders) or below 0.30 (for non-binders).

**Virtual Screening.** To identify putative ligands, validated consensus *k*NN-Dragon models generated for 5-HT<sub>6</sub>R ligands were used for virtual screening of the 59000 molecules within both the WDI chemical library (Daylight 2004) and 1300 DrugBank compounds included in the cmap database. The identified hits (by consensus agreement between all accepted *k*NN-Dragon models) were then evaluated additionally using CBA-SG classifier when it was a need to reduce the size of the VS library generated with *k*NN-Dragon models.

## (2) Biological Network Mining

**Querying the cmap with Alzheimer's Disease Gene Signatures.** The cmap (Lamb et al. 2006) was used to discover unexpected connections between chemicals, genes and the Alzheimer's disease by generating a detailed map that links gene patterns associated with Alzheimer's to corresponding patterns produced by drug candidates and a variety of genetic perturbations included in the cmap database. The effects of different drugs and diseases are

described using “genomic signatures” — the full complement of genes that are turned on and off by a particular drug or disease. We start by querying the online database (cmap: <http://www.broadinstitute.org/cmap/>) with Alzheimer’s disease gene signatures. Then, a computer program, that uses sophisticated pattern-matching methods, matches the barcodes based on the patterns shared among Alzheimer’s gene signature and drugs included in the cmap.

**Alzheimer’s Disease Gene Signatures.** In order to query the cmap, a disease gene signature should exist. Two lists of genes are required to perform the query: a list of up-regulated genes and a list of down-regulated genes characteristic of a disease. Query signatures can be obtained from two major sources: (1) biological literature: gene signatures of diseases can be extracted through the National Library of Medicine’s PubMed system ([http:// www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed)), (2) GEO(Edgar et al. 2002, Barrett & Edgar 2006) database: a gene expression/molecular abundance repository supporting MIAME (Brazma et al. 2001) (Minimum Information About a Microarray Experiment) compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. For the purposes of this study, two independent reports of gene-expression changes in brain tissues from Alzheimer’s patients were used to derive gene signatures (i.e., lists of genes up- and down- regulated in Alzheimer’s disease) to query the cmap. Signature 1 (from hippocampus) consisted of 40 genes reported by Hata, R. *et al* (Hata et al. 2001), and signature 2 (from cerebral cortex) consisted of 25 genes reported by Ricciarelli, R. *et al* (Ricciarelli et al. 2004). NetAffx was then used to map gene symbols and Unigene identifiers to populate gene signature lists with Affymetrix U133A probe sets to query the cmap.

### **(3) Hypothesis Fusion**

Data fusion is the process of combining multiple data in order to produce new information that improves the performance of the system (i.e., the *in silico* model or predictor). This fusion approach was first developed for applications in signal processing (Klien 1999) and later it was applied in VS efforts to enable better decisions as to which small number of molecules should go further for biological testing (Sukumar et al. 2008, Whittle et al. 2006a, Whittle et al. 2006b). Herein, we cross-examined and fused structural hypotheses generated independently from both QSAR-based VS and biological network mining efforts to identify and accept common hits only. This step of merging hypotheses was based on structural identity comparisons. All chemical structures of cmap compounds were retrieved from DrugBank (Wishart et al. 2006, Wishart et al. 2008) using their DrugBank identifiers. Identical structures only were then accepted for further analysis. All chemical structures labeled as identical were also subjected to a manual curation step where structures and names of the chemical compounds were compared in different databases to make sure they both refer to the same chemical entity. Common hits were then considered for further experimental validation.

#### **Experimental Validation in Radiologand Binding Assays**

Final common hit compounds from QSAR-based VS and cmap negative connections with Alzheimer's were purchased and submitted to PDSP for experimental target validation. The experimental details are discussed in Chapter 2.

## Results and Discussion

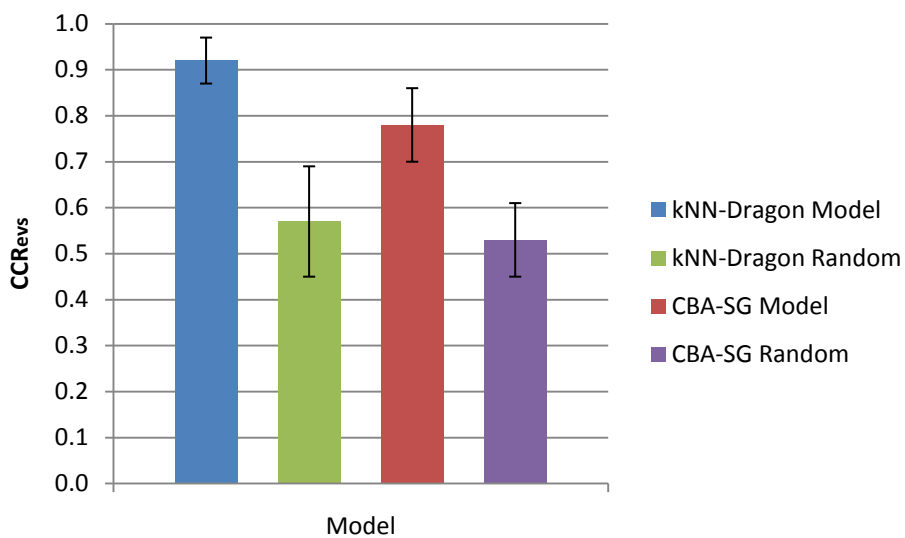
### QSAR Modeling of 5-HT<sub>6</sub>R Binders and Non-binders

*k*NN with Dragon descriptors was employed to classify modeling set compounds into 5-HT<sub>6</sub>R binders vs. non-binders. The five modeling sets derived from applying the external 5-fold CV technique were divided into multiple training and test sets (28-40 divisions) using the Sphere Exclusion algorithm as described in Methods. Multiple QSAR models were generated independently for all training sets and applied to the test sets. Generally, we accept models with CCR values above or equal to 0.70 for both the training and test sets. However, because we were able to generate thousands of acceptable models, we used more conservative criteria (i.e., CCR<sub>train</sub> and CCR<sub>test</sub> above or equal to 0.90) for model selection to predict external compounds. Results of Y-randomization tests confirmed that *k*NN-Dragon classification models with CCR<sub>train</sub> and CCR<sub>test</sub> values above or equal to 0.90 were robust. None of the models with randomized class labels of the training set compounds had CCR<sub>train</sub> and CCR<sub>test</sub> above 0.65 or CCR<sub>evs</sub> above 0.55 for any split.

The CBA method was used to classify the dataset using SG descriptors. The dataset was initially divided, using external 5-fold cross validation technique, into modeling sets with about 155 compounds each and external validation sets containing about 39 compounds each. The modeling sets were then used to build the classifier in CBA(Liu et al. 2001) using an initial pool of about 400 SG descriptors. The classifier gave an average CCR<sub>train</sub> of 0.92 (i.e., the average resulted from five different tests). Then, the external validation set consisting of 39 compounds was used to assess the robustness of the classifier. The average CCR<sub>test</sub> was 0.78, which is not as high the CCR value for the training set, but is still statistically acceptable.

Clearly, *k*NN (mean  $CCR_{\text{evs}} = 0.92$ ) performed better than CBA (mean  $CCR_{\text{evs}}$  equal to 0.78) on the external validation sets (see Fig. 5.3). Therefore, we chose to use *k*NN-Dragon models for VS of external drug libraries. Nevertheless, we maintained CBA-SG models as an additional filter to suggest smaller sets of compounds as 5-HT<sub>6</sub>R putative binders selected from the list of virtual hits obtained with *k*NN-Dragon models and therefore predicted by both models as putative binders.

**Figure 5.3.** Comparison of the QSAR approaches to classify 5-HT<sub>6</sub>R binders vs. non-binders based on CCR<sub>evs</sub>.



## QSAR-Based Virtual Screening

Since our models proved reasonably accurate based on external validation sets, we used the best models to mine two external databases of approved and potential drugs for putative 5-HT<sub>6</sub>R ligands. The use of AD assures reliable predictions by the models. Therefore, we used two types of ADs in the virtual screening of compound databases. First, we used a global AD that acted as a filter and ensured some level of global similarity between the predicted compounds and the compounds in the modeling set. Second, we defined a local AD for each of the individual classification models.

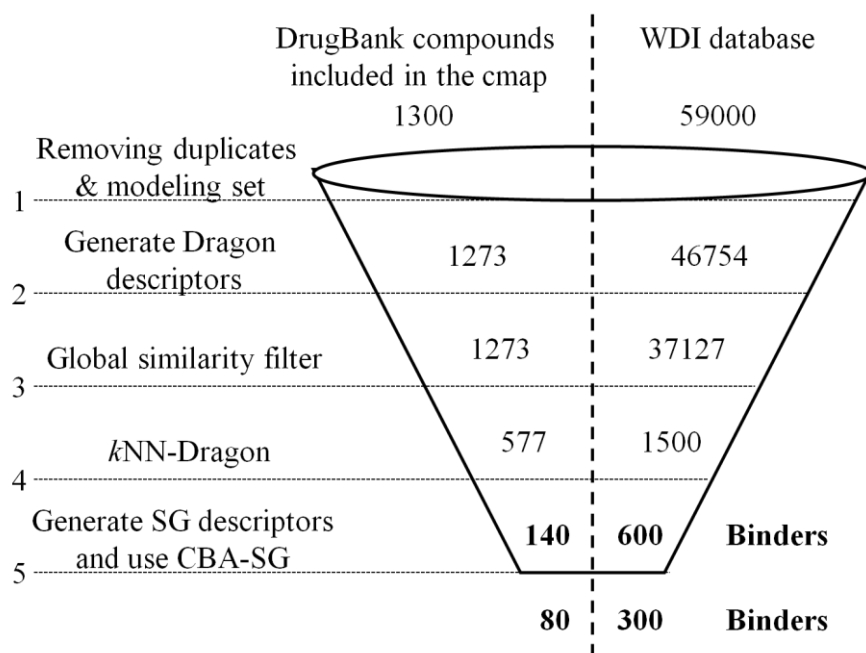
We first screened the WDI database of about 59000 compounds (approved or investigational drugs) (Fig. 5.4). This original collection had many duplicates (i.e., many salt forms for the same chemical entity), and these duplicates were removed using MOE. We also removed all compounds included in our modeling and external validation sets. Dragon descriptors were generated for the remaining 46859 unique compounds in the database; of these, 9732 compounds were excluded because Dragon was unable to calculate at least one of the descriptors generated for the modeling set. The remaining 37127 compounds were then subjected to a global AD filter for the modeling set using a strict Z cutoff of 0.5 (which formally places the allowed pairwise distance threshold at the mean of all pairwise distance distribution for the training set plus one-half of the standard deviation). Then, all *k*NN-Dragon models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  above or equal to 0.70 were employed in consensus fashion to predict 1500 compounds remaining after several filtering steps, which resulted in the identification of the 600 predicted binders. In an effort to reduce the number of hits, we have generated SG descriptors for these 600 molecules and applied the CBA-SG classifier which filtered out half of these compounds, leaving 300 compounds as putative binders for 5-

HT<sub>6</sub>R. None of these hits was tested in this manuscript since we explicitly focused on compounds from DrugBank that were employed in the cmap project. These VS hits from WDI should be viewed as structural hypotheses awaiting the experimental confirmation; the report on these studies is reserved for separate future publication.

Additionally, we screened 1300 DrugBank compounds included in the cmap database. Dragon descriptors were computed for 1273 unique compounds. These compounds were then subjected to a global AD filter for the modeling set using a strict *Z* cutoff of 0.5. Consequently, we placed the allowed pairwise distance threshold at the mean of all pairwise distance distribution for the training set plus one-half of the standard deviation which resulted in 577 predictions within the applicability domain. Next, validated consensus *k*NN-Dragon models (i.e., all models with  $CCR_{\text{train}}$  and  $CCR_{\text{test}}$  above or equal to 0.90) were used to predict these 577 compounds, resulting in the identification of 140 unique compounds predicted to be 5-HT<sub>6</sub>R binders. We did not apply the CBA-filter here because, for the subsequent integration with the cmap mining results, we wanted to explore a larger set of all 140 compound hits (i.e., putative 5-HT<sub>6</sub>R binders) included in the cmap datasets predicted by *k*NN-Dragon models.



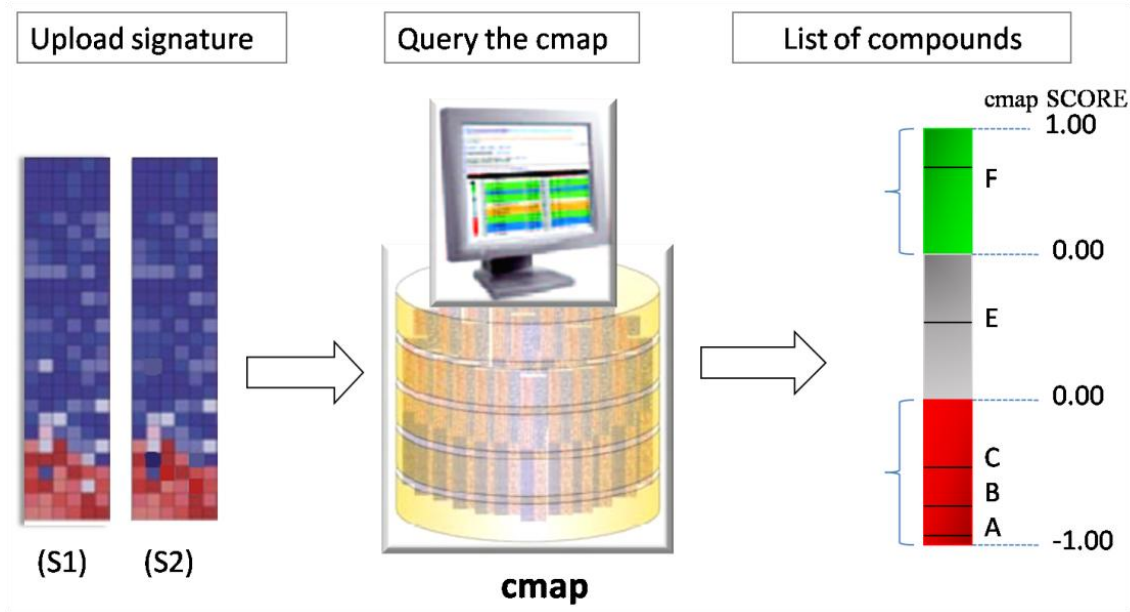
**Figure 5.4.** A representation for QSAR-based virtual screening steps of two chemical databases: the WDI and DrugBank compounds included in the cmap.



## Searching the Connectivity Map for Potential Anti-Alzheimer's Agents

We used two gene signatures for the Alzheimer's disease (designated a S1 and S2) to query the cmap database in an attempt to link genes associated with the disease to potential therapeutic agents. These two signatures were based on two independent rank-ordered gene lists provided by two different Gene Set Enrichment Analysis (GSEA) studies (Hata et al. 2001, Ricciarelli et al. 2004). The two disease signatures were compared with predefined signatures of therapeutic compounds included in the cmap and ranked according to a connectivity score (ranging from +1 to -1), representing relative similarity to the disease gene lists. The connectivity score itself is derived using a nonparametric, rank-based, pattern-matching strategy based on the Kolmogorov-Smirnov statistic (Hollander & Wolfe 1999). Connectivity scores are calculated using the online tools available at the cmap (<http://www.broadinstitute.org/cmap/>). All instances in the database are then ranked according to their connectivity scores; those at the top (+) are most strongly correlated to the query signature and looked at as disease causes, and those at the bottom (-) are most strongly anti-correlated and considered as possible therapeutics (see Fig. 4.5 for concept).

**Figure 5.5.** Querying the connectivity map with Alzheimer's disease gene signatures (S1 and S2).



The majority of chemicals included in the cmap database are represented by multiple independent replicates. Most compounds are profiled in three different cell lines, some at different concentrations. These are called ‘instances’ for the same chemical which are defined as “a treatment and control pair and the list of probe sets ordered by their extent of differential expression between this treatment and control pair” (The connectivity map 2010). The instance is the basic unit of data and metadata in cmap. Instances of the same compound might have similar or dissimilar connectivity scores with the query signature. We have higher confidence in the derived connections when gene signatures are conserved across diverse cell types and experimental settings. However, Lamb and colleagues (Lamb et al. 2006, Lamb 2007) indicated that the non-consistent scoring of different instances of the same chemical may represent either (1) a cellular-context dependent difference in activity, (2) a concentration-discriminated effect, or (3) poor reproducibility between replicates. Therefore, ‘best’ connections are those where multiple, autonomous instances of the same chemical have consistently high (or low) scores. However, inconsistently scoring compounds should not necessarily be dismissed since their significance as potential treatments for a disease can be boosted by additional evidence, such as predictions from QSAR models.

In this study, we were interested in compounds whose chemogenomics profiles were negatively correlated with the Alzheimer’s disease gene signatures. Hits with statistically significant, negative connectivity scores could be potential treatments for the Alzheimer’s disease; however, the list of negatively correlated molecules might be long and must be analyzed carefully before suggesting hypotheses of possible mechanisms for controlling or mediating the disease. Examples of top negative connections with both signatures S1 and S2

are shown in Tables 5.1 and 5.2, respectively. Although the two gene signatures (i.e., for the Alzheimer's disease) used to query the cmap shared no common genes, both queries resulted in a common list of negative connections which were given a higher confidence in our studies. All chemical structures for each chemical compound included in the cmap were obtained from the DrugBank and mapped based on the DrugBank identifiers provided by the cmap database.

**Table 5.1.** Top twenty negative connections from the cmap with S1.

Compound	Rank <sup>a</sup>	Cell	Score	Instance_ID
Naproxen	6100	PC3	-1	7146
Sulfacetamide	6099	MCF7	-0.990	1695
Amprolium	6098	HL60	-0.930	1979
Aminogluthetimide	6097	MCF7	-0.913	7463
Ioxaglic acid	6096	HL60	-0.897	2966
Dexpanthenol	6095	MCF7	-0.871	7455
Suxibuzone	6094	MCF7	-0.870	7163
Chlorphenesin	6093	HL60	-0.862	1432
Metixene	6092	HL60	-0.853	2451
Fulvestrant	6091	MCF7	-0.843	5565
Seneciophylline	6090	MCF7	-0.841	2797
Troglitazone	6089	MCF7	-0.839	6991
Dicloxacillin	6088	HL60	-0.834	2445
Phentolamine	6087	HL60	-0.831	2362
Monocrotaline	6086	MCF7	-0.828	6771
Lymecycline	6085	HL60	-0.823	2953
Bezafibrate	6084	PC3	-0.815	6653
6-Benzylaminopurine	6083	HL60	-0.812	2351
Terbutaline	6082	MCF7	-0.811	3202
Clorgiline	6081	MCF7	-0.805	3219

<sup>a</sup>The rank order is generated from estimating the connectivity scores of 6100 individual treatment instances with S1. A rank order of 6100 corresponds to the compound with the strongest negative connectivity S1, while a rank order of 1 corresponds to the compound with the strongest positive connectivity with S1.

**Table 5.2.** Top twenty negative connections from the cmap with S2

Compound	Rank <sup>a</sup>	Cell	Score	Instance_ID
Trifluoperazine	6100	HL60	-1	2389
Clomifene	6099	MCF7	-0.982	4994
Ethotoin	6098	HL60	-0.977	2196
Sulfafurazole	6097	HL60	-0.973	1603
Quercetin	6096	MCF7	-0.964	4846
Triflusal	6095	HL60	-0.925	1717
Alfuzosin	6094	PC3	-0.903	4644
Metitepine	6093	HL60	-0.890	1616
Trioxysalen	6092	MCF7	-0.885	6216
LY-294002	6091	MCF7	-0.883	258
Tanespimycin	6090	HL60	-0.873	6184
Spirolactone	6089	MCF7	-0.871	6255
Nifurtimox	6088	MCF7	-0.859	4953
Iobenguane	6087	HL60	-0.847	1729
U0125	6086	PC3	-0.845	663
Monorden	6085	MCF7	-0.841	5947
Primidone	6084	PC3	-0.833	6723
Calcium pantothenate	6083	MCF7	-0.828	4775
Phthalylsulfathiazole	6082	HL60	-0.826	3033
Ceforanide	6081	PC3	-0.824	6751

<sup>a</sup>The rank order is generated from estimating the connectivity scores of 6100 individual treatment instances with S2. A rank order of 6100 corresponds to the compound with the strongest negative connectivity S2, while a rank order of 1 corresponds to the compound with the strongest positive connectivity with S2.

## **Hypothesis Generation: Integrating and Fusing Independent Hypotheses from QSAR-Based VS and cmap Analysis**

We fused hypotheses produced from two different datasets and using two different computational methods (Fig. 5.1): (1) QSAR-based datamining of chemical databases in an effort to identify novel ligands for 5-HT<sub>6</sub>R, and (2) Network-mining using two signatures for Alzheimer's disease to query the cmap and identify possible anti-Alzheimer's therapeutics. Our procedure for hypothesis fusion was based on structural identity for chemical compounds derived from both approaches mentioned above. Compounds with negative connectivity scores, representing genes expressed in an opposite fashion to the imported Alzheimer's disease query—which implies their potential benefits to be candidate treatments, were compared with 5-HT<sub>6</sub>R hits predicted from QSAR-based VS.

The primary goal for fusing hypotheses in this study was initially to overcome some of the inherent hit scoring problems in classification QSAR, and to achieve higher success rates in experimental testing of the VS hits. In other words; we often select for further experimental validation those QSAR hits with consensus scores above or equal to 0.90 (refer to consensus scores in methods). However, many novel scaffolds that are significantly different (i.e., structurally and therapeutically) from the training set compounds, might have lower consensus scores ranging from 0.50 to 0.90 despite the fact that they might be binders too. Thus, this process of fusing hypotheses derived independently from different types of data and using multiple prediction methods, allowed us to fish out these low-confidence QSAR hits (that yet could be highly important ligands) for further analysis. As a result, we posit that fusing independent hypotheses is likely to improve the overall success rates of *in silico* lead identification.

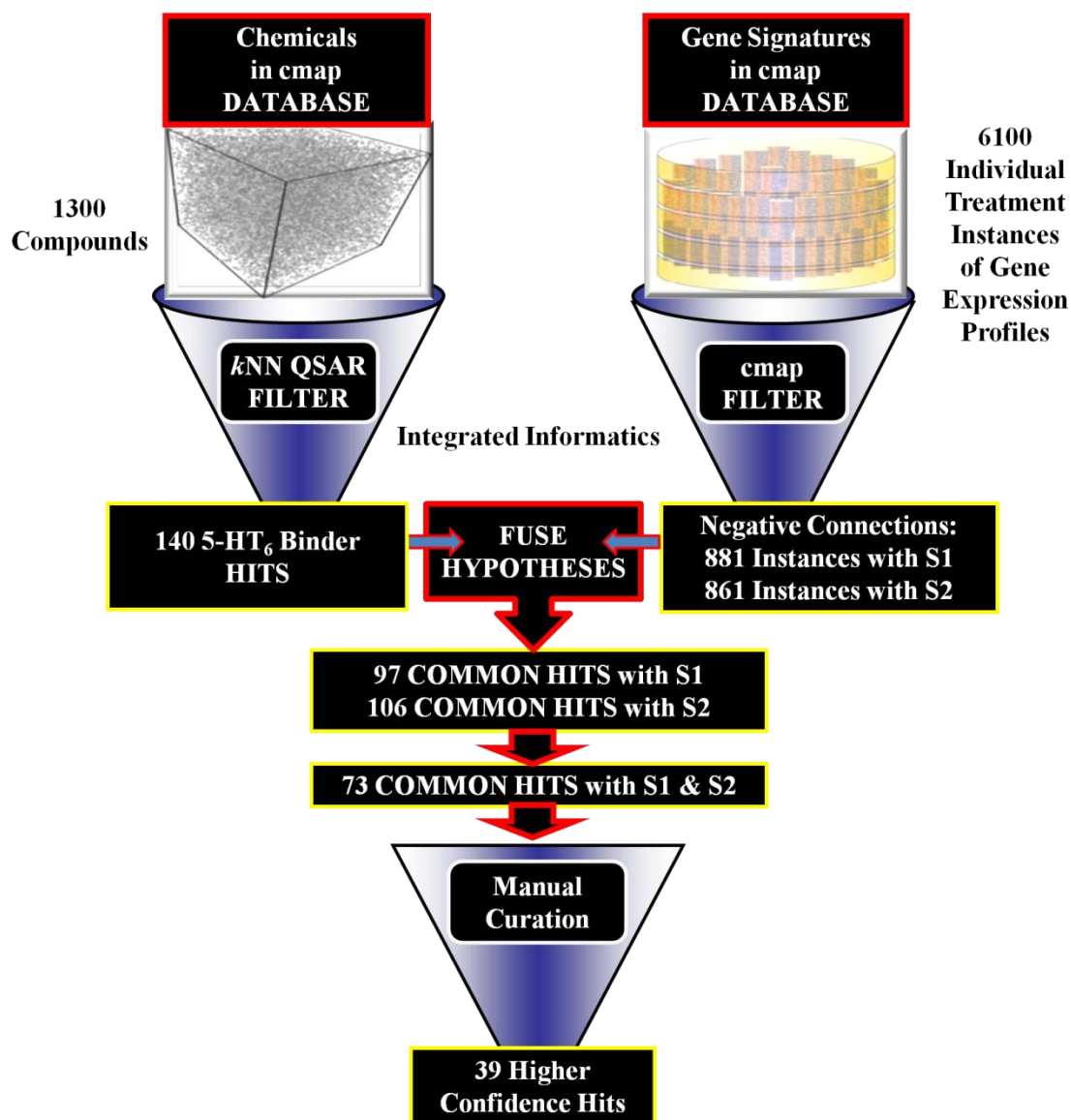


Additionally, our approach could be used to aid in the process of prioritizing connections from the cmap that might be difficult to call otherwise, especially as the size of the database continues to grow. In diseases like the Alzheimer's, with little knowledge about specific etiology, and the lack of drug gene signatures generated from neuronal cell lines, it is hard to decide *a priori* which negative connections are more important to be viewed as potential therapeutics. Thus, fusing hypotheses derived independently from cmap and QSAR should enable us to increase the confidence in recovered connections.

### **Scoring and Fusing Structural Hypotheses to Identify Anti-Alzheimer's Agents**

Our method for decision fusion was derived from a combination of voting and statistical metrics. In the first step, we used two different scoring functions to rank the computational hits generated independently from both QSAR and cmap. In the QSAR study, we used the *k*NN 'consensus score' which takes into account the total number of models used to predict compound's activity, and the number of models that predicted the compound to belong to a specific class correctly. We considered all computational hits that had an average predicted value (i.e., consensus score) above or equal to 0.50 for further inspection. Our analysis resulted in 140 putative 5-HT<sub>6</sub>R binders among cmap compounds and with *k*NN consensus scores ranging from 0.50-1.00 (see Fig. 5.6).

**Figure 5.6.** The workflow for fusing hypotheses from QSAR modeling and cmap negative connections.



On the other hand, we used the connectivity scores (Lamb et al. 2006) to rank the hits resulted from querying the cmap with Alzheimer's disease gene signatures. Because we were interested in identifying novel treatments for Alzheimer's disease, we ranked hits with larger (-) connectivity scores at the top and gave them higher confidence. Such compounds were hypothesized to have higher chances to reverse the Alzheimer's gene signatures and therefore might have immense therapeutic value in Alzheimer's disease. We considered for further analysis all compounds that had at least one instance of negative connection with any of the two gene signatures used to query the cmap (S1 and S2) so that not to miss any important connections. Our analysis resulted in identifying 881 negative connectivity instances with S1 and 861 instances with S2 (Fig. 5.6).

Finally, we fused the hypotheses generated from both QSAR and cmap analyses and accepted common hits only. We identified 97 compounds that were both predicted to be active at 5-HT<sub>6</sub>R and had at least one instance of negative connectivity with S1 and 106 compounds that had at least one instance of negative connectivity with S2. Accepting only common hits among S1 and S2 resulted in 73 putative hits (see Fig. 5.6). At this stage we applied a manual curation where we inspected all available data for these 73 hits. Each of the 73 common hits had three scores (*k*NN consensus score, cmap connectivity score with S1, and cmap connectivity score with S2) to be considered in the final decision to prioritize hits for further testing. Therefore, we estimated the average connectivity scores for all predicted hits across all treatment instances for each of the S1 and S2 hits. Then we excluded those compounds that had high positive connectivity scores in some treatment instances of the same compound. Finally, we retained 39 compounds that had acceptable negative average connectivity scores at least with one signature (see Fig. 5.6). We hypothesized that these

compounds could be looked at as putative 5-HT<sub>6</sub>R hits and potential cognitive enhancements. One of the final 39 hits, vinpocetine, worth special attention as there is new evidence that has just emerged indicating its potential role in the treatment of Parkinson's disease and Alzheimer's disease (Jeon et al. 2010, Medina 2010). Details on all these 39 VS hits are provided in Tables 5.3 and 5.4.

Each of the 39 common hits had three scores (*k*NN consensus score, cmap connectivity score with S1, and cmap connectivity score with S2) to be considered in the final decision to prioritize hits for further experimental testing. We plotted the mean connectivity scores vs. *k*NN QSAR consensus scores generating separate plots for S1 and S2 (see Fig. 5.7) to analyze these hits in further details.

**Table 5.3.** Final thirty nine computational hits from QSAR-based VS and cmap.

<b>cmap Name</b>	<b>cmap Score 1</b>	<b>cmap Score 2</b>	<b>Num. <i>k</i>NN Models</b>	<b><i>k</i>NN CS</b>	<b><i>k</i>NN Pred</b>	<b>CBA Pred</b>
Acepromazine	-0.528	-0.496	441	0.93	B	B
Alimemazine	-0.121	-0.117	438	1.00	B	B
Astemizole	-0.349	-0.237	328	0.91	B	B
Bepidil	-0.134	-0.409	393	0.89	B	B
Bromperidol	-0.239	-0.213	428	0.83	B	B
Cetirizine	-0.495	-0.327	421	0.92	B	B
Chlorprothixene	-0.277	-0.298	442	0.90	B	B
Cinchocaine	-0.004	-0.335	423	0.58	B	B
Cinnarizine	-0.349	-0.149	414	0.98	B	B
Citalopram	-0.003	-0.260	429	0.71	B	NB
<b>1</b> (Clomiphene)	-0.378	-0.265	409	0.91	B	B
<b>2</b> (Clomipramine)	-0.192	-0.310	437	0.96	B	B
Cloperastine	-0.273	-0.353	443	0.88	B	B
<b>3</b> (Clozpine)	0.093	-0.058	422	0.97	B	B
Diltiazem	-0.128	-0.336	433	0.72	B	NB
<b>4</b> (Doxepin)	0.027	-0.259	444	0.95	B	B
<b>5</b> (Fendiline)	-0.303	-0.228	393	0.84	B	NB
Flavoxate	-0.127	-0.112	403	0.71	B	NB
<b>6</b> (Fluspirilene)	0.055	-0.138	351	0.98	B	B
Imipramine	-0.400	-0.214	427	0.96	B	B
Laudanosine	-0.226	-0.174	411	0.78	B	NB
<b>7</b> (LY-294002)	-0.028	-0.078	428	0.71	B	B
Meclozine	-0.365	-0.171	439	0.95	B	B
Mepacrine	-0.236	-0.301	418	0.53	B	B
Methylergometrine	-0.400	-0.509	441	0.98	B	B
Naftifine	-0.198	-0.148	359	0.85	B	B
<b>8</b> (Nortriptyline)	0.011	-0.354	433	0.93	B	B
Phenoxybenzamine	-0.461	-0.309	444	0.79	B	NB
Piperidolate	-0.168	-0.119	431	0.69	B	NB
<b>9</b> (Prestwick-559)	-0.247	-0.206	435	0.96	B	B
Prestwick-685	-0.086	-0.213	381	0.72	B	B
Promazine	-0.210	-0.307	424	0.97	B	B
<b>10</b> (Raloxifene)	0.047	-0.058	356	0.56	B	NB
<b>11</b> (Tamoxifen)	0.300	-0.220	435	0.93	B	B
Telenzepine	-0.419	-0.114	387	0.68	B	B
Terfenadine	-0.183	-0.512	416	0.51	B	NB
Vanoxerine	-0.450	-0.233	374	1.00	B	NB
Vinpocetine	-0.177	-0.132	376	0.76	B	B
<b>13</b> (Zuclopenthixol)	-0.152	0.144	434	0.98	B	B

**Table 5.4.** Therapeutic classes of the thirty nine final computational hits from QSAR-based VS and cmap.

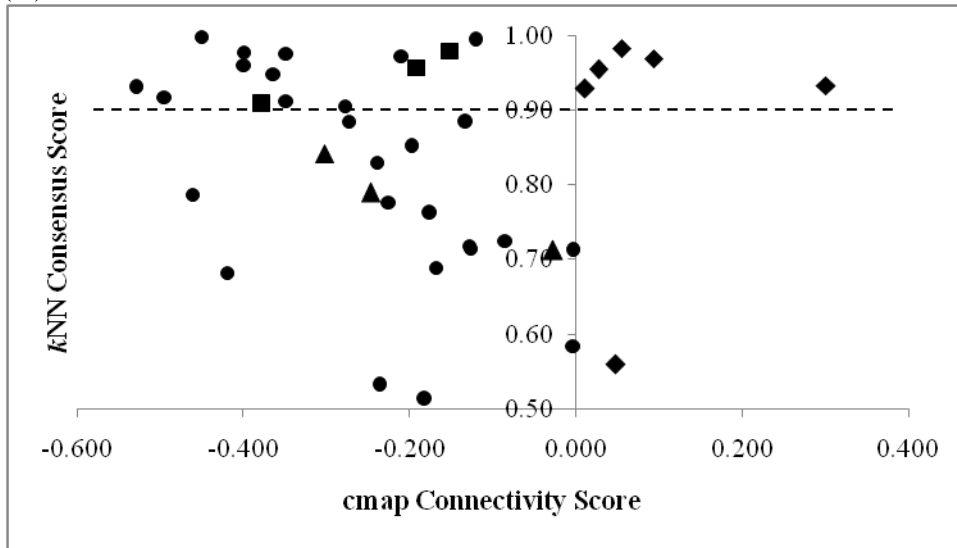
<b>cmap Name</b>	<b>Therapeutic Class/Use</b>
Acepromazine	Antipsychotic
Alimemazine	Antipruritic, sedative, hypnotic and anti-emetic
Astemizole	Anti-Histamine
Bepidil	Calcium channel blocker once used to treat angina
Bromperidol	Neuroleptic, used as an antipsychotic in the treatment of schizophrenia
Cetirizine	Second-generation antihistamine
Chlorprothixene	Typical antipsychotic drug of the thioxanthene class
Cinchocaine	Local anesthetic
Cinnarizine	Antihistamine which is mainly used for the control of nausea and vomiting due to motion sickness
Citalopram	Antidepressant drug of the selective serotonin reuptake inhibitor (SSRI) class
<b>1</b>	SERM
<b>2</b>	Tricyclic antidepressant
Cloperastine	Cough suppressant.
<b>3</b>	Atypical antipsychotics
Diltiazem	Calcium channel blocker
<b>4</b>	Psychotropic agent with tricyclic antidepressant and anxiolytic properties
<b>5</b>	Calcium channel blocker
Flavoxate	Anticholinergic with antimuscarinic effects
<b>6</b>	Antipsychotic
Imipramine	Tricyclic antidepressant
Laudanosine	Benzyltetrahydroisoquinoline alkaloid. Interacts with GABA, opioid, and nicotinic acetylcholine receptors
<b>7</b>	Morpholino derivative of quercetin. It is a potent inhibitor of phosphoinositide 3-kinase s (PI3Ks)
Meclozine	Antihistamine considered to be an antiemetic
Mepacrine	Antiprotozoal, antirheumatic and an intrapleural sclerosing agent. It is known to act as a histamine N-methyltransferase inhibitor
Methylergometrine	Psychedelic alkaloid
Naftifine	Allylamine antifungal drug
<b>8</b>	Second-generation tricyclic antidepressant
Phenoxybenzamine	Non-specific, irreversible alpha antagonist

Piperidolate	Antimuscarinic.
<b>9</b>	Drug used in scientific research which acts as a moderately selective dopamine D <sub>3</sub> receptor partial agonist.
Prestwick-685	Not reported in the literature
Promazine	Antipsychotic
<b>10</b>	SERM
<b>11</b>	SERM
Telenzepine	Anticholinergic or sympatholytic
Terfenadine	Antihistamine formerly used for the treatment of allergic conditions
Vanoxerine	Piperazine derivative which is a potent and selective dopamine reuptake inhibitor (DRI)
Vinpocetine	Vinpocetine has been identified as a potent anti-inflammatory agent that might have a potential role in the treatment of Parkinson's disease and Alzheimer's disease (Jeon et al. 2010, Medina 2010)
<b>13</b>	Typical antipsychotic drug

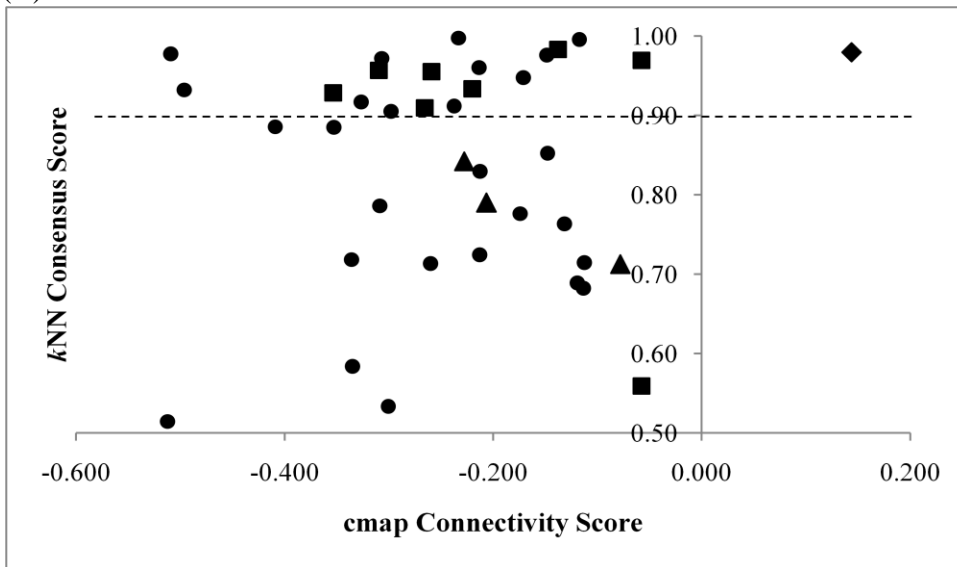
**Figure 5.7.** Plots for *k*NN scores vs. cmap connectivity scores for 39 final common hits from QSAR-based VS and cmap for: (A) Alzheimer's disease signature S1, and (B) Alzheimer's disease signature S2. Squares: compounds predicted and validated as 5-HT<sub>6</sub>R binders having negative connectivity scores with Alzheimer's disease gene signatures; diamonds: compounds predicted and experimentally validated as 5-HT<sub>6</sub>R binders but having positive connectivity scores with one of the Alzheimer's disease gene signatures; triangles: compounds predicted as 5-HT<sub>6</sub>R binders having negative connectivity scores with Alzheimer's disease gene signatures but found non-binders in radioligand binding assays against 5-HT<sub>6</sub>R; circles: compounds predicted as 5-HT<sub>6</sub>R binders which have negative connectivity scores with Alzheimer's disease gene signatures but were not experimentally tested.



(A)



(B)



Additionally, another level of confidence was achieved (besides considering both *k*NN CS and cmap scores) by giving more emphasis to molecules that belonged to the same pharmacological or therapeutic group or had very high structural similarity to hits of higher confidence. This step permitted the retrieval of some compounds that had less significant negative connectivity scores with the disease (e.g., null connectivity or even low positive connectivity scores in few instances). We noticed that the 39 putative binders belonged to several major therapeutic groups (see Table 5.4): antipsychotics, antidepressants, anti-histamines, selective estrogen receptor modulators (SERMs) and calcium channel blockers. Predicting antipsychotics, antidepressants and anti-histamines was not surprising as it is known that many of these compounds are active at 5-HT<sub>6</sub>R receptors (Roth et al. 1994) and because the modeling set of compounds in the QSAR study belonged to these classes of compounds. However, it was unexpected that SERMs are predicted to have activity at 5-HT<sub>6</sub>R.

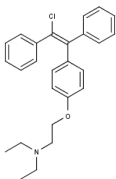
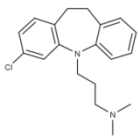
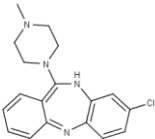
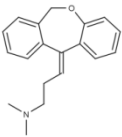
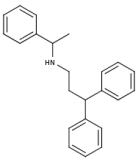
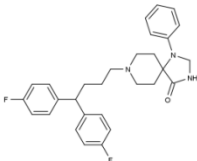
### **Hypothesis Testing: Evaluation of Computational Hits at Human Cloned 5-HT<sub>6</sub> Receptors**

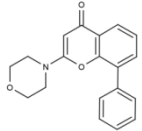
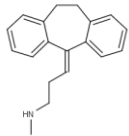
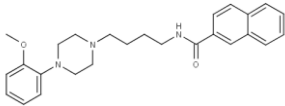
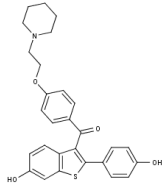
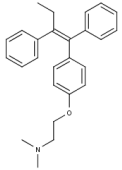
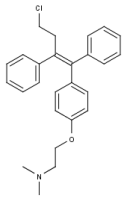
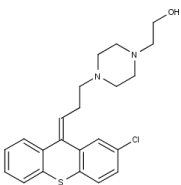
Common hits from QSAR-VS studies and cmap were taken forward for biological validation, in binding assays, for 5-HT<sub>6</sub>R. As discussed above, we identified 39 chemicals, out of 59000 molecules included in the WDI (Daylight 2004), (and out of 1300 compounds included in the cmap), as consensus hits and putative binders for 5-HT<sub>6</sub>R with higher chances of having potential therapeutic effects in Alzheimer's disease; none of these hits was included in the training set used to develop QSAR models. Then, we prioritized thirteen compounds for further experimental validation in 5-HT<sub>6</sub>R radioligand binding assays (Table 5.5). Our final selection was based on different criteria: (1) we tested some compounds with

high consensus scores and stronger negative connectivity with Alzheimer's disease, (2) some compounds were selected because they belonged to the same therapeutic class as several other predicted hits and were not known before to bind to 5-HT<sub>6</sub>R such as selective estrogen receptor modulators (SERMs) (e.g., clomifene, tamoxifen and toremifene), (3) we tested some compounds with low *k*NN CS (e.g., raloxifene having *k*NN CS of 0.58) if other hits that belonged to the same therapeutic class had high consensus scores (e.g., tamoxifen and toremifene having *k*NN CS above to or equal 0.93 and clomiphene having a *k*NN CS of 0.91), (4) we also tried to test predictions that had strong negative connectivity scores with one query signature but had much weaker negative connectivity with the second signature to see if there is one specific signature that was generating better results.

We found that ten of these thirteen predicted actives were confirmed experimentally to inhibit 5-HT<sub>6</sub>R radioligand binding thereby achieving a success hit rate of 77 % in this proof-of-concept study (see Table 5.5).

**Table 5.5.** Experimental validation results for the thirteen computational hits predicted as 5-HT<sub>6</sub>R ligands and had negative connections with Alzheimer's disease gene signatures.

Cp. ID	Compound	PDSP ID CID <sup>a</sup>	Score1 <sup>b</sup> /Cell Score2 <sup>c</sup> /Cell	kNN CS <sup>d</sup>	CBA Pred. <sup>e</sup>	Ki (nM)
1	 Clomifene	13499 1548953	-0.602/PC3 -0.982/ MCF7	0.91	B <sup>f</sup>	1,956.0
2	 Clomipramine	13494 2801	-0.768/PC3 -0.814/MCF7	0.96	B	112.0
3	 Clozapine	24842 2818	-0.590/PC3 -0.652/MCF7	0.97	B	17.0 <sup>g</sup>
4	 Doxepin	13495 667477	-0.463/MCF7 -0.777/HL60	0.95	B	105.0
5	 Fendiline	14821 3336	-0.520/MCF7 -0.683/HL60	0.84	NB <sup>h</sup>	NB
6	 Fluspirilene	14815 3396	-0.493/MCF7 -0.551/HL60	0.98	B	1,188.0

7		13502 3973	-0.790/MCF7 -0.883/MCF7	0.69	B	NB
8	LY-294002 	13503 4543	-0.555/PC3 -0.586/MCF7	0.96	B	214.0
9	Nortriptyline 	13498 3038495	-0.741/MCF7 -0.619/HL60	0.79	B	NB
10	Prestwick-559 	13505 5035	-0.626/HL60 -0.619/HL60	0.56	NB	750.0
11	Raloxifene 	13506 2733526	0/MCF7 -0.531/MCF7	0.93	B	1,041.0
12	Tamoxifen 	16514 3005573	N/A N/A	0.93	B	4,125.0
13	Toremifene <sup>i</sup> 	13510 5311507	-0.609/PC3 -0.746/HL60	0.98	B	169.0

Zuclopenthixol

---

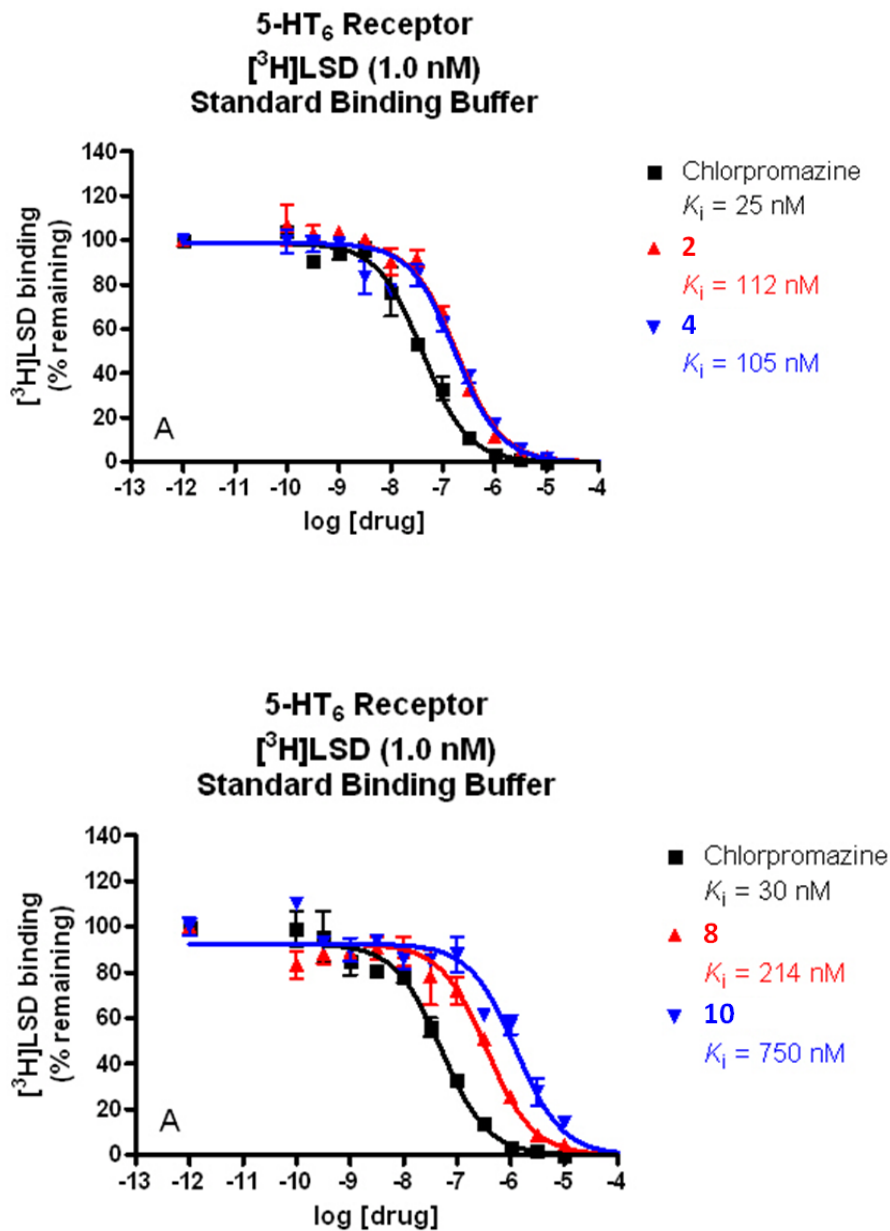
Success Rate	Success Rate
77 % for predictions	100% for predictions
with <i>k</i> NN CS $\geq$ 0.5	with <i>k</i> NN CS $\geq$ 0.9

---

<sup>a</sup>CID, PubChem compound ID; <sup>b</sup>cmap score1, the highest negative connectivity score for this compound with S1 (or the smallest positive in case all other scores are positive); <sup>c</sup>cmap score2, the highest negative connectivity score for this compound with S2 (or the smallest positive in case all other scores are positive); <sup>d</sup>CS, consensus score; <sup>e</sup>CBA pred., predicted binding to 5-HT<sub>6</sub> receptors by CBA; <sup>f</sup>B, binder; <sup>g</sup>PDSP certified data; <sup>h</sup>NB, non-binder; <sup>i</sup>Toremifene was not included in the cmap but was prioritized because 3 other related SERMs were hits from both cmap and QSAR-based VS.

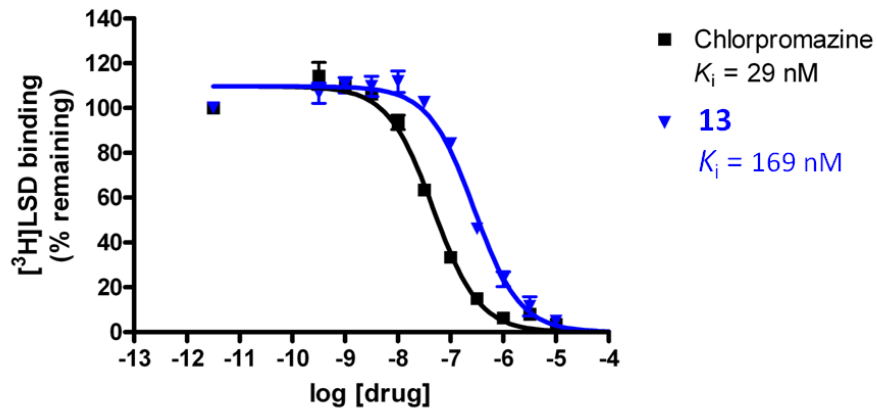
One of these ten confirmed hits was clozapine, which is known to bind 5-HT<sub>6</sub>R but was not included in our training set. Binding affinity ( $K_i$ ) values for the nine predicted hits were in the range 17 - 4125 nM, with six compounds having  $K_i$  values < 1  $\mu$ M. These six highest affinity compounds were: clozapine ( $K_i=17$  nM), doxepin ( $K_i=105$  nM, Fig. 5.8 (A)), clomipramine ( $K_i=112$  nM, Fig. 5.8 (B)), zuclopenthixol ( $K_i=169$  nM, Fig. 5.8 (C)), nortriptyline ( $K_i=214$  nM, Fig. 5.8 (D)) and raloxifene ( $K_i=750$  nM, Fig. 5.8 (E)). Of these, raloxifene was the most surprising.

**Figure 5.8.** Competition binding isotherms at 5-HT<sub>6</sub>R for several predicted actives: (A) clomipramine (2, red triangle) and chlorpromazine (square), and doxepin (4, blue triangle) and chlorpromazine (square); (B) nortipytyline (8, red triangle) and chlorpromazine (square), and raloxifene (10, blue triangle) and chlorpromazine (square); (C) zuclopenthixol (13, triangle) and chlorpromazine (square), versus [3H]LSD.





5-HT<sub>6</sub> Receptor  
[<sup>3</sup>H]LSD (1.0 nM)  
Standard Binding Buffer



Among the tested compounds, we found that compounds having negative connectivity scores and *k*NN CS above 0.90 were all true actives at 5-HT<sub>6</sub>R achieving a success rate of 100%. We also found that lowering the threshold to 0.50 resulted in 3 false positives which decreased the success rate down to 77 %. It was strikingly important that we were able to prioritize a VS hit (i.e., raloxifene) with very low *k*NN CS of 0.56 and insignificant negative connectivity scores with Alzheimer's (see Table 5.5) and validate that this compound was a true binder of 5-HT<sub>6</sub>R and a potential therapeutic for Alzheimer's disease. This is a clear example on the importance of integrating and fusing independent hypotheses to increase the confidence of otherwise 'desperate' computational hits.

Mining of the biomedical literature using ChemoText identified possible neuroprotective, in addition to cognitive- and memory-enhancing effects for most of the computational hits (see Table 5.6), although there is no evidence that 5-HT<sub>6</sub>R –active compounds are neuroprotective. The list of all 39 compounds predicted by our integrative approach as putative 5-HT<sub>6</sub>R binders with possible anti-Alzheimer's effects is shown in Table 5. In addition, the top 100 VS hits (i.e., putative 5-HT<sub>6</sub>R ligands) from the WDI identified by the QSAR/VS approach are provided in the Supporting Information.

**Table 5.6.** The significance of the tested hits in relation to cognition, neuroprotection and anti-Alzheimer's effects.

Compound	Predicted	$K_i$	Significance to Alzheimer's disease prevention/treatment
1	Active	1956.0	Unknown
2	Active	112.0	Neuroprotective (Hwang et al. 2008)
3	Active	17.0	Used in combination therapy for Alzheimer's (PFEIFER et al. 2009)
4	Active	105.0	Unknown
5	Active	NB	GABA receptor modulator (Ong & Kerr 2005, Ong et al. 2005) and may inhibit amyloid-beta protein oligomerization as other related antihypertensives (Zhao et al. 2009)
6	Active	1188.0	Possible anti-Alzheimer's effects (Zhang et al. 2007)
7	Active	NB	Can inhibit central sensitization and neuroinflammation (Horwood et al. 2006, Pezet et al. 2008)
8	Active	214.0	Possible anti-Alzheimer's effects (Doraiswamy et al. 2003)
9	Active	NB	Unknown
10	Active	750.0	Possible anti-Alzheimer's effects (Yaffe et al. 2005)
11	Active	1041.0	Neuroprotective (O'Neill et al. 2004)
12	Active	4125.0	Unknown
13	Active	169.0	Facilitates memory in rats (Khalifa 2003)

### **SERMs Identified as 5-HT<sub>6</sub>R Ligands**

Several selective estrogen receptor modulators (SERMs) were predicted as 5-HT<sub>6</sub>R ligands and also had negative connections with the Alzheimer's disease gene signatures. Clomifene, raloxifene and tamoxifen had negative connections with Alzheimer's disease gene signatures in the cmap database (Lamb et al. 2006). Toremifene was not included in the cmap but was predicted as 5-HT<sub>6</sub>R binder by QSAR-based VS. Although anti-Alzheimer's effects of these drugs were observed previously and attributed to their modulation of estrogen receptors (ERs), the evidence about ER modulators or hormone replacement therapy in postmenopausal women to prevent or treat the Alzheimer's disease has been inconclusive and sometimes even contradictory (Asthana et al. 2009, Henderson 2009, Shumaker et al. 2003). Although postmenopausal estrogen depletion is a known risk factor for Alzheimer's disease, estrogen-containing hormone therapy initiated during late postmenopausal period does not improve episodic memory (an important early symptom of Alzheimer's disease), leads to no improvement or adverse effect on overall cognitive performance and Alzheimer's disease in postmenopausal women (Pinkerton & Henderson 2005, Rapp et al. 2003, Shumaker et al. 2003), and it increases the risk of dementia (Henderson 2009, Shumaker et al. 2003). Be that as it may, there is still substantial evidence from both pre-clinical and human studies that ovarian steroids have significant effects on neuroregulatory pathways (Schmidt & Rubinow 2009, Benmansour et al. 2009, Frye 2009, Ledoux et al. 2009, Woolley 2007b, Hart et al. 2007, Woolley 2007a, Woolley & Schwartzkroin 1998). However, critical gaps exist in our knowledge of both the effects on brain function of declining ovarian steroid secretion during reproductive aging, and the role of ovarian steroid hormone therapy in the prevention or treatment of brain diseases (Asthana et al. 2009).

## **Raloxifene Identified as a 5-HT<sub>6</sub>R Ligand and Agent with Potential Utility in Alzheimer's Disease**

Raloxifene is a selective estrogen receptor modulator used to prevent or treat osteoporosis; recently it was also approved by the FDA as an anti-cancer drug for reducing the risk of invasive breast cancer in postmenopausal women (FDA 2007). It was one of the low confidence QSAR-based VS hits because of the low structural similarity with modeling set compounds. Therefore, we would have avoided testing this compound if it had not been predicted from the cmap to have a decent negative connection with Alzheimer's disease. Another level of confidence was obtained from having other compounds that belonged to the same pharmacological group (SERMs) which were predicted as 5-HT<sub>6</sub>R actives with high confidence (i.e., consensus scores above 0.90) and had negative connections with Alzheimer's disease. This example highlights the value of the integrated informatics approach in increasing the hit rates of QSAR-based VS. Experimental testing had indeed confirmed that raloxifene binds to 5-HT<sub>6</sub>R with a  $K_i$  of 750 nM (Table 5.5, Fig. 5.8 (E)).

Yaffe and coworkers examined the data from the Multiple Outcomes of Raloxifene Evaluation (MORE) trial and indicated that raloxifene given at a dose of 120 mg/day, but not 60 mg/day, led to reduced risk of cognitive impairment in postmenopausal women. The maximum plasma concentration ( $C_{max}$ ) at such high doses indicated that administration of a single dose of 185 mg of raloxifene hydrochloride to four healthy volunteers resulted in a maximum plasma concentration ( $C_{max}$ ) of 12.5  $\mu$ g/L (~26 nM) (Morello et al. 2003).

We hypothesize that our studies identified raloxifene's putative cognition enhancing effects by using QSAR-assisted analysis of cmap connectivity scores. In the same time cmap analysis helped in prioritizing raloxifene, despite its very low  $kNN$  CS, as 5-HT<sub>6</sub>R binder

which turned out to have a decent binding to a very important potential target for cognition enhancement. Although raloxifene was not considered before as an attractive CNS drug due to its pharmacodynamic profile, the current findings of the MORE study, and others (Littleton-Kearney et al. 2002) pointed out that raloxifene enters the brain in relevant quantities and exerts a measurable effects in humans.

However, it is very possible that raloxifene's anticipated anti-Alzheimer's effects could be due to complex polypharmacological profile effecting several protein targets and signaling pathways involved in memory, cognition, inflammation, oxidative control and other important biological processes to Alzheimer's disease etiology, and not limited to its canonical targets (i.e. estrogen receptors). Future animal studies are required to validate the mechanism(s) of action for raloxifene's anti-Alzheimer's effects. Currently, raloxifene is in phase II clinical trials for Alzheimer's disease (NIA.NIH 2010).

### **Predict and Validate Polypharmacology of SERMs**

Our receptor family models suggested polypharmacological effects for SERMs through their predicted activities against 5-HT, adrenergic alpha, dopaminergic, histamine, muscarinic, and Sigma receptors (see Table 5.7). However, a closer look at the chemical structures of these compounds revealed some level of chemical dissimilarity between raloxifene and the three other SERMs studied here (i.e., clomiphene ,tamoxifen and toremifene).

**Table 5.7.** Predicting polypharmacology of SERMs using receptor family-based QSAR models described in Chapter 3.

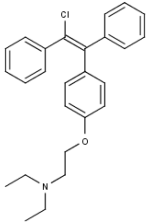
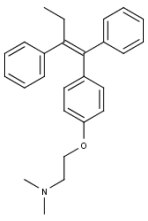
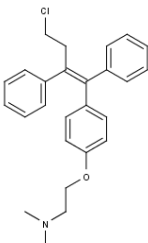
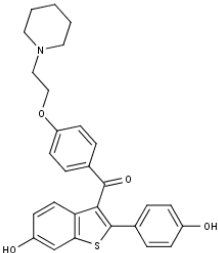
Compound	Serotonin		Alpha		Dopamine		Muscarinic		Histamine		Sigma	
	Pred <sup>a</sup>	Exp <sup>b</sup>	Pred	Exp	Pred	Exp	Pred	Exp	Pred	Exp	Pred	Exp
Clomiphene	B <sup>c</sup>	B	B	B	B	B	B	B	B	B	B	B
Raloxifene	B	B	B	B	B	B	B	B	B	B	B	B
Tamoxifen	B	B	B	B	B	B	B	NB <sup>d</sup>	B	B	B	B
Toremifene	B	B	B	B	B	B	B	NB	B	B	B	B
Success Rate	100%		100%		100%		50%		100%		100%	

<sup>a</sup>Pred, predicted binding using QSAR models; <sup>b</sup>Exp, experimental result using secondary radioligand binding assays; <sup>c</sup>B, binder; <sup>d</sup>NB, non-binder.

Structural similarity evaluation based on MACCS structural keys and Tanimoto coefficients indicated that raloxifene's similarity to the other SERMs (clomiphene, tamoxifen and toremifene) is  $< 30\%$ . This suggests this molecule might have some distinct pharmacological profile and might interact with different molecular targets in a distinct manner (at least with some of these molecular targets). However, the structural similarity between the three other SERMs studied here is  $> 78\%$  (see Table 5.8). Additionally, analysis of SERMs-protein interaction networks using STITCH (Kuhn et al. 2008) indicated that raloxifene has a different set of nearest neighbor proteins than the other SERMs (see Fig. 5.9). This chemical and biological dissimilarity suggests that raloxifene might have a different polypharmacological profile that's not limited to its action on estrogen receptors.

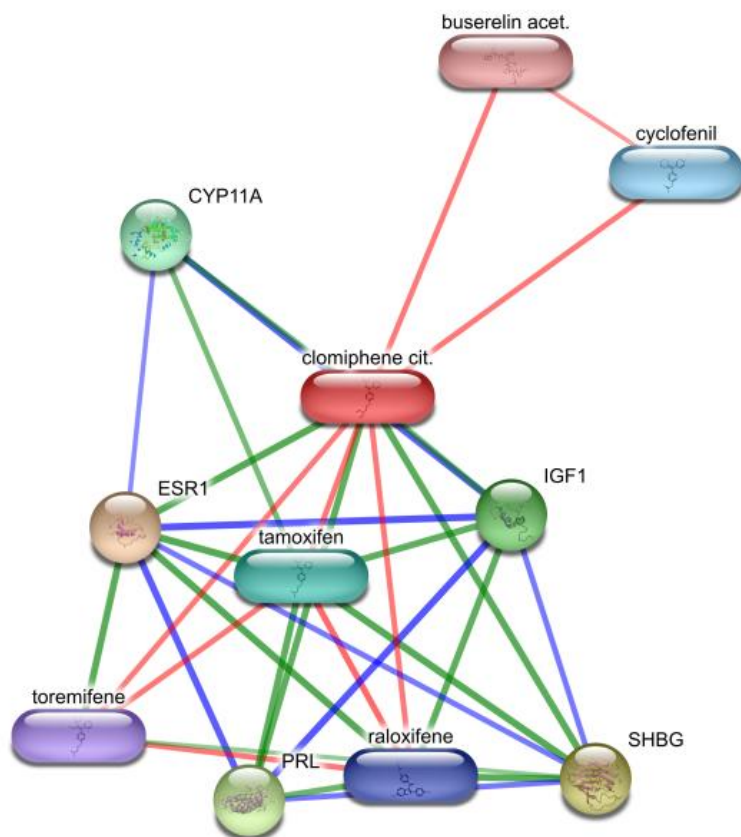


**Table 5.8.** Tanimoto similarities between SERMs based on MACCS structural keys.

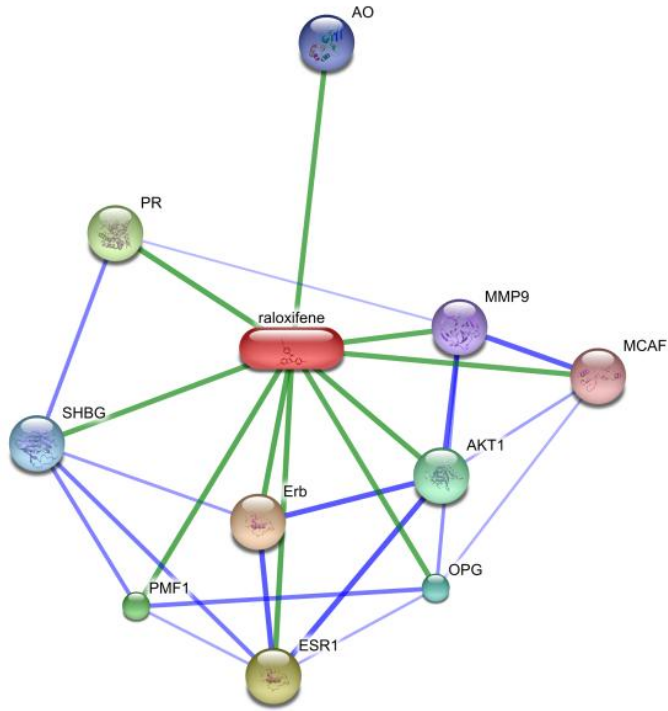
SERMs	Clomiphene	Tamoxifen	Toremifene	Raloxifene
 <b>Clomiphene</b>	100.0	78.6	81.0	29.4
 <b>Tamoxifen</b>	78.6	100.0	87.5	27.9
 <b>Toremifene</b>	81.0	87.5	100.0	27.5
 <b>Raloxifene</b>	29.4	27.9	27.5	100.0

**Figure 5.9.** Chemical protein interaction networks for SERMs. (A) Network centered at clomiphene, (B) Network centered at raloxifene, (C) Network centered at tamoxifen, and (D) Network centered at toremifene.

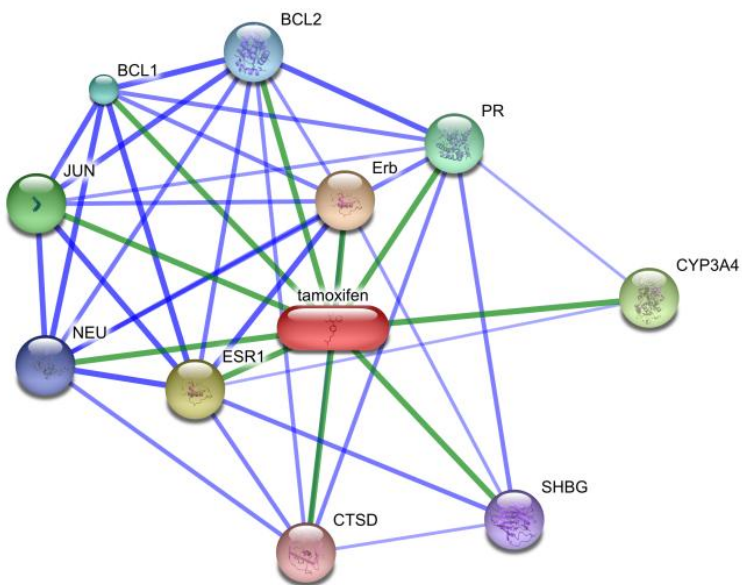
A



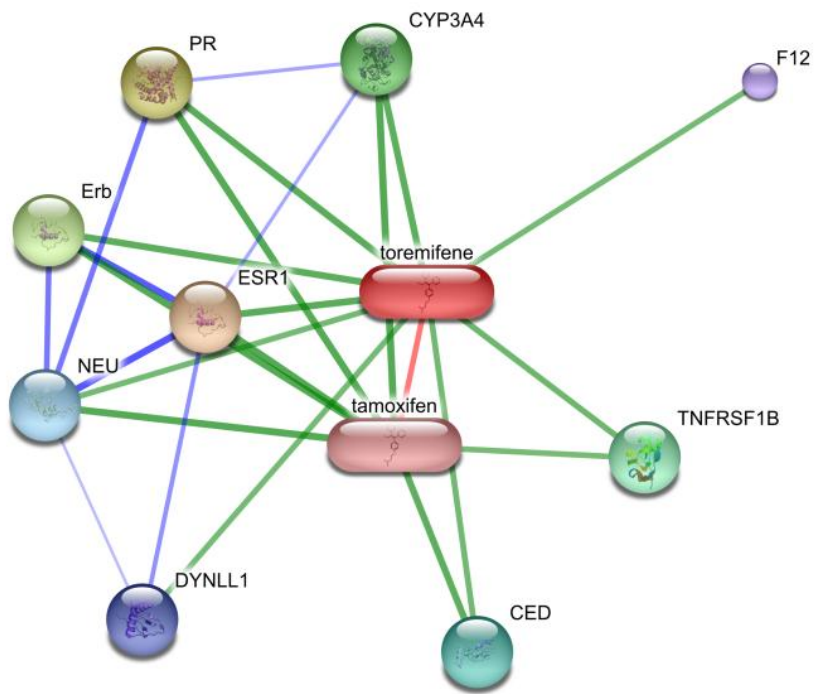
**B**



**C**



D



Additionally, we undertook a large screen of potential targets using the receptorome profiling approach (Armbruster and Roth, 2005). All tests were performed by our collaborators at PDSP. Secondary screening results indicated that raloxifene has nanomolar binding affinities to several GPCRs, ion channels and protein transporters. The highest binding affinities were towards adrenergic  $\alpha_{2C}$  receptors with a  $K_i$  of 61 nM, 5-HT<sub>2B</sub> receptors with  $K_i$  of 69 nM, kappa opioid receptors (KOR) with  $K_i$  of 186, and sigma 1 receptors with  $K_i$  of 247 nM. All binding affinity results are shown in Table 5.9.

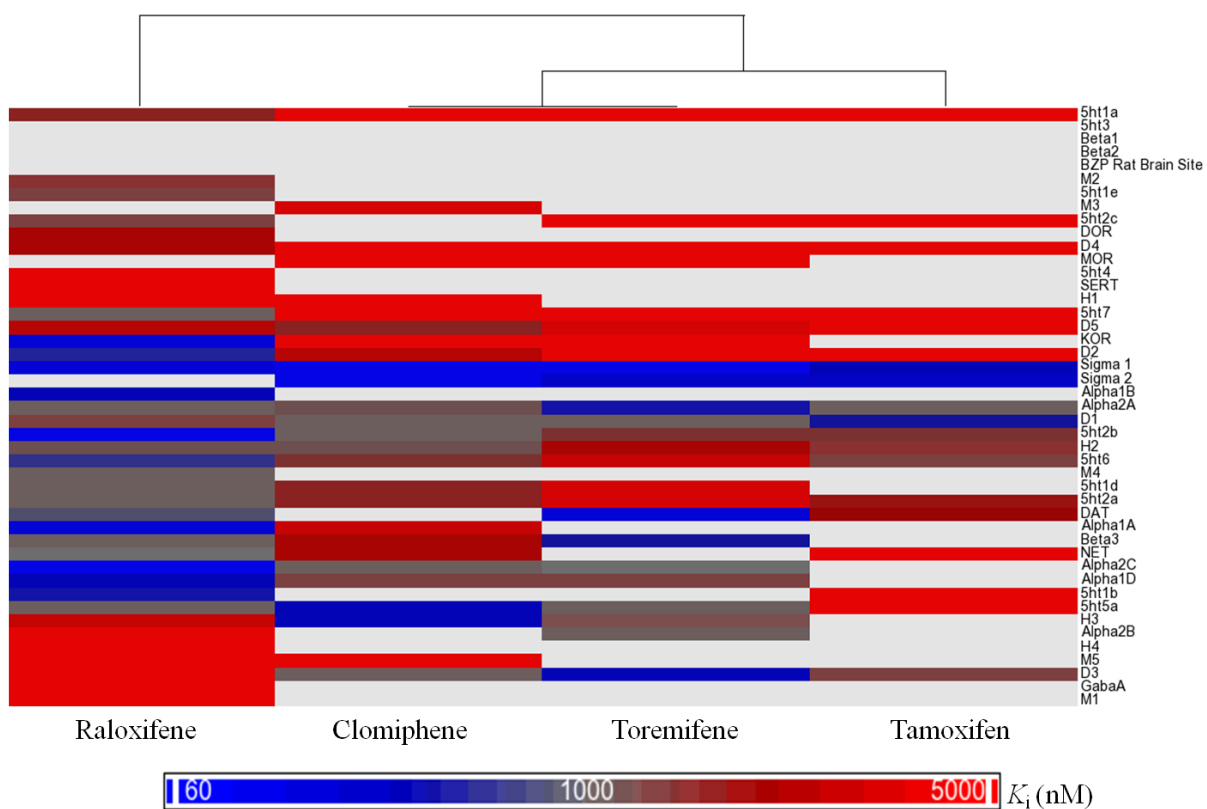
Clustering binding affinities of SERMs across a panel of molecular targets confirmed both QSAR and STITCH predictions that raloxifene's biological profile might more different than that for other SERMs. In figure 5.9 we can see that centering the STITCH network on clomiphene, will pull the rest of the three other SERMs but tamoxifen is closer to clomiphene than to raloxifene. Additionally, in figure 5.9 D, we can see that centering the network on toremifene identified tamoxifen as the nearest neighbor SERM. Binding profiles of SERMs to a panel of GPCRs and other related molecular targets (cf. Fig. 5.10) confirmed that clomiphene, tamoxifen and toremifene indeed cluster together, while raloxifene's binding affinity profile seemed a bit more different. From this example we can foresee the great potential for combining the knowledge derived from both genomic and protein binding profiles with the knowledge derived from QSAR-based VS to identify unexpected connections between molecular targets and chemical compounds.

**Table 5.9.**  $K_i$  estimates for SERMs (i.e., clomiphene, raloxifene, tamoxifen and toremifene) at a large panel of cloned receptors.

Receptor	$K_i$ for Clomiphene (PDSP 13499)	$K_i$ for Raloxifene (PDSP 13505)	$K_i$ for Tamoxifen (PDSP 678)	$K_i$ for Tamoxifen (PDSP 10572)	$K_i$ for Toremifene (PDSP 16514)
5ht1a	>10000	2330	3477	>10000	>10000
5ht1b	Primary < 50%	624	1618	7857	Primary < 50%
5ht1d	2171	1222	N/A	Primary < 50%	4431
5ht1e	Primary < 50%	1868	Primary < 50%	Primary < 50%	Primary < 50%
5ht2a	2281	1049	2596	2720	4520
5ht2b	1210	69	N/A	1952	1916
5ht2c	Primary < 50%	1642	4282	5787	>10000
5ht4	N/A	5050	N/A	N/A	N/A
5ht5a	506	1219	2123	7821	1283
5ht6	1956	750	931.1	1698	4125
5ht7	6615	1220	1077	>10000	5428
Alpha1A	4085	247.7	N/A	N/A	Primary < 50%
Alpha1B	Primary < 50%	534.6	N/A	N/A	Primary < 50%
Alpha1D	1625	478.2	N/A	Primary < 50%	1633
Alpha2A	1440	1288.2	N/A	1211	622.8
Alpha2B	Primary < 50%	7556	N/A	N/A	1105
Alpha2C	1255	61	N/A	Primary < 50%	1004
D1	1252	1626	1508	657	1138
D2	3543.0(AVE)	683	1682	5517	5122
D3	1302	>10000	498	1740	514
D4	6191 (AVE)	3023	7817	>10000	6209
D5	2298	3803	>10000	>10000	4550
DAT	N/A	928	4328	2820	263
H1	5853	5356	N/A	Primary < 50%	Primary < 50%
H2	1403	1436	N/A	1980	3067
H3	550.1	3847	N/A	N/A	1520
H4	Primary < 50%	7072	N/A	Primary < 50%	Primary < 50%
M1	Primary < 50%	>10,000	N/A	Primary < 50%	Primary < 50%
M2	Primary < 50%	2037	N/A	Primary < 50%	Primary < 50%
M3	4476	Primary < 50%	N/A	Primary < 50%	Primary <

					50%
M4	N/A	1229	N/A	Primary < 50%	Primary < 50%
M5	6816	8127	N/A	Primary < 50%	Primary < 50%
Sigma 1	128	247.8	N/A	481	183
Sigma 2	19	Primary < 50%	N/A	331	377

**Figure 5.10.** The heatmap of binding affinities ( $K_i$ ) for several SERMs (clomiphene, raloxifene, tamoxifen and toremifene), across a panel of GPCRs and other transmembrane molecular targets, analyzed by hierarchical clustering of the pairwise similarities in binding affinities using Euclidean distances. The bar-view is a key for coloring according to normalized descriptor frequency based on binding affinities where blue color indicates most potent binding affinities while red color denotes least potent binding affinities.





## Conclusions

We have developed a novel integrative chemocentric informatics approach that could be used as a tool for generating and cross-validating drug discovery hypotheses. Our approach integrates different *in silico* strategies and different data types and sources to increase the confidence in the final hypotheses. The study design was composed of three major parts: (1) QSAR-based datamining of chemical libraries to identify new ligands for target proteins, (2) Network-mining to identify chemicals that could treat specific diseases; and (3) Hypothesis fusion between (1) and (3).

This approach has been applied to study the 5-HT<sub>6</sub>R system in relation to cognition enhancement strategies which may be useful for Alzheimer's and similar diseases with impaired cognition (e.g., schizophrenia). Disease gene signatures for Alzheimer's disease have been used to query the cmap database to formulate testable hypotheses about potential treatments. Common compound hits from QSAR/VS studies against 5-HT<sub>6</sub>R and the cmap were tested in at 5-HT<sub>6</sub>R. Our approach identified 39 drugs, as potential 5-HT<sub>6</sub>R antagonists, out of 59000 molecules included in the WDI. Thirteen hits with higher confidence level were tested in binding assays and ten compounds were confirmed as 5-HT<sub>6</sub>R ligands achieving a success rate of 77%. We noticed that this study design can be applied to many other protein targets and families of targets involved in the etiology of Alzheimer's disease.

Herein, we hypothesized and proved that integrating results generated from the cmap with predictions generated from QSAR-based VS increased the confidence level in the computational hits generated from QSAR-based VS. It also increased our confidence in some cmap negative connections with Alzheimer's (i.e., raloxifene) that would have been neglected based on either their cmap weak scores or lower confidence predictions from

QSAR models. Therefore, we foresee this design as a promising tool to identify molecule-molecular target-disease (phenotype) associations.

## CHAPTER 6

### FUTURE DIRECTIONS

#### Summary

In order to pursue a rational approach for the discovery of ‘magic shotguns’ and to end our reliance on serendipity as the major driving force for discovering such compounds, both genomics-based physical screening and *in silico* receptoromics should come together in the interplay. Herein, we attempted for the first time to establish a compendium of computational predictors, completely based on 2D QSAR methods, could be used simultaneously for the identification of potential leads. Once leads are discovered, potential toxicities could be also predicted in a similar manner by virtually screening such compounds against anti-targets (e.g., 5-HT<sub>2B</sub> agonists) (Setola & Roth 2005). Later, after confirming activities against the predicted molecular targets, structure activity effects can be tweaked by medicinal chemists.

In this study, we succeeded in accumulating a large number of computational predictors for several receptor families (5-HT, adrenergic, dopamine, histamine, muscarinic and sigma receptors) and subtypes (5-HT<sub>2A</sub>, 5-HT<sub>2B</sub>, 5-HT<sub>6</sub>, 5-HT<sub>7</sub>, adrenergic Alpha<sub>2A</sub>, adrenergic Alpha<sub>2B</sub>, adrenergic Alpha<sub>2C</sub>, D<sub>1</sub>, D<sub>2</sub>, D<sub>3</sub>, D<sub>4</sub>, and D<sub>5</sub>). Our classification models had high accuracies (CCR<sub>evs</sub> in the range 0.74 to 0.97) estimated on

external validation sets. Continuous models using actual binding affinity data had lower accuracies however, with  $R^2$  ranging from 0.59-0.67. This might be partially due to some inconsistencies in binding data extracted from different sources and generated by different groups. It might be also indicative that binding data per se might not be as effective as activity data to generate highly predictive structure activity models. In other words, our experience with testing actives indicated that many compounds that have very good nanomolar binding affinities are totally inactive in functional assays. Therefore, using high-quality functional data instead of binding data might be more appropriate to achieve our goals in predicting highly active compounds across several receptors of interest. Consequently, this might lead to a new era of successful QSAR modeling of polypharmacological effects. So far, we are still very short in high-quality activity data deposited in public repositories.

One of the most important predictors generated in our studies was based on QSAR models of 5-HT<sub>2B</sub> receptor ligands that can be used for virtual screening to identify potential valvulopathic compounds. Our results indicated the reliability of our computational models as efficient predictors of compounds' affinity towards 5-HT<sub>2B</sub> receptors. We suggest that the computational models developed in this study could be used as drug liability predictors similar to commonly used predictors of other undesired side effects such as carcinogenicity, mutagenicity, PGP binding, or hERG binding. Our models can be used to flag compounds that are expected to bind to 5-HT<sub>2B</sub> receptors but they cannot distinguish agonists from antagonists. Nevertheless, as demonstrated in this study, these putative 5-HT<sub>2B</sub> binders can be tested in functional assays for their potential to activate 5-HT<sub>2B</sub> receptors to further assess their valvulopathic potential.

We have also developed a novel integrative chemocentric informatics approach that could be used as a tool for generating and cross-validating drug discovery hypotheses. Our approach integrates different *in silico* strategies and different data types and sources to increase the confidence in the final hypotheses. The study design was composed of three major parts: (1) QSAR-based datamining of chemical libraries to identify new ligands for target proteins, (2) Network-mining to identify chemicals that could treat specific diseases; and (3) Hypothesis fusion between (1) and (3).

This approach has been applied to study the 5-HT<sub>6</sub>R system in relation to cognition enhancement strategies which may be useful for Alzheimer's and similar diseases with impaired cognition (e.g., schizophrenia). Herein, we hypothesized and proved that integrating results generated from the cmap with predictions generated from QSAR-based VS increased the confidence level in the computational hits generated from QSAR-based VS. It also increased our confidence in some cmap negative connections with Alzheimer's (i.e., raloxifene) that would have been neglected based on either their cmap weak scores or lower confidence predictions from QSAR models. Therefore, we foresee this design as a promising tool to identify molecule-molecular target-disease (phenotype) associations. We also believe that our recent and future studies into integrated chemocentric informatics will provide new successful avenues for the development of novel multi-targeted therapeutics capable of treating and/or preventing complex diseases such as neurodegenerative diseases, cancer and diabetes. This approach could be extended to many similar receptor systems and many different diseases serving as a cost-effective *in silico* tool for the discovery of novel biologically active compounds acting via clinically relevant targets.

### ***In silico* Receptoromics**

All methods discussed above share potential pitfalls of any QSAR methodology such as over-fitting and potential danger associated with extrapolation. However, with the continuing advancements in this field, on the level of machine learning methods and descriptor types, there is a great room for improving the expected outcome. Future work in this field will include experimenting with various types of descriptors as well as consider additional combinations of classification techniques in the context of combinatorial QSAR modeling.

Experimental validation procedures should be designed carefully to study the possibilities and the limitations of family-based models, especially because we are dealing with closely related proteins (i.e., GPCRs). For example, common hits predicted to bind several families of GPCRs should be experimentally validated. We should be cautioned however, that even though the binding profile might look promising but the real therapeutic effects are related to the actual functional activities of these ligands on specific receptor subtypes within these families. At the same time, the higher is the predicted promiscuity level, the higher is the risk for having adverse effects *in vivo* and the higher possibility for complications in drug design efforts in general; trying to optimize a lead's activity on several molecular targets at the same time would be highly difficult.

Additionally, to be able to make the best use of family-based models we should design the proper tools for: (1) calculation of the appropriate consensus scores across family-based models, (2) estimating the hit rates in actual binding assays. Currently, we are calculating consensus scores separately from models generated for each family. Then we select for further experimental testing those hits that possess acceptable consensus scores in each case. It might be useful to find new ways to calculate one consensus score across all

families which will make it easier for us to perform activity profiling across families and to calculate corresponding hit rates. However, calculating hit rates for hits predicted to bind several families of GPCRs could be highly complicated. It should be kept in mind that although we predict a compound to be a ligand for a specific receptor family, but this prediction does not imply in any way that this compound will bind to several or all receptor subtypes in that specific family. This result is dictated by the nature of the datasets used for model building. We will find that some chemicals will bind one specific subtype and does not bind to any other subtypes in the same family and vice versa.

Additionally, as more data becomes available where we have full matrices for large number of compounds tested in binding and functional assays against a panel of receptors, it will be highly useful to experiment with Multi-Task learning (MTL) methods; unlike the case with conventional QSAR calculations using Single Task Learning (STL), where the models are developed for a single property, Multi-task Learning (MTL) approaches train the models simultaneously for several related properties (i.e., binding profiles against several receptor families or subtypes). Such approach has the potential to predict polypharmacology against a multitude of receptor families and subtypes and/or subtype selectivity. MTL is expected to increase the predictive power of the QSAR models in comparison to STL models (Varnek et al. 2009), because “what is learned for each task can help the other tasks learned better” (Caruana 1997).

GPCRs are highly promiscuous targets, and the majority of GPCR ligands hit multiple GPCRs at the same time, therefore, it makes more sense to use modeling methods that could analyze the ligand binding profile to families and subtypes of GPCRs at once. This process will allow the model to take into account all the commonalities and differences that

exist between families of receptors and specific receptor subtypes. We foresee MTL as a promising approach to study both polypharmacology and selectivity of GPCR ligands; single task models are surely missing an important layer on information embedded in the ligands affinity to all other families of receptors or receptor subtypes. In MTL, the modeled property consists of all the binding affinities of compounds across different families of receptors. Such property can be coded in a bit string of zeros and ones (i.e., binding is encoded by 1 and non-binding is encoded by 0).

One can use Associative Neural Networks (ASNN) (ASNN 2010, Tetko 2002) as a machine learning method to conduct MTL on polypharmacological datasets. This method represents a combination of an ensemble of feed-forward neural networks and the k-nearest neighbor technique. ASNN uses the correlation between ensemble binding responses as a measure of distance among the analyzed cases for the nearest neighbor technique. Using this method provides an improved prediction by the bias correction of the neural network ensemble. An associative neural network has a memory that can coincide with the training set. If new data becomes available, the network further improves its predictive ability and provides a reasonable approximation of the unknown function without a need to retrain the neural network ensemble. This feature of the method dramatically improves its predictive ability over traditional neural networks and k-nearest neighbor techniques. Another important feature of ASNN is the possibility to interpret neural network results by analysis of correlations between data cases in the space of models.

### **Chemocentric Informatics Approach**

One of the major limitations of using our integrative approach effectively is the relatively small number of compounds contained in the cmap. Consequently, this will limit



our ability to identify novel hits; novel hits are better identified from larger libraries with a multitude of diverse compounds. But this resource is constantly growing, and it aims to cover all FDA approved drugs and other chemicals of clinical potential. However, we might be able to tackle this problem computationally; where the negative connections (with a disease state) having the highest confidence can be used as similarity probes to search chemical databases.

There are also limitations that originate from the disease gene-signatures. It is common that patients having certain diseases or who are older than 50, are already taking some medications for different reasons. These medications might affect the quality of generated disease gene signatures so that changes in gene signatures might be more correlated to drugs rather than disease states (i.e. it might mask disease signatures in postmortem signatures from human brain) and consequently could affect the result we obtain from querying the cmap. However, the amount of information contained in these genomic signatures is enormous. Therefore, the earnest analysis of all predicted negative and positive connections with a disease state and considering the wide range of connectivity scores (i.e. negative, positive and null) should be considered and studied carefully. Experiments should be also designed to test both positive and negative hypotheses.

Another kind of analysis for cmap predictions depends on the specific scores for gene signature similarities (with the disease gene signature) in each of the cell lines used to generate the chemical's gene expression profiles. Herein, we raise the question whether specific cell types are more reliable than others in a disease state of interest? Do we have to analyze scores obtained from each cell line separately to derive hypotheses based on the differences between cell lines and what protein targets are over-expressed in which cell lines? Or shall we consider inconsistencies in connectivity scores among cell lines as an

evidence for non-reliable gene expression profiles and therefore should be neglected? So these are all important issues to keep in mind for further improvements on the way we analyze data from this important chemogenomic resource. It would be wise though to consider all possibilities, derive hypotheses, design appropriate tests and validate predictions and methods. Finally, fusing all generated hypotheses is expected to provide us with the most optimal solutions.

### **Conclusion**

In conclusion, it is likely that our computational efforts described herein and other efforts by different groups to study polypharmacology *in silico* will eventually result in useful and reliable tools aimed at enriching chemical libraries in compounds that have affinities for more than a single desired molecular target. We also think that a combination of *in silico* methods will be more powerful than a single method alone as our expertise in the computational field has indicated time after time.

## Appendix I:

Final VS Hits as 5-HT<sub>2B</sub> Actives

WDI Compound Name	No. of <i>k</i> NN	Consensus
	Models	Score
VEIUTAMINE	104	1.00
DROMIA	112	1.00
BRL-56905	121	1.00
CLAUSINE-E	123	0.99
ARAPROFEN	112	0.99
FURPROFEN	104	0.99
ADRENOGLOMERULOTROPIN	122	0.98
6-FLUOROMELATONIN	115	0.98
HYDROXYTRYPTAMINE- GLUTAMYL	111	0.98
LIQUIRITIGENIN	110	0.98
PIDOBENZONE	123	0.98
GALANGIN	103	0.97
EMD-47020	102	0.97
SYRIACUSIN-C	119	0.97
MICROMINUTIN	117	0.97
NPC-14692	113	0.96
SALSALIC-ACID	124	0.96
CP-123800	116	0.96

METHYLDOPA-RACEMIC	115	0.96
SR-4452	114	0.96
CEAROIN	111	0.95
CL-08-A	109	0.95
MNA-279	102	0.95
2-HYDROXYESTRONE	102	0.95
5-METHOXYTRYPTOLINE	115	0.95
EMD-57283	111	0.95
SALICYL-TYROSINE	110	0.95
HYDROXYMESOCARB-BETA	108	0.94
RHAPONTIGENIN	114	0.94
AFFININE	108	0.94
6-METHOXYMELLEIN	122	0.93
R-53309	105	0.93
BARCELONEIC-ACID-B	103	0.93
17-ALPHA-ESTRADIOL-ACETATE	116	0.93
RH-34	112	0.93
LEK-8827	109	0.93
PAXAMATE	122	0.93
ACETYLCARANINE	106	0.92
METHYLERGOMETRINE	118	0.92
FEBRIFUGINE	117	0.92
K-182	116	0.92

2-METHOXYESTRONE	113	0.92
METHYLDOPA	107	0.92
CGP-13698	114	0.91
BW-826-C	101	0.91
BE-6143	110	0.91
AM-40	112	0.90
BUTOLAME	110	0.90
L-TYROSINE	104	0.89
LAPACHONE-BETA	107	0.89
TOFETRIDINE	124	0.89
FUSAROCHROMANONE	103	0.88
DOISYNESTROL	119	0.88
METBUFEN	108	0.88
H-195-60	122	0.88
AMINOHIPPURATE-SODIUM	102	0.87
SR-4895	100	0.87
XANTHANOIC-ACID	123	0.87
HEXACYPRONE	115	0.87
BENZALBUTYRATE-SODIUM	115	0.87
LY-248510	121	0.87
BENZOYLNORECGONINE	110	0.86
PYRIDAZOMYCIN	102	0.86
VALLDEMOSSINE	115	0.86

CARFIMATE	107	0.86
TL-404	121	0.86
RO-03-9024	115	0.85
4-BENZYLOXYPHENYLACETATE	121	0.85
NB-355	119	0.85
MENADIONE	102	0.84
PIM-35	101	0.84
TRANS-2-HYDROXYLOMUSTINE	106	0.84
FLUTIMIDE	111	0.84
661-U-88	117	0.84
SENKYUNOLIDE-J	104	0.84
ERYTHRININ-B	103	0.83
LY-193326	120	0.83
YM-992	120	0.83
CARBESTROL	123	0.83
RETICULATINE-A	103	0.83
CLAVIROLIDE-D	102	0.82
3-HYDROXYPRAZEPAM	110	0.82
NSC-350102	114	0.82
POLYMONINE	114	0.82
C-883901	102	0.81
PISIFERIC-ACID	107	0.81
IMIDOCARB	100	0.81

NPC-15199	115	0.81
ISO-BUTYLNAPHTHYLACETATE	121	0.80
PROPIONYLPHENETIDINE	123	0.80
RALGIN	121	0.79
NPC-15667	119	0.79
HYDROXYPHENYLGLYCINE	104	0.79
BUTYLPHENAMIDE	112	0.79
ARPHAMENINE-B	111	0.78
METAMIVANE	123	0.78
RICCARDIPHENOL-C	120	0.77
WAY-122331	120	0.77
TROXIPIDE	114	0.77
BW-306-U	117	0.77
MELINONINE-F	105	0.76
GLYCOCTRINE-II	109	0.76
AMICLENOMYCIN	104	0.76
METHYLSALICYLATE	103	0.75
L-4035	118	0.75
BUTACETIN	113	0.74
O-ACETYLPROPRANOLOL	116	0.74
L-372460	114	0.74
ETERSALATE	113	0.73
FUCHSIN	100	0.73

4-HYDROXYDERRICIN	111	0.73
FPL-13950	124	0.73
KWD-2131	116	0.72
K-7731	118	0.72
CD-417	117	0.72
MC-207110	106	0.72
ENALAPRIL	109	0.72
PHENOXYACETATE-METHYL-ESTER	107	0.71
BENZASTATIN-B	114	0.70
NSC-319848	107	0.70
NSC-645306	100	0.70
ASPERGILLAMIDE-A	100	0.70



## Appendix II

146 5-HT<sub>2B</sub> Actives Used in the Modeling Studies

<b>Cp.</b>	<b>PubChe</b>	<b>SMILES</b>
<b>ID</b>	<b>CID</b>	
1	1001	<chem>NCCc1ccccc1</chem>
2	1065	<chem>O(C)c1cc2c(nccc2C(O)C2N3CC(C(C2)CC3)C=C)cc1</chem>
3	1150	<chem>[nH]1cc(c2c1cccc2)CCN</chem>
4	1224	<chem>OC1(CC2N(CC1)CC1c3c2cccc3CCc2c1cccc2)C(C)(C)C</chem>
5	1229	<chem>Ic1cc(OC)c(cc1OC)CC(N)C</chem>
6	1243	<chem>BrC1cc2c(cc1O)C(CN(CC2)C)c1ccccc1</chem>
7	1250	<chem>Oc1cc2c(N(C3N(CCC23C)C)C)cc1</chem>
8	1355	<chem>Clc1cc(N2CCNCC2)ccc1</chem>
9	1614	<chem>O1c2cc(ccc2OC1)CC(N)C</chem>
10	1615	<chem>O1c2cc(ccc2OC1)CC(NC)C</chem>
11	1832	<chem>O(C)c1cc2c([nH]cc2CCN(C)C)cc1</chem>
12	2099	<chem>O=C1N(CCCc2n(c3c(c12)cccc3)C)Cc1nc[nH]c1C</chem>
13	2159	<chem>S(=O)(=O)(CC)c1cc(C(=O)NCC2N(CCC2)CC)c(OC)cc1N</chem>
14	2160	<chem>N(CCC=C1c2c(CCC3c1cccc3)cccc2)(C)C</chem>
15	2170	<chem>Clc1cc2c(Oc3c(N=C2N2CCNCC2)cccc3)cc1</chem>
16	2196	<chem>O(C)c1ccc(cc1)C(=O)N1CCCC1=O</chem>
17	2247	<chem>Fe1ccc(cc1)Cn1c2c(nc1NC1CCN(CC1)CCc1ccc(OC)cc1)cccc2</chem>
18	2267	<chem>Clc1ccc(cc1)CC1=NN(C2CCCN(CC2)C)C(=O)c2c1cccc2</chem>
19	2308	<chem>ClC12C(C3CC(C)C(O)(C(=O)CO)C3(CC1O)C)CCC1=CC(=O)C=CC12C</chem>

20 2318 FC(F)(F)c1cc(ccc1)CC(NCCOC(=O)c1ccccc1)C  
 21 2326 O(CC(N1CCCCC1)C)c1ccccc1Cc1ccccc1  
 22 2344 O(C(c1ccccc1)c1ccccc1)C1CC2N(C(C1)CC2)C  
 23 2377 O(CCCNC)c1ccccc1Cc1ccccc1  
 24 2443 Brc1[nH]c2c3c1CC1N(CC(C=C1c3ccc2)C(=O)NC1(OC2(O)N(C(CC(C)C)C(=O)N3C2CCC3)C1=O)C(C)C)C  
 25 2477 O=C1N(CCCCN2CCN(CC2)c2ncccn2)C(=O)CC2(C1)CCCC2  
 26 2512 O=C(N(CCCN(C)C)C(=O)NCC)C1CC2C(N(C1)CC=C)Cc1c3c2cccc3[nH]c1  
 27 2520 O(C)c1cc(ccc1OC)C(C(C)C)(CCCN(CCc1cc(OC)c(OC)cc1)C)C#N  
 28 2585 O(CCNCC(O)COc1c2c3c([nH]c2ccc1)cccc3)c1ccccc1OC  
 29 2726 Clc1cc2N(c3c(Sc2cc1)cccc3)CCCN(C)C  
 30 2769 Clc1cc(C(=O)NC2CCN(CC2OC)CCCOc2ccc(F)cc2)c(OC)cc1N  
 31 2771 Fc1ccc(cc1)C1(OCc2c1ccc(c2)C#N)CCCN(C)C  
 32 2780 Clc1cc(C(=O)NC2CCN(CC2)Cc2ccccc2)c(OC)cc1N  
 33 2781 Clc1ccc(cc1)C(OCCC1N(CCC1)C)(C)c1ccccc1  
 34 2782 Clc1ccc(cc1)Cn1c2c(nc1CN1CCCC1)cccc2  
 35 2801 Clc1cc2N(c3c(Cc2cc1)cccc3)CCCN(C)C  
 36 2818 Clc1cc2NC(N3CCN(CC3)C)=C3C(=Nc2cc1)C=CC=C3  
 37 2820 Clc1cc2NC(N3CCNCC3)=C3C(=Nc2cc1)C=CC=C3  
 38 2866 OC1CCC2C(CC3N(C2)CCc2c3[nH]c3c2cccc3)C1C(OC)=O  
 39 2895 N(CCC=C1c2c(C=Cc3c1cccc3)cccc2)(C)C  
 40 2913 N1(CCC(CC1)=C1c2c(C=Cc3c1cccc3)cccc2)C

41 3065 O1C(NC(=O)C2CC3C(N(C2)C)Cc2c4c3cccc4[nH]c2)(C(C)C)C(=O)N2C(Cc3cccc3)C(=O)N3C(CCC3)C12O

42 3066 O1C(NC(=O)C2CC3C(N(C2)C)Cc2c4c3cccc4[nH]c2)(C)C(=O)N2C(Cc3cccc3)C(=O)N3C(CCC3)C12O

43 3117 S(SC(=S)N(CC)CC)C(=S)N(CC)CC

44 3250 OCC(NC(=O)C1C=C2C(N(C1)C)Cc1c3c2cccc3[nH]c1)C

45 3251 O1C(NC(=O)C2C=C3C(N(C2)C)Cc2c4c3cccc4[nH]c2)(C)C(=O)N2C(Cc3cccc3)C(=O)N3C(CCC3)C12O

46 3341 Clc1c2c(cc(O)c1O)C(CNCC2)c1ccc(O)cc1

47 3372 S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)CCO)cc(cc2)C(F)(F)F

48 3386 FC(F)(F)c1ccc(OC(CCNC)c2cccc2)cc1

49 3463 O(CCCC(C(O)=O)(C)C)c1cc(ccc1C)C

50 3519 Clc1cccc(Cl)c1CC(=O)N=C(N)N

51 3646 O1c2c(OC1)cc1c(CN(CC1)C)c2OC

52 3675 N(N)CCc1cccc1

53 3689 Oc1ccc(cc1)C(O)C(N1CCC(CC1)Cc1cccc1)C

54 3827 s1c2c(cc1)C(c1c(CC2=O)cccc1)=C1CCN(CC1)C

55 3878 Clc1c(cccc1Cl)-c1nnc(nc1N)N

56 3938 O=C(NC1C=C2C(N(C1)C)Cc1c3c2cccc3[nH]c1)N(CC)CC

57 3964 Clc1cc2c(Oc3c(N=C2N2CCN(CC2)C)cccc3)cc1

58 4090 O(Cc1cccc1)C(=O)NCC1CC2C(N(C1)C)Cc1c3c2cccc3n(c1)C

59 4106 S1c2c(cc(SC)cc2)C(N2CCN(CC2)C)Cc2c1cccc2

60 4140 OCC(NC(=O)C1C=C2C(N(C1)C)Cc1c3c2cccc3[nH]c1)CC

61	4184	<chem>N12C(c3c(Cc4c1cccc4)cccc3)CN(CC2)C</chem>
62	4205	<chem>n1c2N3C(c4c(Cc2ccc1)cccc4)CN(CC3)C</chem>
63	4296	<chem>FC(F)(F)c1cc(N2CCNCC2)ccc1</chem>
64	4418	<chem>O(C)c1cccc1N1CCN(CC1)CC(O)COc1c2c(ccc1)cccc2</chem>
65	4449	<chem>Clc1cc(N2CCN(CC2)CCCN2N=C(N(CCOc3cccc3)C2=O)CC)ccc1</chem>
66	4475	<chem>BrC1cc(cnc1)C(OCC1CC2(OC)C(N(C1)C)Cc1c3c2cccc3n(c1)C)=O</chem>
67	4543	<chem>N(CCC=C1c2c(CCc3c1cccc3)cccc2)C</chem>
68	4585	<chem>S1C2=Nc3c(NC(N4CCN(CC4)C)=C2C=C1C)cccc3</chem>
69	4595	<chem>O=C1c2c(n(c3c2cccc3)C)CCC1Cn1cnc1C</chem>
70	4636	<chem>Oc1c(C)c(CC=2NCCN=2)c(cc1C(C)(C)C)C</chem>
71	4658	<chem>Ic1ccc(cc1)C(=O)N(CCN1CCN(CC1)c1cccc1OC)c1cccc1</chem>
72	4691	<chem>Fc1ccc(cc1)C1CCNCC1COc1cc2OCOc2cc1</chem>
73	4745	<chem>S(CC1CC2C(N(C1)CCC)Cc1c3c2cccc3[nH]c1)C</chem>
74	4748	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCN1CCN(CC1)CCO</chem>
75	4847	<chem>Fc1ccc(cc1)C(=O)C1CCN(CC1)CCC=1C(=O)N2C(=NC=1C)C=CC=C2</chem>
76	4893	<chem>o1cccc1C(=O)N1CCN(CC1)c1nc(N)c2cc(OC)c(OC)cc2n1</chem>
77	4917	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCN1CCN(CC1)C</chem>
78	4927	<chem>S1c2c(N(c3c1cccc3)CC(N(C)C)C)cccc2</chem>
79	5002	<chem>S1c2c(cccc2)C(=Nc2c1cccc2)N1CCN(CC1)CCOCCO</chem>
80	5011	<chem>n1c2c(ccc1N1CCNCC1)cccc2</chem>
81	5018	<chem>Clc1cc2c(cc1O)C(CN(CC2)C)c1cccc1</chem>
82	5073	<chem>Fc1cc2onc(c2cc1)C1CCN(CC1)CCC=1C(=O)N2C(=NC=1C)CCCC2</chem>
83	5074	<chem>S1C=CN2C1=NC(C)=C(CCN1CCC(CC1)=C(c1ccc(F)cc1)c1ccc(F)cc1)C</chem>

2=O

84	5095	<chem>O=C1Nc2c(C1)c(ccc2)CCN(CCC)CCC</chem>
85	5163	<chem>O=C(Nc1ccnc1)N1CCc2c1cc1c(n(cc1)C)c2</chem>
86	5202	<chem>Oc1cc2c([nH]cc2CCN)cc1</chem>
87	5406	<chem>O=C(NC1CC2C(N(C1)C)Cc1c3c2cccc3[nH]c1)N(CC)CC</chem>
88	5452	<chem>S1c2c(N(c3c1cccc3)CCC1N(CCCC1)C)cc(SC)cc2</chem>
89	5533	<chem>Clc1cc(N2CCN(CC2)CCCN2N=C3N(C=CC=C3)C2=O)ccc1</chem>
90	5566	<chem>S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)C)cc(cc2)C(F)(F)F</chem>
91	5684	<chem>O(C)c1cccc1N1CCN(CC1)CCN(C(=O)C1CCCCC1)c1ncccc1</chem>
92	5709	<chem>N1CCN=C1Cc1c(cc(cc1C)C(C)(C)C)C</chem>
93	5736	<chem>Clc1cc2c(Sc3c(C=C2OCCN(C)C)cccc3)cc1</chem>
94	6018	<chem>O(C)c1cc2C3N(CC(CC(C)C)C(=O)C3)CCc2cc1OC</chem>
95	7638	<chem>O(Cc1cccc1)c1ccc(O)cc1</chem>
96	8794	<chem>N#CCc1cccc1</chem>
97	15443	<chem>Fc1ccc(cc1)C(=O)CCCN1CCN(CC1)c1ncccc1</chem>
98	15641	<chem>P(OCCCCCCCC)(OCCCCCCCC)(=O)c1cccc1</chem>
99	15897	<chem>FC(F)(F)c1cc(ccc1)CC(N)C</chem>
100	16106	<chem>N=1c2c(Cc3c(cccc3)C=1N1CCN(CC1)C)cccc2</chem>
101	16118	<chem>S(=O)(=[NH])(CCC(N)C(O)=O)C</chem>
102	16414	<chem>Clc1cc2N(c3c(Sc2cc1)cc(O)cc3)CCCN(C)C</chem>
103	21722	<chem>n1(c2CCCCCCc2c2c1cccc2)CCCN(C)C</chem>
104	23897	<chem>O1CCN(CC1)CC1CCc2[nH]c(C)c(c2C1=O)CC</chem>
105	27400	<chem>s1c2c(cc1)C(c1c(CC2)cccc1)=C1CCN(CC1)C</chem>

106 37586 Clc1[nH]c2c3c1CC1N(CC(CC1c3ccc2)CC#N)C  
 107 49381 Fc1cc2N=C(N3CCN(CC3)C)c3c(Cc2cc1)cccc3  
 108 54940 Brc1ccc(OC)c(C(=O)NCC2N(CCC2)CC)c1OC  
 109 60149 Clc1cc2c(n(cc2C2CCN(CC2)CCN2CCNC2=O)-c2ccc(F)cc2)cc1  
 110 60795 Clc1c(N2CCN(CC2)CCCCOc2cc3NC(=O)CCc3cc2)cccc1Cl  
 111 60809 s1nc(C=2CN(CCC=2)C)c(OCCCCC)n1  
 112 65489 Clc1cc2NC(=O)N(c2cc1)CCCN1CCC(CC1)C(=O)c1ccc(F)cc1  
 113 73333 Fc1ccc(cc1)C(CCCN1CCN(CC1)C(=O)NCC)c1ccc(F)cc1  
 114 91613 [nH]1c(cnc1C)CCN  
 115 99049 O1C(NC(=O)C2C=C3C(N(C2)C)Cc2c4c3cccc4[nH]c2)(C(C)C)C(=O)N2  
C(CC(C)C)C(=O)N3C(CCC3)C12O  
 116 107992 Clc1nc(N2CCNCC2)cnc1  
 117 108029 O(C)c1cc2c([nH]cc2C=2CCNCC=2)cc1  
 118 123932 S(Oc1ccc(cc1O)CCN)(O)(=O)=O  
 119 125085 Fc1ccc(cc1)C(=O)CCCN1CCC2(N(CN(CCc3ccc(N)cc3)C2=O)c2cccc2)  
CC1  
 120 131747 Clc1nc(N2CCC(N)CC2)ccc1  
 121 132564 O1CCCc2c3c(n(cc3CCN)C)ccc12  
 122 146224 Clc1cc(Cl)ccc1CNN\C=N/C(=O)c1nc(Cl)c(nc1N)N  
 123 291264 O1C2C(CCC(=C3C2C(=CC3=O)C)C)C(C)C1=O  
 124 589768 Oc1c2c([nH]cc2CC(N)C(O)=O)ccc1  
 125 592735 O1C(NC(=O)C2CC3C(N(C2)C)Cc2c4c3cccc4[nH]c2)(C(C)C)C(=O)N2C(  
C(CC)(C)C)C(=O)N3C(CCC3)C12O

126	627310	<chem>BrC1[nH]c2c3c1CC1N(CC(C=C1c3ccc2)C(=O)N(CC)CC)C</chem>
127	667466	<chem>Clc1cc\2c(Sc3c(cccc3)/C/2=C\CCN(C)C)cc1</chem>
128	667468	<chem>O1Cc2c(cccc2)\C(\c2c1cccc2)=C\CCN(C)C</chem>
129	1548943	<chem>O(C)c1cc(ccc1O)CNC(=O)CCCC\C=C\C(C)C</chem>
130	3065828	<chem>O(C)c1c(cccc1OC)C(=O)NC1CC2N(C(C1)CC2)Cc1cccc1</chem>
131	3408722	<chem>O(C)c1c2c(cccc2)c(cc1C(=O)NCC1N(CCC1)CCCC)C#N</chem>
132	3906894	<chem>Clc1cc2NC(=O)C3N(CCNC3)c2cc1Cl</chem>
133	4284720	<chem>s1cccc1COc1cc2c([nH]cc2CC(N)C)cc1</chem>
134	5251926	<chem>S(=O)(=O)(NC1CC2C(N(C1)C)Cc1c3c2cccc3n(c1)C)N(C)C</chem>
135	5281881	<chem>S1c2c(cc(cc2)C(F)(F)F)\C(\c2c1cccc2)=C/CCN1CCN(CC1)CCO</chem>
136	5474706	<chem>O(C)C=1C=CC2=NC=3C(=C2C=1)CCNC=3C</chem>
137	5487301	<chem>OCc1cc\2c(N=C/C/2=C\NN\C(=N\CCCC)\N)cc1</chem>
138	6422124	<chem>O(CCCc1[nH]cnc1)c1ccc(cc1)C(=O)C1CC1</chem>
139	6446980	<chem>Clc1cc2C=C(N3CCN(CC3)C)c3c(cccc3)\C(\c2cc1)=C\C#N</chem>
140	6510284	<chem>s1cc2c(c1)\C(\c1c(N=C2N2CCN(CC2)C)csc1)=C\C#N</chem>
141	6713986	<chem>Clc1cc(CC)c(O)c(C(=O)NCC2N(CCC2)CC)c1OC</chem>
142	9908697	<chem>OC1CCCC1NC(=O)C1CC2C(N(C1)C)Cc1c3c2cccc3n(c1)C(C)C</chem>
143	13995788	<chem>O(C(=O)C1CC2C(N(C1)C)Cc1c3c2cccc3n(c1)C(C)C)C1CCC(OC)CC1</chem>
144	14096625	<chem>s1nc(N2CCN(CC2)CCCCN2C(=O)C3C(CCCC3)C2=O)c2c1cccc2</chem>
145	14665495	<chem>Ic1cc(C(=O)NCC2N(CCC2)CC)c(OC)c(OC)c1</chem>
146	21982952	<chem>O=C(NC1CCCCC1)C1CC2C(N(C1)C)Cc1c3c2cccc3n(c1)C(C)C</chem>

### Appendix III

#### 608 5-HT<sub>2B</sub> Inactives Used in the Modeling Studies

Cp.	PubChe	SMILES
ID	CID	
147	89	<chem>Oc1cccc(C(=O)CC(N)C(O)=O)c1N</chem>
148	143	<chem>O=C1N=C(NC=2NCC(N(C1=2)C=O)CNc1ccc(cc1)C(=O)NC(CCC(O)=O)C(O)=O)N</chem>
149	191	<chem>O1C(CO)C(O)C(O)C1n1c2ncnc(N)c2nc1</chem>
150	199	<chem>N(CCCCN)=C(N)N</chem>
151	204	<chem>O=C1NC(=O)NC1NC(=O)N</chem>
152	225	<chem>OC1CC2CCC3C4CCC(=O)C4(CCC3C2(CC1)C)C</chem>
153	235	<chem>O1C(C(O)CO)C(=O)C(O)=C1O</chem>
154	253	<chem>S1CC2NC(=O)NC2C1CCCC(O)=O</chem>
155	275	<chem>O(N=C(N)N)CCC(N)C(O)=O</chem>
156	288	<chem>OC(CC(=O)[O-])C[N+](C)(C)C</chem>
157	298	<chem>ClC(Cl)C(=O)NC(C(O)c1ccc([N+](=O)[O-])cc1)CO</chem>
158	408	<chem>O=C1N(C)C(CC1)c1cccnc1</chem>
159	450	<chem>OC1CCC2C3C(CCC12C)c1c(cc(O)cc1)CC3</chem>
160	554	<chem>O=C([O-])C1[N+](CCC1)(C)C</chem>
161	653	<chem>O1c2c(cccc2)C(=O)C(CC=2C(=O)c3c(OC=2O)cccc3)=C1O</chem>
162	815	<chem>O1C(CN)C(O)C(O)C(O)C1OC1C(O)C(OC2OC(CO)C(O)C(N)C2O)C(N)CC1N</chem>



163 853 Ic1cc(cc(I)c1Oc1cc(I)c(O)c(I)c1)CC(N)C(O)=O  
 164 861 Ic1cc(cc(I)c1Oc1cc(I)c(O)cc1)CC(N)C(O)=O  
 165 925 P(OCC1OC([n+]2cc(ccc2)C(=O)N)C(O)C1O)(OP(OCC1OC(n2c3ncn  
 c(N)c3nc2)C(O)C1O)(O)=O)(=O)[O-]  
 166 932 O1c2c(C(=O)CC1c1ccc(O)cc1)c(O)cc(O)c2  
 167 1027 OC1CC2=CCC3C4CCC(C(=O)C)C4(CCC3C2(CC1)C)C  
 168 1046 O=C(N)c1ncnc1  
 169 1054 Oc1c(CO)c(cnc1C)CO  
 170 1072 O=C1NC(=O)N=C2N(c3cc(C)c(cc3N=C12)C)CC(O)C(O)C(O)CO  
 171 1130 s1c[n+](Cc2enc(nc2N)C)c(C)c1CCO  
 172 1203 O1c2c(CC(O)C1c1cc(O)c(O)cc1)c(O)cc(O)c2  
 173 1207 N1C2Cc3c(ccc3)C1(c1c2ccc1)C  
 174 1211 O1c2cc3C(N(CCc3cc2OC)C)Cc2ccc(Oc3c4C([N+](CCc4cc(OC)c3O  
 )(C)C)Cc3cc1c(O)cc3)cc2  
 175 1258 O1C2C3[N+](O-)(C(CC(OC(=O)C(CO)c4cccc4)C3)C12)C  
 176 1302 O(C)c1cc2c(cc(cc2)C(C(O)=O)C)cc1  
 177 1691 O1C(C)C(O)C(N)CC1OC1CC(O)(Cc2c1c(O)c1c(C(=O)c3c(C1=O)c(  
 OC)ccc3)c2O)C(=O)CO  
 178 1805 O1C(CO)C(O)C(O)C1N1C=NC(=NC1=O)N  
 179 1892 O=C1N(C)C(=O)N(c2ncn(c12)CCO)C  
 180 1978 O(CC(O)CNC(C)C)c1ccc(NC(=O)CCC)cc1C(=O)C  
 181 1981 Clc1ccc(cc1)C(=O)n1c2c(cc(OC)cc2)c(CC(OCC(O)=O)=O)c1C  
 182 1983 Oc1ccc(NC(=O)C)cc1

183 1986 s1c(nnc1S(=O)(=O)N)NC(=O)C  
 184 1989 S(=O)(=O)(NC(=O)NC1CCCCC1)c1ccc(cc1)C(=O)C  
 185 1990 O=C(NO)C  
 186 1993 O(C(C[N+](C)(C)C)C)C(=O)C  
 187 2012 O(C)C1C2(O)CC3C(C(OC(=O)C)(C4C5N(CC6(C(C35C(OC)CC6O)C4OC)COC)CC)C1O)C2OC(=O)c1cccc1  
 188 2022 O=C1N=C(Nc2n(cnc12)COCCO)N  
 189 2073 OC1N2C3C4C(CC2C2N(c5c(C2(C3)C4O)cccc5)C)C1CC  
 190 2082 S(CCC)c1cc2[nH]c(nc2cc1)NC(OC)=O  
 191 2083 Oc1ccc(cc1CO)C(O)CNC(C)(C)C  
 192 2094 O=C1N=CN=C2NNC=C12  
 193 2116 O1c2c(CCC1(CCCC(CCCC(CCCC(C)C)C)C)c(C)c(O)c(C)c2C  
 194 2130 NC12CC3CC(C1)CC(C2)C3  
 195 2142 O1C(CN)C(O)C(O)C(O)C1OC1C(O)C(OC2OC(CO)C(O)C(N)C2O)C(NC(=O)C(O)CCN)CC1N  
 196 2151 O=C1N(N(C)C(C)=C1N)c1cccc1  
 197 2153 O=C1N(C)C(=O)N(c2nc[nH]c12)C  
 198 2157 Ic1cc(cc(I)c1OCCN(CC)CC)C(=O)c1c2c(oc1CCCC)cccc2  
 199 2158 O1C2OC(OC2C(OCCCN(C)C)C1C(O)CO)(C)C  
 200 2173 S1C2N(C(C(=O)[O-])C1(C)C)C(=O)C2NC(=O)C(N)c1cccc1  
 201 2178 [n+]1(ccccc1C)Cc1enc(nc1N)CCC  
 202 2199 O(C(=O)C)C1C(NCC1O)Cc1ccc(OC)cc1  
 203 2202 Oc1c2c(Cc3c(C2=O)c(O)ccc3)ccc1

204 2206 O=C1N(N(C)C(=C1)C)c1ccccc1  
 205 2227 N(CCCCN=C(N)N)=C(N)N  
 206 2230 O(C(=O)C=1CN(CCC=1)C)C  
 207 2240 O1C2OC3(OOC24C(CCC(C4CC3)C)C(C)C1=O)C  
 208 2255 S(OC1C(OS(O)(=O)=O)C(OC(OC2CC3(C(CCC45C3CCC(C4)C(=C  
 )C5O)C(C2)C(O)=O)C)C1OC(=O)CC(C)C)CO)(O)(=O)=O  
 209 2265 S(c1n(cnc1[N+](=O)[O-])C)c1ncnc2nc[nH]c12  
 210 2271 S1C2N(C(C(O)=O)C1(C)C)C(=O)C2NC(=O)C(NC(=O)N1CCNC1=  
 O)c1ccccc1  
 211 2282 S1C2N(C(C(OC(OC(OCC)=O)C)=O)C1(C)C)C(=O)C2NC(=O)C(N)  
 c1ccccc1  
 212 2284 Clc1ccc(cc1)C(CC(O)=O)CN  
 213 2315 S(=O)(=O)(N)c1cc2S(=O)(=O)NC(Nc2cc1C(F)(F)F)Cc1ccccc1  
 214 2333 Brc1cc(cc(Br)c1O)C(=O)c1c2c(oc1CC)cccc2  
 215 2337 O(C(=O)c1ccc(N)cc1)CC  
 216 2343 Clc1cc2NC(=NS(=O)(=O)c2cc1S(=O)(=O)N)CSCc1ccccc1  
 217 2353 O1c2c(OC1)cc-1c(CC[n+])3c-1cc1c(c3)c(OC)c(OC)cc1)c2  
 218 2356 O1C2C(OC(=O)c3c2c(O)c(OC)c(O)c3)C(O)C(O)C1CO  
 219 2366 n1ccccc1CCNC  
 220 2371 OC1CCC2(C(CCC3(C2CCC2C4C(CCC23C)(CCC4C(C)=C)C(O)=O  
 )C)C1(C)C)C  
 221 2376 O1C(c2c(c3OCOc3cc2)C1=O)C1N(CCc2c1cc1OCOc1c2)C  
 222 2391 O(C(=O)C)c1ccc(cc1)C(c1ccc(OC(=O)C)cc1)c1ncccc1

223 2448 Brc1ccc(cc1)C1(O)CCN(CC1)CCCC(=O)c1ccc(F)cc1  
 224 2462 O1C2(C(OC1CCC)CC1C3C(C4(C(=CC(=O)C=C4)CC3)C)C(O)CC1  
 2C)C(=O)CO  
 225 2466 O(CCCC)c1ccc(cc1)CC(=O)NO  
 226 2471 S(=O)(=O)(N)c1cc(cc(NCCCC)c1Oc1ccccc1)C(O)=O  
 227 2478 S(OCCCCOS(=O)(=O)C)(=O)(=O)C  
 228 2482 O(C(=O)c1ccc(N)cc1)CCCC  
 229 2485 O1C(CN)C(O)C(O)C(N)C1OC1C(OC2OC(CO)C(O)C2O)C(O)C(NC  
 (=O)C(O)CCN)CC1N  
 230 2550 SCC(C(=O)N1CCCC1C(O)=O)C  
 231 2551 O(CC[N+](C)(C)C)C(=O)N  
 232 2554 O=C(N)N1c2c(C=Cc3c1cccc3)cccc2  
 233 2560 S1C2N(C(C(O)=O)C1(C)C)C(=O)C2NC(=O)C(C(O)=O)c1ccccc1  
 234 2561 O(C(=O)CCC(O)=O)C1CCC2(C3C(CCC2C1(C)C)(C)C1(C(C2CC(C  
 CC2(CC1)C)(C(O)=O)C)=CC3=O)C)C  
 235 2562 O(C(=O)C1(CCCC1)c1ccccc1)CCOCCN(CC)CC  
 236 2564 Clc1ccc(cc1)C(OCCN(C)C)c1ncccc1  
 237 2574 O=C(NCCc1[nH]cnc1)CCN  
 238 2576 O(CC(CCC)(COC(=O)N)C)C(=O)NC(C)C  
 239 2578 ClCCN(N=O)C(=O)NCCCCl  
 240 2609 ClC=1CSC2N(C(=O)C2NC(=O)C(N)c2ccccc2)C=1C(O)=O  
 241 2610 S1C2N(C(=O)C2NC(=O)C(N)c2ccc(O)cc2)C(C(O)=O)=C(C1)C  
 242 2615 S1C2N(C(=O)C2NC(=O)C(OC=O)c2ccccc2)C(C(O)=O)=C(C1)CS c1

nnnn1C

243 2617 s1c(nnc1SCC=1CSC2N(C(=O)C2NC(=O)Cn2nnc2)C=1C(=O)[O-]  
 ])C

244 2625 S1C2N(C(=O)C2(OC)NC(=O)CSCC#N)C(C(=O)[O-]  
 )=C(C1)CSc1nnnn1C

245 2630 S1C2N(C(=O)C2NC(=O)C(NC(=O)N2CCN(CC)C(=O)C2=O)c2ccc(  
 O)cc2)C(C(O)=O)=C(C1)CSc1nnnn1C

246 2637 s1cccc1CC(=O)NC1(OC)C2SCC(COC(=O)N)=C(N2C1=O)C(=O)[O  
 -]

247 2665 O(C)c1cc2C3N(CC(CC)C(C3)CC3NCCc4c3cc(OC)c(O)c4)CCc2cc1  
 OC

248 2666 S1C2N(C(=O)C2NC(=O)C(N)c2ccccc2)C(C(O)=O)=C(C1)C

249 2670 s1cccc1CC(=O)NC1C2SCC(COC(=O)C)=C(N2C1=O)C(=O)[O-]

250 2672 S1C2N(C(=O)C2NC(=O)CSc2ccncc2)C(C(=O)[O-]  
 )=C(C1)COC(=O)C

251 2678 Clc1ccc(cc1)C(N1CCN(CC1)CCOCC(O)=O)c1cccc1

252 2717 Clc1ccc(cc1)C1S(=O)(=O)CCC(=O)N1C

253 2719 Clc1cc2nccc(NC(CCCN(CC)CC)C)c2cc1

254 2720 Clc1cc2NC=NS(=O)(=O)c2cc1S(=O)(=O)N

255 2724 Clc1ccc(OCC(O)COC(=O)N)cc1

256 2727 Clc1ccc(S(=O)(=O)NC(=O)NCCC)cc1

257 2732 Clc1ccc(cc1S(=O)(=O)N)C1(O)NC(=O)c2c1cccc2

258 2733 Clc1cc2NC(Oc2cc1)=O

259 2749 O=C1N(O)C(=CC(=C1)C)C1CCCCC1  
 260 2757 OC(C1N2CC(C(C1)CC2)C=C)c1c2c(ncc1)cccc2  
 261 2762 O1c2c(OC1)cc1N(N=C(C(O)=O)C(=O)c1c2)CC  
 262 2764 Fc1cc2c(N(C=C(C(O)=O)C2=O)C2CC2)cc1N1CCNCC1  
 263 2784 O(C(=O)C(O)(c1ccccc1)c1ccccc1)C1C2CC[N+](C1)(CC2)C  
 264 2786 ClC(C(NC(=O)C1N(CC(C1)CCC)C)C1OC(SC)C(O)C(O)C1O)C  
 265 2791 ClCC(=O)C1(OC(=O)CC)C2(CC(O)C3(F)C(C2CC1C)CCC1=CC(=O)C=CC13C)C  
 266 2794 Clc1ccc(N2C3=C\C(=N/C(C)C)\C(Nc4ccc(Cl)cc4)=CC3=Nc3c2cccc3)cc1  
 267 2797 Clc1ccc(OC(C(O)=O)(C)C)cc1  
 268 2798 Clc1ccc(cc1)CCCC[N+](CCCCCCC)(CC)CC  
 269 2812 Clc1ccccc1C(n1ccnc1)(c1ccccc1)c1ccccc1  
 270 2813 Clc1ccccc1-c1noc(C)c1C(=O)NC1C2SC(C)(C)C(N2C1=O)C(=O)[O-]  
 ]  
 271 2833 O(C)C1=CC=C2C(=CC1=O)C(NC(=O)C)CCc1c2c(OC)c(OC)c(OC)c1  
 272 2860 N(C)(C)C1CC2=CCC3C(CCC45C(CCC34)C(N(C5)C)C)C2(CC1)C  
 273 2882 O1c2c(C(=O)C=C1C(O)=O)c(OCC(O)COc1c3c(OC(=CC3=O)C(O)=O)ccc1)ccc2  
 274 2900 O=C1C(CC(CC1C)C)C(O)CC1CC(=O)NC(=O)C1  
 275 2907 ClCCN(P1(OCCCN1)=O)CCCl  
 276 2949 o1ncc2CC3(C4C(C5CCC(O)(C#C)C5(CC4)C)CCC3=Cc12)C

277 2955 S(=O)(=O)(c1ccc(N)cc1)c1ccc(N)cc1  
 278 2958 O1C(C)C(O)C(N)CC1OC1CC(O)(Cc2c1c(O)c1c(C(=O)c3c(C1=O)c(OC)ccc3)c2O)C(=O)C  
 279 2966 NC(=N)N1CCc2c(C1)cccc2  
 280 2973 O=C(N(O)CCCCNC(=O)CCC(=O)N(O)CCCCCN)CCC(=O)NCCC  
 CCN(O)C(=O)C  
 281 2975 O=C1CC2C(C3CCC(C(CCC(O)=O)C)C13C)C(=O)CC1CC(=O)CCC  
 12C  
 282 2993 [n+]1(c2c(cccc2)c(N)cc1C)CCCCCCCC[n+]1c2c(cccc2)c(N)cc1C  
 283 2995 N(CCCN1c2c(CCc3c1cccc3)cccc2)C  
 284 3003 FC12C(C3CC(C)C(O)(C(=O)CO)C3(CC1O)C)CCC1=CC(=O)C=CC  
 12C  
 285 3008 O(C)c1cc2C34C(C(N(CC3)C)Cc2cc1)CCCC4  
 286 3019 Clc1cc2S(=O)(=O)N=C(Nc2cc1)C  
 287 3025 O(CCCC)c1nc2c(cccc2)c(c1)C(=O)NCCN(CC)CC  
 288 3038 Clc1c(S(=O)(=O)N)cc(S(=O)(=O)N)cc1Cl  
 289 3040 Clc1cccc(Cl)c1-  
 c1noc(C)c1C(=O)NC1C2SC(C)(C)C(N2C1=O)C(=O)[O-]  
 290 3052 O=C(N(CC)CC)N1CCN(CC1)C  
 291 3059 Fe1cc(F)ccc1-c1cc(C(O)=O)c(O)cc1  
 292 3062 O1C(C)C(OC2OC(C)C(O)C(O)C2)C(O)CC1OC1C(OC(OC2CC3CC  
 C4C(CC(O)C5(C)C(CCC45O)C4=CC(OC4)=O)C3(CC2)C)CC1O)C  
 293 3069 O1C(CO)C(O)C(O)C(NC)C1OC1C(O)(CO)C(OC1OC1C(N=C(N)N)

C(O)C(N=C(N)N)C(O)C1O)C  
 294 3076 S1c2c(N(CCN(C)C)C(=O)C(OC(=O)C)C1c1ccc(OC)cc1)cccc2  
 295 3081 O1C(C)(C)C(=O)NC1=O  
 296 3108 OCCN(CCO)c1nc(N2CCCCC2)c2nc(nc(N3CCCCC3)c2n1)N(CCO)  
CCO  
 297 3110 S(=O)(=O)([O-])CN(C)C=1C(=O)N(N(C)C=1C)c1cccc1  
 298 3114 O=C(N)C(CCN(C(C)C)C(C)C)(c1cccc1)c1ncccc1  
 299 3132 SCC(Cc1cccc1)C(=O)NCC(O)=O  
 300 3162 O(C(C)(c1cccc1)c1ncccc1)CCN(C)C  
 301 3168 Fc1ccc(cc1)C(=O)CCCN1CCC(N2c3c(NC2=O)cccc3)=CC1  
 302 3169 OC(CN1CCN(CC1)c1cccc1)CO  
 303 3180 O(CCCC)c1ccc(cc1)C(=O)CCN1CCCCC1  
 304 3182 O=C1N(C)C(=O)N(c2ncn(c12)CC(O)CO)C  
 305 3195 O=C1n2c3C4N(CCCC4(C1)CC)CCc3c1c2cccc1  
 306 3198 Clc1cc(Cl)ccc1C(OCc1ccc(Cl)cc1)Cn1ccnc1  
 307 3202 Oc1cc([N+](CC)(C)C)ccc1  
 308 3222 O(C(=O)C(NC(C(=O)N1CCCC1C(O)=O)C)CCc1cccc1)CC  
 309 3242 O(C)c1n(nc(c1)C)-c1nc(cc(OC)n1)C  
 310 3247 Oc1cc2CC=C3C4CCC(=O)C4(CCC3c2cc1)C  
 311 3255 O1C(CC)C(O)(C)C(O)C(C)C(=O)C(CC(O)(C)C(OC2OC(CC(N(C)C)C2O)C)C(C)C(OC2OC(C)C(O)C(OC)(C2)C)C(C)C1=O)C  
 312 3276 OCCNC(=O)Cn1ccnc1[N+](=O)[O-]  
 313 3279 OCC(NCCNC(CC)CO)CC



314 3280 O(CC)c1cc(ccc1OCC)Cc1nccc2c1cc(OCC)c(OCC)c2  
 315 3288 OC1(CCC2C3C(CCC12C)C1(C=CC(=O)CC1)CC3)C#C  
 316 3291 O=C1NC(=O)CC1(CC)C  
 317 3308 O1CCc2c([nH]c3c2cccc3CC)C1(CC(O)=O)CC  
 318 3310 O1C2C(OC(OC2)C)C(O)C(O)C1OC1C2C(C(c3c1cc1OCOc1c3)c1cc  
 (OC)c(O)c(OC)c1)C(OC2)=O  
 319 3333 Clc1c(cccc1Cl)C1C(C(OCC)=O)=C(NC(C)=C1C(OC)=O)C  
 320 3335 OC(=O)CCC(=O)c1ccc(cc1)-c1ccccc1  
 321 3342 O(c1cc(ccc1)C(C(O)=O)C)c1ccccc1  
 322 3343 Oc1cc(cc(O)c1)C(O)CNC(Cc1ccc(O)cc1)C  
 323 3351 Clc1ccc(OCC(=O)N2CCN(CC2)Cc2cc3OCOc3cc2)cc1  
 324 3354 O1c2c(cccc2C(OCCN2CCCC2)=O)C(=O)C(C)=C1c1ccccc1  
 325 3371 FC(F)(F)c1cc(Nc2ccccc2C(O)=O)ccc1  
 326 3374 Fc1cc2c3N(C=C(C(O)=O)C2=O)C(CCc3c1)C  
 327 3375 FC12C(C3CC(C)C(O)(C(=O)CO)C3(CC1O)C)CC(F)C1=CC(=O)C=  
 CC12C  
 328 3379 FC1C2=CC(=O)C=CC2(C2C(C3CC4OC(OC4(C(=O)CO)C3(CC2O)  
 C)(C)C)C1)C  
 329 3382 FC12C(C3CC4OC(OC4(C(=O)COC(=O)C)C3(CC1O)C)(C)C)CC(F)  
 C1=CC(=O)C=CC12C  
 330 3384 FC12C(C3CCC(O)(C(=O)C)C3(CC1O)C)CC(C1=CC(=O)C=CC12C  
 )C  
 331 3392 FC1C2=CC(=O)CCC2(C2C(C3CC4OC(OC4(C(=O)CO)C3(CC2O)C

)(C)C)C1)C  
 332 3405 O=C1N=C(Nc2ncc(nc12)CNc1ccc(cc1)C(=O)NC(CCC(O)=O)C(O)=  
 O)N  
 333 3414 P(=O)([O-])([O-])C(=O)[O-]  
 334 3418 P(Oc1ccccc1C(O)=O)(O)(O)=O  
 335 3442 OC(=O)c1ncc(cc1)CCCC  
 336 3446 OC(=O)CC1(CCCCC1)CN  
 337 3449 O1c2c3C4(C1CC(O)C=C4)CCN(Cc3ccc2OC)C  
 338 3454 O=C1N=C(Nc2n(cnc12)COC(CO)CO)N  
 339 3467 O1C(OC2C(O)C(OC3OCC(O)(C)C(NC)C3O)C(N)CC2N)C(N)CCC1  
 C(NC)C  
 340 3475 S(=O)(=O)(NC(=O)NN1CC2C(CCC2)C1)c1ccc(cc1)C  
 341 3478 S(=O)(=O)(NC(=O)NC1CCCC1)c1ccc(cc1)CCNC(=O)c1ncc(nc1)C  
 342 3488 Clc1cc(C(=O)NCCc2ccc(S(=O)(=O)NC(=O)NC3CCCC3)cc2)c(OC  
 )cc1  
 343 3503 Oc1c(O)c(c2c(cc(C)c(-  
 c3c(cc4c(c(C=O)c(O)c(O)c4C(C)C)c3O)C)c2O)c1C(C)C)C=O  
 344 3516 O(CC(O)CO)c1ccccc1OC  
 345 3553 ClCC(=O)C12OC(OC1CC1C3CCC4=CC(=O)CCC4(C)C3(F)C(O)C  
 C12C)(C)C  
 346 3561 IC#CCOc1cc(Cl)c(Cl)cc1Cl  
 347 3573 O1CC(CCC12OC1C(C3(C(C4C(CC3=O)C3(C(CC(O)CC3)CC4)C)C  
 1)C)C2C)C

348 3590 OC(CCCC(N)C)(C)C  
 349 3604 [N+](CCCCC[N+](C)(C)C)(C)(C)C  
 350 3606 Oc1ccc(cc1)C(C(CC)c1ccc(O)cc1)CC  
 351 3607 NC1(CN(CN(C1)CC(CCCC)CC)CC(CCCC)CC)C  
 352 3623 O(C(=O)C(O)c1ccccc1)C1CC2N(C(C1)CC2)C  
 353 3637 n1ncc2c(cccc2)c1NN  
 354 3639 Clc1cc2NCNS(=O)(=O)c2cc1S(=O)(=O)N  
 355 3640 OC1(CCC2C3C(C4(C(=CC(=O)CC4)CC3)C)C(O)CC12C)C(=O)CO  
 356 3647 S(=O)(=O)(N)c1cc2S(=O)(=O)NCNc2cc1C(F)(F)F  
 357 3649 O(C)c1cc2c(nccc2C(O)C2N3CC(C(C2)CC3)CC)cc1  
 358 3661 O(C(=O)C(CO)c1ccccc1)C1CC2N(C(C1)CC2)C  
 359 3676 O=C(Nc1c(cccc1C)C)CN(CC)CC  
 360 3677 N(C(Cc1ccccc1)(C)C)C  
 361 3687 IC1=CN(C2OC(CO)C(O)C2)C(=O)NC1=O  
 362 3698 O=C1NC=C(C=C1N)c1ccncc1  
 363 3718 O=C1N(Cc2c1ccccc2)c1ccc(cc1)C(C(O)=O)C  
 364 3730 Ic1c(C(=O)NCC(O)CO)c(I)c(N(C(=O)C)CC(O)CO)c(I)c1C(=O)NCC  
 (O)CO  
 365 3746 O(C(=O)C(CO)c1ccccc1)C1CC2[N+](C(C1)CC2)(C(C)C)C  
 366 3748 O=C(NNC(C)C)c1ccncc1  
 367 3760 Clc1cccc(Cl)c1COC(Cn1ccnc1)c1ccc(Cl)cc1Cl  
 368 3762 Oc1cc(ccc1O)C(O)C(NC(C)C)CC  
 369 3767 O=C(NN)c1ccncc1

370 3775 O=C(N)C(CC[N+](C(C)C)(C(C)C)C)(c1ccccc1)c1ccccc1  
 371 3780 O1C2C(OCC2O[N+](=O)[O-])C(O[N+](=O)[O-])C1  
 372 3783 O(CC(NC(C(O)c1ccc(O)cc1)C)C)c1ccccc1  
 373 3823 Clc1cc(Cl)ccc1C1(OC(CO1)COc1ccc(N2CCN(CC2)C(=O)C)cc1)Cn  
 1ccnc1  
 374 3825 OC(=O)C(C)c1cc(ccc1)C(=O)c1ccccc1  
 375 3830 o1ccccc1CNc1ncnc2nc[nH]c12  
 376 3879 O1C(CO)C(O)C(O)C(O)C1OC1C(OC(OC2C(OC(OC3C(OC(OC4CC  
 5CCC6C(CC(O)C7(C)C(CCC67O)C6=CC(OC6)=O)C5(CC4)C)CC3  
 O)C)CC2O)C)CC1OC(=O)C)C  
 377 3888 O1C(CC(C)C1C(C(=O)C(C(O)C(CCc1ccc(C)c(O)c1C(O)=O)C)C)C  
 C)(CC)C1OC(C)C(O)(CC1)CC  
 378 3913 S1CCN2CC(N=C12)c1ccccc1  
 379 3917 Oc1cc(ccc1O)C(O)C(N)C  
 380 3928 S(C)C1OC(C(NC(=O)C2N(CC(C2)CCC)C)C(O)C)C(O)C(O)C1O  
 381 3937 OC(=O)C1N(CCC1)C(=O)C(NC(CCc1ccccc1)C(O)=O)CCCCN  
 382 3945 OC(CC1N(C)C(CCC1)CC(=O)c1ccccc1)c1ccccc1  
 383 3948 Fc1c2N(C=C(C(O)=O)C(=O)c2cc(F)c1N1CC(NCC1)C)CC  
 384 3955 Clc1ccc(cc1)C1(O)CCN(CC1)CCC(C(=O)N(C)C)(c1ccccc1)c1ccccc  
 1  
 385 3962 O1C(CC(O)CC1=O)CCC1C2C(=CC(CC2OC(=O)C(CC)C)C)C=CC1  
 C  
 386 3966 N12C(C3CC(C4N(C3)CCCC4)C1)CCCC2

387 3978 O1c2c(OC1)cc1c(C3C4N(CCC4=CC(O)C3O)C1)c2  
 388 3998 S(=O)(=O)(N)c1ccc(cc1)CN  
 389 4030 O(C(=O)Nc1[nH]c2cc(ccc2n1)C(=O)c1cccc1)C  
 390 4034 Clc1ccc(cc1)C(N1CCN(CC1)Cc1cc(ccc1)C)c1cccc1  
 391 4036 Clc1c(Nc2ccccc2C(=O)[O-])c(Cl)ccc1C  
 392 4039 Clc1ccc(OCC(OCCN(C)C)=O)cc1  
 393 4043 OC1C2C(C3CCC(C(=O)C)C3(C1)C)CC(C1=CC(=O)CCC12C)C  
 394 4045 O(CC(=O)NCCN(CC)CC)c1ccc(OC)cc1  
 395 4046 FC(F)(F)c1c2nc(cc(c2ccc1)C(O)C1NCCCC1)C(F)(F)F  
 396 4053 ClCCN(CCCl)c1ccc(cc1)CC(N)C(O)=O  
 397 4055 O=C1c2c(cccc2)C(=O)C=C1C  
 398 4057 O(C(=O)C(O)(c1cccc1)c1cccc1)C1CCC[N+](C1)(C)C  
 399 4059 O(CC(O)CO)c1cccc1C  
 400 4077 S(=O)(=O)([O-])CCS  
 401 4078 S1c2c(N(c3c1cccc3)CCC1N(CCCC1)C)cc(S(=O)C)cc2  
 402 4080 O(C)c1cc2CCC3C4CCC(O)(C#C)C4(CCC3c2cc1)C  
 403 4086 Oc1cc(cc(O)c1)C(O)CNC(C)C  
 404 4087 Oc1cc(ccc1)C(O)C(N)C  
 405 4091 N(C(N=C(N)N)=N)(C)C  
 406 4098 s1cccc1CN(CCN(C)C)c1ncccc1  
 407 4100 S1\C(=N/C(=O)C)\N(N=C1S(=O)(=O)N)C  
 408 4101 N12CN3CN(C1)CN(C2)C3  
 409 4107 O(CC(O)COC(=O)N)c1cccc1OC

410 4112 OC(=O)C(NC(=O)c1ccc(N(Cc2nc3c(nc(nc3N)N)nc2)C)cc1)CCC(O)=O  
 411 4114 O1c2c(C=CC1=O)cc1c(occ1)c2OC  
 412 4122 s1cccc1C(=O)c1cc2[nH]c(nc2cc1)NC(OC)=O  
 413 4138 Oc1cc(ccc1O)CC(N)(C(O)=O)C  
 414 4139 S1C2=CC(=[N+](C)C)C=CC2=Nc2c1cc(N(C)C)cc2  
 415 4140 OCC(NC(=O)C1C=C2C(N(C1)C)Cc1c3c2cccc3[nH]c1)CC  
 416 4159 OC1(CCC2C3C(C4(C(=CC(=O)C=C4)C(C3)C)C)C(O)CC12C)C(=O)CO  
 417 4165 S(=O)(=O)(N)c1cc2S(=O)(=O)CCCc2cc1C  
 418 4170 Clc1cc2NC(N(c3ccccc3C)C(=O)c2cc1S(=O)(=O)N)C  
 419 4171 O(CC(O)CNC(C)C)c1ccc(cc1)CCOC  
 420 4189 Clc1cc(Cl)ccc1C(OCc1ccc(Cl)cc1Cl)Cn1ccnc1  
 421 4195 O(C)c1ccc(OC)cc1C(O)CNC(=O)CN  
 422 4196 OC1(CCC2C3C(=C4C(=CC(=O)CC4)CC3)C(CC12C)c1ccc(N(C)C)cc1)C#CC  
 423 4201 ON1C(N)=CC(=NC1=N)N1CCCCC1  
 424 4211 Clc1cccc1C(C(Cl)Cl)c1ccc(Cl)cc1  
 425 4212 Oc1c2c(C(=O)c3c(C2=O)c(NCCNCCO)ccc3NCCNCCO)c(O)cc1  
 426 4240 ClC12C(C3CC(C)C(OC(=O)c4occc4)(C(=O)CCl)C3(CC1O)C)CCC1=CC(=O)C=CC12C  
 427 4243 O1C(C(CC(C)C1(O)CO)C)C1OC(C2(OC(CC2)C2(OC3(OC(C(C(OC)C(C(O)=O)C)C)C(C)C(O)C3)CC2)C)CC)C(C1)C

428 4246 O1C2C3N(CC2)CC=C3COC(=O)C(O)(C)C(O)(C)C(C)C1=O  
 429 4260 O(C(=O)C)c1cc(C(C)C)c(OCCN(C)C)cc1C  
 430 4411 O(CC(O)CNC(C)(C)C)c1c2CC(O)C(O)Cc2ccc1  
 431 4419 O1C2C34CCN(C(Cc5c3c1c(O)cc5)C4(O)CCC2O)CC1CCC1  
 432 4425 O1C2C34CCN(C(Cc5c3c1c(O)cc5)C4(O)CCC2=O)CC=C  
 433 4428 O1C2C34CCN(C(Cc5c3c1c(O)cc5)C4(O)CCC2=O)CC1CC1  
 434 4436 N1CCN=C1Cc1c2c(ccc1)cccc2  
 435 4441 O1C(CO)C(O)C(O)C(OC2OC(C)C(O)C(O)C2O)C1Oc1cc(O)c2c(OC  
 (CC2=O)c2ccc(O)cc2)c1  
 436 4454 O1C(CN)C(O)C(O)C(N)C1OC1C(O)C(OC1CO)OC1C(OC2OC(CN)  
 C(O)C(O)C2N)C(N)CC(N)C1O  
 437 4456 O(C(=O)N(C)C)c1cc([N+](C)(C)C)ccc1  
 438 4474 O(C(=O)C=1C(C(C(OC)=O)=C(NC=1C)C)c1cc([N+](=O)[O-  
 ])ccc1)CCN(Cc1cccc1)C  
 439 4477 Clc1cc([N+](=O)[O-])ccc1NC(=O)c1cc(Cl)ccc1O  
 440 4487 O=C1N(N(C)C(C)=C1NC(=O)c1ccnc1)c1cccc1  
 441 4488 FC(F)(F)c1cc(Nc2ncccc2C(O)=O)ccc1  
 442 4495 S(=O)(=O)(Nc1ccc([N+](=O)[O-])cc1Oc1cccc1)C  
 443 4528 Nc1c2c(ccc1)C(CN(C2)C)c1cccc1  
 444 4536 OC1(CCC2C3C(C4C(=CC(=O)CC4)CC3)CCC12C)C#C  
 445 4537 OC1(CCC2C3C(C4=C(CC(=O)CC4)CC3)CCC12C)C#C  
 446 4539 Fc1cc2c(N(C=C(C(O)=O)C2=O)CC)cc1N1CCNCC1  
 447 4542 OC1(CCC2C3C(C4C(=CC(=O)CC4)CC3)CCC12CC)C#C

448 4544 O1C(c2c(c(OC)c(OC)cc2)C1=O)C1N(CCC2c1c(OC)c1OCOc1c2)C  
 449 4583 Fc1cc2c3N(C=C(C(O)=O)C2=O)C(COc3c1N1CCN(CC1)C)C  
 450 4587 O1C(C)C(C)C(O)C(C)C(=O)C2(OC2)CC(C)C(OC2OC(CC(N(C)C)C  
 2O)C)C(C)C(OC2OC(C)C(O)C(OC)C2)C(C)C1=O  
 451 4594 S(=O)(Cc1ncc(C)c(OC)c1C)c1[nH]c2cc(OC)ccc2n1  
 452 4605 O1C(C)C(O)C(O)C(O)C1OC1CC2(O)CCC3C(C2(CO)C(O)C1)C(O)  
 CC1(C)C(CCC13O)C1=CC(OC1)=O  
 453 4621 OCCN(CC(=O)N(C(Cc1cccc1)(C)C)C)CC(=O)N(C(Cc1cccc1)(C)  
 C)C  
 454 4628 O1c2c(OC1)cc1N(C=C(C(O)=O)C(=O)c1c2)CC  
 455 4666 O1C2CC(O)C3(C(C(OC(=O)c4cccc4)C4(O)CC(OC(=O)C(O)C(NC(  
 =O)c5cccc5)c5cccc5)C(=C(C4(C)C)C(OC(=O)C)C3=O)C)C2(OC(  
 =O)C)C1)C  
 456 4678 OC(C(CO)(C)C)C(=O)NCCCCO  
 457 4680 O(C)c1cc(ccc1OC)Cc1nccc2c1cc(OC)c(OC)c2  
 458 4688 N(Cc1cccc1)(CC#C)C  
 459 4689 O1C(CN)C(O)C(O)C(N)C1OC1C(O)C(OC1CO)OC1C(OC2OC(CO)  
 C(O)C(O)C2N)C(N)CC(N)C1O  
 460 4735 O(CCCCCOc1ccc(cc1)C(N)=N)c1ccc(cc1)C(N)=N  
 461 4740 O=C1N(CCCCC(=O)C)C(=O)N(c2ncn(c12)C)C  
 462 4742 OC(C(NC(=O)C(NC(=O)C(NC(=O)CC(C)C)C(C)C)C(C)C)CC(C)C  
 CC(=O)NC(C(=O)NC(C(O)CC(O)=O)CC(C)C)C  
 463 4746 N1CCCCC1CC(C1CCCCC1)C1CCCCC1



464 4754 O(CC)c1ccc(NC(=O)C)cc1  
 465 4758 S1C2N(C(C(=O)[O-])C1(C)C)C(=O)C2NC(=O)C(Oc1ccccc1)C  
 466 4760 O=C1c2c(cccc2)C(=O)C1c1ccccc1  
 467 4761 n1ccccc1C(CCN(C)C)c1ccccc1  
 468 4806 s1ccnc1NS(=O)(=O)c1ccc(NC(=O)c2ccccc2C(O)=O)cc1  
 469 4811 O(C(=O)NC)c1cc2c(N(C3N(CCC23C)C)C)cc1  
 470 4814 O(C)c1ccc(cc1C(=O)NCc1ccnc1)C(=O)NCc1ccnc1  
 471 4816 O1C2C3OC(=O)C45OC4CC(O)(C(C2C(C)=C)C1=O)C35C  
 472 4819 O1CC(Cc2n(cnc2)C)C(CC)C1=O  
 473 4828 O(CC(O)CNC(C)C)c1c2c([nH]cc2)ccc1  
 474 4832 O(C(=O)C(O)(c1ccccc1)c1ccccc1)C1CCC[N+](C1)(CC)C  
 475 4843 O=C1N(CCC1)CC(=O)N  
 476 4848 O=C1Nc2cccnc2N(c2c1cccc2)C(=O)CN1CCN(CC1)C  
 477 4855 O=C1c2c(nc(nc2)N2CCCC2)N(C=C1C(O)=O)CC  
 478 4865 O1CC2C(C(c3c(cc4OCOc4c3)C2O)c2cc(OC)c(OC)c(OC)c2)C1=O  
 479 4883 O(CC(O)CNC(C)C)c1ccc(NC(=O)C)cc1  
 480 4886 O1CCN(CC1)CCCOc1ccc(OCCCC)cc1  
 481 4894 OC1(CCC2C3C(C4(C(=CC(=O)C=C4)CC3)C)C(O)CC12C)C(=O)C  
 O  
 482 4900 OC1(CCC2C3C(C4(C(=CC(=O)C=C4)CC3)C)C(=O)CC12C)C(=O)  
 CO  
 483 4904 OC(CCN1CCCC1)(c1ccccc1)c1ccccc1  
 484 4906 O=C(Nc1ccccc1C)C(NCCC)C

485 4908 O(C)c1cc(NC(CCCN)C)c2ncccc2c1  
 486 4911 S(=O)(=O)(N(CCC)CCC)c1ccc(cc1)C(O)=O  
 487 4919 OC(CCN1CCCC1)(C1CCCCC1)c1cccc1  
 488 4920 O=C1CCC2(C3C(C4CCC(C(=O)C)C4(CC3)C)CCC2=C1)C  
 489 4922 OC(=O)CCC(NC(=O)c1cccc1)C(=O)N(CCC)CCC  
 490 4934 O1c2c(cccc2)C(c2c1cccc2)C(OCC[N+](C(C)C)(C(C)C)C)=O  
 491 4974 O1C23C(C4(O)C(C5C(C(O)C4OC(=O)C(CC)C)C(O)(C4N(CC(CC4)C)C5)C)C2)C(OC(=O)C)C(OC(=O)C)C2C1(O)C(OC(=O)C(O)(CC)C)CCC23C  
 492 4984 O1C(CO)C(NC(=O)C(N)Cc2ccc(OC)cc2)C(O)C1n1c2ncnc(N(C)C)c2nc1  
 493 4993 Clc1ccc(cc1)-c1c(nc(nc1N)N)CC  
 494 4994 O=C1C=CNC(=O)C1(CC)CC  
 495 5037 O=C1N(N(C)C(C)=C1NC(C)C)c1cccc1  
 496 5052 O(C)C1C(C2C(CC1OC(=O)c1cc(OC)c(OC)c(OC)c1)CN1C(C2)c2[nH]c3cc(OC)ccc3c2CC1)C(OC)=O  
 497 5066 O1C(CN)C(O)C(O)C(N)C1OC1C(OC2OC(CO)C(O)C2O)C(O)C(N)CC1N  
 498 5070 s1c2cc(OC(F)(F)F)ccc2nc1N  
 499 5102 O1c2c(CC1C(C)=C)c1OC3C(c4cc(OC)c(OC)cc4OC3)C(=O)c1cc2  
 500 5184 O1C2C3N(C(CC(OC(=O)C(CO)c4cccc4)C3)C12)C  
 501 5195 N(C(Cc1cccc1)C)(CC#C)C  
 502 5198 ClCCN(N=O)C(=O)NC1CCC(CC1)C

503 5215 S(=O)(=O)(Nc1ncccn1)c1ccc(N)cc1  
 504 5224 O1C(OC2C(O)C(OC3OCC(O)(C)C(NC)C3O)C(N)CC2N)C(N)CC=C1CN  
 505 5250 O1C2C(C(C)C13NCC(CC3)C)C1(C(C3C(CC1)C1(C(CC(O)CC1)=C3)C)C2)C  
 506 5267 S(C(=O)C)C1C2C3CCC4(OC(=O)CC4)C3(CCC2C2(C(C1)=CC(=O)CC2)C)C  
 507 5275 O(C)c1ccc(cc1)-c1n[n+](CCCC(O)=O)c(N)cc1  
 508 5297 O1C(CO)C(O)C(O)C(NC)C1OC1C(O)(C=O)C(OC1OC1C(N=C(N)N)C(O)C(N=C(N)N)C(O)C1O)C  
 509 5299 OC(C(NC(=O)N(N=O)C)C=O)C(O)C(O)CO  
 510 5303 O1CC(=CC1=O)C1CCC2(O)C3C(CCC12C)C1(CCC(O)CC1(O)CC3)C=O  
 511 5315 s1ccnc1NS(=O)(=O)c1ccc(NC(=O)CCC(O)=O)cc1  
 512 5318 Clc1cc(Cl)ccc1C(SCc1ccc(Cl)cc1)Cn1ccnc1  
 513 5319 S(=O)(=O)(NC(=O)c1ccccc1)c1ccc(N)cc1  
 514 5323 S(=O)(=O)(Nc1nc(OC)nc(OC)c1)c1ccc(N)cc1  
 515 5324 S(=O)(=O)(N=C(N)N)c1ccc(N)cc1  
 516 5325 S(=O)(=O)(Nc1nc(ccn1)C)c1ccc(N)cc1  
 517 5326 S(=O)(=O)(Nc1ncc(OC)cn1)c1ccc(N)cc1  
 518 5328 s1c(nnc1NS(=O)(=O)c1ccc(N)cc1)C  
 519 5329 S(=O)(=O)(Nc1noc(c1)C)c1ccc(N)cc1  
 520 5330 S(=O)(=O)(Nc1nnc(OC)cc1)c1ccc(N)cc1

521 5332 S(=O)(=O)(Nc1ncnc(OC)c1)c1ccc(N)cc1  
 522 5333 S(=O)(=O)(N)c1ccc(N)cc1  
 523 5335 S(=O)(=O)(Nc1n(ncc1)-c1ccccc1)c1ccc(N)cc1  
 524 5336 S(=O)(=O)(Nc1ncccc1)c1ccc(N)cc1  
 525 5340 s1ccnc1NS(=O)(=O)c1ccc(N)cc1  
 526 5342 S(=O)(CCC1C(=O)N(N(C1=O)c1ccccc1)c1ccccc1)c1ccccc1  
 527 5353 S(=O)(C)c1cc(OC)c(cc1)-c1[nH]c2ccnc2n1  
 528 5354 S(C(C)C)c1ccc(cc1)C(O)C(NCCCCCCCC)C  
 529 5355 S(=O)(=O)(N)c1cc(C(=O)NCC2N(CCC2)CC)c(OC)cc1  
 530 5359 s1cccc1C(=O)c1ccc(cc1)C(C(O)=O)C  
 531 5362 O=C1N(N(C(=O)C1(CCCC)COC(=O)CCC(O)=O)c1ccccc1)c1ccccc  
 1  
 532 5367 O(C)C1C(C2C(CC1OC(=O)c1cc(OC)c(OC(OCC)=O)c(OC)c1)CN1C  
 (C2)c2[nH]c3cc(OC)ccc3c2CC1)C(OC)=O  
 533 5387 s1cc2c(N(c3c(NC2=O)cccc3)C(=O)CN2CCN(CC2)C)c1C  
 534 5401 O1CCCC1C(=O)N1CCN(CC1)c1nc(N)c2cc(OC)c(OC)cc2n1  
 535 5403 Oc1cc(cc(O)c1)C(O)CNC(C)(C)C  
 536 5404 Clc1cc(Cl)ccc1C1(OC(CO1)CO)c1ccc(N2CCN(CC2)C(C)C)cc1)Cn1n  
 cnc1  
 537 5410 O(C(=O)CC)C1CCC2C3C(CCC12C)C1(C(=CC(=O)CC1)CC3)C  
 538 5419 N1CCN=C1C1CCCc2c1cccc2  
 539 5424 OC=1C(=O)C(O)=C(O)C(=O)C=1O  
 540 5426 O=C1NC(=O)CCC1N1C(=O)c2c(cccc2)C1=O

541 5429 O=C1NC(=O)N(c2nen(c12)C)C  
 542 5430 s1cc(nc1)-c1[nH]c2c(n1)cccc2  
 543 5433 ClC(Cl)C(=O)NC(C(O)c1ccc(S(=O)(=O)C)cc1)CO  
 544 5468 s1c(ccc1C(C(O)=O)C)C(=O)c1cccc1  
 545 5472 Clc1cccc1CN1CCc2sccc2C1  
 546 5479 S(=O)(=O)(CCn1c(ncc1[N+](=O)[O-])C)CC  
 547 5496 O1C(CO)C(O)C(N)C(O)C1OC1C(O)C(OC2OC(CN)C(O)CC2N)C(N  
 )CC1N  
 548 5501 O(CC)C(=O)NNc1nncc2c1cccc2  
 549 5503 S(=O)(=O)(NC(=O)NN1CCCCC1)c1ccc(cc1)C  
 550 5504 N1CCN=C1Cc1cccc1  
 551 5505 S(=O)(=O)(NC(=O)NCCCC)c1ccc(cc1)C  
 552 5507 Clc1ccc(Nc2cccc2C(O)=O)cc1C  
 553 5508 O=C(c1ccc(cc1)C)c1n(C)c(cc1)CC(=O)[O-]  
 554 5510 S=C(Oc1cc2c(cc1)cccc2)N(C)c1cc(ccc1)C  
 555 5526 OC(=O)C1CCC(CC1)CN  
 556 5530 NC1CC1c1cccc1  
 557 5534 N1(CCCC1)CC#CCN1CCCC1  
 558 5544 FC12C(C3CC(O)C(O)(C(=O)CO)C3(CC1O)C)CCC1=CC(=O)C=CC  
 12C  
 559 5546 n1c(N)c2nc(-c3cccc3)c(nc2nc1N)N  
 560 5560 Clc1cc2NC(NS(=O)(=O)c2cc1S(=O)(=O)N)C(Cl)Cl  
 561 5571 OC(=O)c1ccc[n+](c1)C

562 5572 OC(CCN1CCCCC1)(C1CCCCC1)c1ccccc1  
 563 5576 O1C(C)(C)C(=O)N(C)C1=O  
 564 5577 O(C)c1c(OC)cc(cc1OC)C(=O)NCc1ccc(OCCN(C)C)cc1  
 565 5585 O1c2c(cc3c(oc(c3)C)c2C)C(=CC1=O)C  
 566 5593 OCC(C(=O)N(Cc1ccncc1)CC)c1ccccc1  
 567 5635 O1C2C(C3N(CCC3=CC2OC)C)c2c(cc3OCOc3c2)C1=O  
 568 5645 OC1C2C3CCC(C(CCC(O)=O)C)C3(CCC2C2(C(C1)CC(O)CC2)C)C  
 569 5651 Clc1c2Oe3cc4C(NC(=O)C(NC(=O)C(NC(=O)C(NC)CC(C)C)C(O)c(
 c1)cc2)CC(=O)N)C(=O)NC1c2cc(-
 c5c(cc(O)cc5O)C(NC(=O)C(NC1=O)C(O)c1cc(Cl)c(Oc(c4)c3OC3O
 C(CO)C(O)C(O)C3OC3OC(C)C(O)C(N)(C3)C)cc1)C(O)=O)c(O)cc2  
 570 5665 OC(=O)CCC(N)C=C  
 571 5666 O1CCNCC1COc1ccccc1OCC  
 572 5668 OC1(n2c3C4N(CCCC4(C1)CC)CCc3c1c2ccccc1)C(OC)=O  
 573 5673 O(C(=O)C=1n2c3C4N(CCCC4(C=1)CC)CCc3c1c2ccccc1)CC  
 574 5707 S1CCCN=C1Nc1c(cccc1C)C  
 575 5803 Ic1cc(cc(I)c1Oe1cc(I)c(O)cc1)CC(O)=O  
 576 5850 [N+]1(CCCCC1)(CCCCC[N+])1(CCCCC1)C)C  
 577 5853 ClC(Cl)(Cl)C(P(OC)(OC)=O)O  
 578 5917 n12nnnc1CCCCC2  
 579 6077 S1c2c(N(c3c1ccccc3)CCCN(C)C)cc(cc2)C(=O)C  
 580 6082 O(C)c1ccc(OC)cc1C(O)C(N)C  
 581 6093 s1c([N+](=O)[O-])cnc1N1CCNC1=O

582 6103 Clc1cc2nc(oc2cc1)N  
 583 6127 [N+]1(CCC(CC1)=C(c1ccccc1)c1ccccc1)(C)C  
 584 6603 N1(C)C(CCCC1(C)C)(C)C  
 585 6634 Clc1nnc(NS(=O)(=O)c2ccc(N)cc2)cc1  
 586 6726 N1(CCN(CC1)C)C(c1ccccc1)c1ccccc1  
 587 6834 BrC1ccc(cc1)C(CCN(C)C)c1ncccc1  
 588 6890 [nH]1cc(c2c1cccc2)CN(C)C  
 589 7534 O(C(=O)C(O)c1ccccc1)C1CC(N(C)C(C1)C)(C)C  
 590 8249 N(=C(\N=C(N)N)/N)/CCc1ccccc1  
 591 8513 O=C1c2c(cccc2N)C(=O)c2c1cccc2N  
 592 8646 O=C1N=C(N=C2NNN=C12)N  
 593 9052 O1c2c(cccc2)C(=O)C(C(CC(=O)C)c2ccc([N+](=O)[O-])cc2)=C1O  
 594 9341 OC1C2N(CC1)CC=C2COC(=O)C(O)(C(C)C)C(OC)C  
 595 9363 O(C)c1cc(ccc1O)C(=O)N(CC)CC  
 596 9429 S1c2c(N(c3c1cccc3)CCCN1CCN(CC1)C)cc(S(=O)(=O)N(C)C)cc2  
 597 9458 FC(F)(F)c1cc(OC(=O)C)c(cc1)C(O)=O  
 598 9801 N(C(Cc1ccccc1)C)CCC(c1ccccc1)c1ccccc1  
 599 9985 N1CCCCC1CCC  
 600 10147 O1c2c3c(C4C(N(C3)C)c3c(CC4O)cc4OCOc4c3)ccc2OC1  
 601 10302 O(C)c1cc2c(cc1OC)CCNC2C  
 602 10428 O1CCC(O)(CC1=O)C  
 603 10666 O(C)C=1C=CN(C)C(=O)C=1C#N  
 604 10745 O(C(=O)c1ccccc1OC(=O)C)c1ccccc1C(O)=O

605	10767	<chem>O(C(=O)c1ccccc1)C(CC)(CN(C)C)C</chem>
606	11066	<chem>O1c2c(OC1)cc1C=3N(CCc1c2)C(=O)c1c(C=3)ccc(OC)c1OC</chem>
607	11096	<chem>O=C1CC2N(C(C1)CCC2)C</chem>
608	11289	<chem>ClC(=C(c1ccc(OC)cc1)c1ccc(OC)cc1)c1ccc(OC)cc1</chem>
609	12550	<chem>O1C(C)C(NC(=O)c2ccccc(NC=O)c2O)C(OC(C)C(OC(=O)CC(C)C)C (CCCCC)C1=O)=O</chem>
610	13729	<chem>S1C2N(C(=O)C2N)C(C(O)=O)=C(C1)COC(=O)C</chem>
611	13738	<chem>o1nc(nc1CCN(CC)CC)-c1ccccc1</chem>
612	14520	<chem>S1CCC(NC(=O)C)C1=O</chem>
613	15548	<chem>O(C)c1cc(ccc1OC)CC1N(CCc2c1cc(OC)c(OC)c2)C</chem>
614	16231	<chem>Clc1nc(C(=O)N=C(N)N)c(nc1N)N</chem>
615	16363	<chem>Fc1ccc(cc1)C(=O)CCCN1CCC(N2c3c(NC2=O)cccc3)CC1</chem>
616	18999	<chem>O(C)c1c(OC)c2C=3C(=CC(=O)C(O)=CC=3)C(N)CCc2cc1OC</chem>
617	19009	<chem>O(C)c1c2c(ccc1OC)cc-1[n+](CCc3cc(OC)c(OC)cc-13)c2</chem>
618	19659	<chem>S=C1N([O-])C=CC=C1</chem>
619	20710	<chem>O(C(=O)C(CO)c1ccccc1)C1CC2[N+](O-)(C(C1)CC2)C</chem>
620	21100	<chem>O(C)c1cc(ccc1O)C(O)CNC</chem>
621	21109	<chem>O(C)c1c(OC)c(OC)ccc1CN1CCNCC1</chem>
622	22407	<chem>O=C1N2C(C3CC(C2)CNC3)=CC=C1</chem>
623	22955	<chem>O(C)c1ccc2c(CN3C(C2)c2cc(O)c(OC)cc2CC3)c1O</chem>
624	23307	<chem>O(C)c1cc2c(cc1OC)cc1[n+](ccc3cc(OC)c(OC)cc13)c2C</chem>
625	23831	<chem>S(O)(=O)(=O)CCN1CCN(CC1)CCO</chem>
626	28061	<chem>ClCC(O)Cn1c(ncc1[N+](=O)[O-])C</chem>



627 31072 S=C1N(C=CN1C)C(OCC)=O  
 628 31729 Oc1ccc(cc1)C(O)C(NCCc1ccc(O)cc1)C  
 629 37338 [N+](CCCc1[n+](c2cc(N)ccc2c2c1cc(N)cc2)-c1ccccc1)(CC)(CC)C  
 630 40634 O1c2c(CCC1(C(O)=O)C)c(C)c(O)c(C)c2C  
 631 42616 O=C1C(N2CC2)=C(NC(OCC)=O)C(=O)C(N2CC2)=C1NC(OCC)=O  
 632 44097 S1C2N(C(=O)C2NC(=O)C(S(O)(=O)=O)c2ccccc2)C(C(=O)[O-])=C(C1)C[n+]1ccc(cc1)C(=O)N  
 633 48704 O(C)c1c2-c3c(CC4N(CCc(cc1OC)c24)C)ccc(OC)c3O  
 634 54456 Oc1cc2c(cc1O)CCNC2C  
 635 60793 O1C(CO)C(O)C(O)C(O)C1OC(C(O)C(O)C(O)=O)C(O)CO  
 636 62389 [nH]1c2c(ncnc2NCc2ccccc2)nc1  
 637 64961 [nH]1c2c(c3c1cncc3)cccc2  
 638 67425 O(CCN(CC)CC)c1c(OCCN(CC)CC)cccc1OCCN(CC)CC  
 639 68094 O=C1C=C2NC=3C(=C2C=C1)C=CNC=3C  
 640 68843 N1C(CCCC1C)C  
 641 69216 O=C1NC(=O)NC1C  
 642 71655 O1CCN(CC1)C(N=C(N)N)=N  
 643 71771 Clc1cccc(Cl)c1Nc1cccc1CC(OCC(O)=O)=O  
 644 91522 [nH]1c2c(CCNC2C)c2c1cccc2  
 645 92118 O(C)C1C2C34C(N(CC2(CCC3O)COC)CC)C1(O)C1(O)C2C4CC(C2OC)C(OC)C1  
 646 96946 OC(CC1N(C)C(CCC1)CC(O)c1ccccc1)c1ccccc1  
 647 97508 OC(=O)C(NC(=O)C)CC(O)=O

648 98889 OC=1C(=O)N(N(C)C=1C)c1cccc1

649 99114 O1C2C(C(C)C1=O)C(O)CC(=C1C2C(=CC1=O)C)C

650 107751 S(C(=O)C)CC(Cc1cccc1)C(=O)NCC(OCc1cccc1)=O

651 122642 P(OC1C(O)C(OC1n1c2ncnc(N)c2nc1)CO)(OCC1OC(n2c3ncnc(N)c3nc2)C(OP(OCC2OC(n3c4ncnc(N)c4nc3)C(O)C2O)(O)=O)C1O)(O)=O

652 161120 O1C2C(O)(C34OC5OC(=O)C(O)C56C3(C(OC4=O)C(O)C6C(C)(C)C)C2O)C(C)C1=O

653 165537 O1C23C4(OC1)C1C(O)(C56C(C(CCC5OC)(CN(C26)CC)C)C3OC(=O)C)CC(C1OC)C(OC)C4

654 170157 O(C)C1C2CC3C(C(O)(C1)C1C4N(CC5(C(C34C(O)CC5)C1)C)CC)C2O

655 170344 OC(=O)CN(CC(=O)Nc1c(ccc1CC)CC)CC(O)=O

656 179850 OC1CCc2c1nc1c(ccc1)c2N

657 180933 O=C1C=CC(=O)c2c1c1c(c3c(cc1)cccc3)cc2

658 195165 O1CCN(CC1)CCNc1nnc-2c(c1)CCc1c-2cccc1

659 201400 O(C(=O)C1(CCC(c2c1cccc2)C(OC1CC2[N+](C(C1)CC2)(CC)C)=O)c1cccc1)C1CC2[N+](C(C1)CC2)(CC)C

660 220774 OC1CCC2(C(CCC3(C2CC=C2C4C(C)C(CCC4(CCC23C)C(O)=O)C)C)C1(C)C)C

661 248507 O(C)c1c2-c3cc(OC)c(O)cc3CC3N(CCc(cc1O)c23)C

662 251561 O1C=C(C2C(CN3C(C2)c2[nH]c4c(c2CC3)cccc4)C1C)C(OC)=O

663 260807 O(Cc1cccc1)c1c2[nH]cc(c2ccc1)CN(C)C

664 264115 N1C2N(CCC3(C24CCN(C3Nc2c4cccc2)C)c2c1cccc2)C

665 267769 O1C23C(=CC1=O)C=CC(N1C2CCCC1)C3

666 271325 BrC1cc(OC)c(OC)cc1CC1N(CCCc2c1cc(OC)c(OC)c2)C

667 274159 O1CC(=CC1=O)C1CCC2(O)C3C(CCC12C)C1(C(CC(O)CC1)CC3)C

668 279057 O1C2CC3C(C4N(CC3(C4C23c2c(NC3=O)cccc2)C=C)C)C1

669 287691 S(C)C1=CC=C2C(=CC1=O)C(NC(=O)C)CCc1c2c(OC)c(OC)c(OC2OC(CO)C(O)C(O)C2O)c1

670 312827 O1C2C(C(C)C13NCC(CC3)C)C1(C(C3C(CC1)C1(C(CC(O)CC1)CC3)C)C2)C

671 342467 Ic1c(C(=O)NC2C(O)C(O)C(OC2O)CO)c(I)c(NC(=O)C)c(I)c1N(C(=O)C)C

672 413349 O1C(CN)C(O)C(O)C(N)C1OC1C(O)C(O)C(N)CC1N

673 418931 O1C(C)C(C)C(OC(=O)C)C(C)C(=O)C2(OC2)CC(C)C(OC2OC(CC(N(C)C)C2OC(=O)C)C)C(C)C(OC2OC(C)C(OC(=O)C)C(OC)C2)C(C)C1=O

674 420422 O(C)c1cc(ccc1OC)C(OC1CC2N(C(C1)CC2)C)=O

675 439739 OC1C2C(C3CCC(C(=O)CO)C3(C1)C)CCC1=CC(=O)CCC12C

676 442649 n1cc(ccc1)C1=NCCC1

677 443007 O(C)c1cc(ccc1O)C(OC1CC2N(C(C1)CC2)C)=O

678 446541 O1Cc2c(c(O)c(C\C=C(\CCC(O)=O)/C)c(OC)c2C)C1=O

679 457906 OC1CC([N+](C1)(C)C)C(=O)[O-]

680 500165 OC1CC2CCC3C4CCC(C(CCC(O)=O)C)C4(CCC3C2(CC1)C)C

681 521440 OC1(CCC2C3C(C4(C(=CC(=O)CC4)CC3)C)C(=O)CC12C)C(=O)C  
O

682 523975 O=C1N2C(=Nc3c1cccc3)CCCC2

683 542212 O1C(CO)C(O)C(OC(C(O)=O)C)C(NC(=O)C)C1O

684 580552 O1C2C(C3N(CCC3=CC2O)C)c2c(cc3OCOc3c2)C1=O

685 630532 O1C2CC(=O)N3C4C5(C6[N+])([O-  
])(CC(C(C24)C6)=CC1)CC5)c1c3cccc1

686 630921 OC1CCC2C(CC3N(C2)CCc2c3[nH]c3c2cccc3)C1C(O)=O

687 643764 O(C)c1c(OC)cc(cc1OC)\C=C/C(=O)N1CCC=CC1=O

688 657298 S=C1NC(=CC(=O)N1)CCC

689 657345 S(Cc1oc(cc1)CN(C)C)CCN\C(\NC)=C/[N+](=O)[O-]

690 667493 S=C1NC(=CC(=O)N1)C

691 688585 n1cccc1/C(=C\CN1CCCC1)/c1ccc(cc1)C

692 719408 s1ccnc1NC(=S)Nc1cccc1

693 1349907 S=C1NC=CN1C

694 1548885 S(=O)(C)c1ccc(cc1)\C=C/1\c2c(cc(F)cc2)C(CC(O)=O)=C\1C

695 1548912 O1c2cc(ccc2OC1)\C=C/C=C\C(=O)N1CCCCC1

696 1548942 O(C)c1cc(ccc1O)CNC(=O)CCCC\C=C/C(C)C

697 1548955 Cl\C(=C(\c1ccc(OCCN(CC)CC)cc1)/c1cccc1)\c1cccc1

698 2761171 S=C(N)c1cc(ncc1)CC

699 3032771 S(\C(=C/N(Cc1cnc(nc1N)C)C=O)\C)\CCOP(O)(O)=O)C(=O)c1cccc  
c1

700 3133561 OC1C23C(C45C6CC2C4N(CC6(CCC5O)C)CC)CC(O)C(C3)C1=C

701 3634067 O=CC=1C2CC3[N+](CC2=CC)(CCC23C=1Nc1c2cccc1)C

702 3649142 O=C1NC(CC(C)C)C(=O)NC(CCN)C(=O)NC(CCN)C(=O)NC(C(O)C)C(=O)NCCC(NC(=O)C(NC(=O)C(NC(=O)C(NC(=O)C(CCCC(CC)C)CCN)C(O)C)CCN)C(=O)NC(CCN)C(=O)NC1CC(C)C

703 3672427 [nH]1c2C3N(C4C5C(NCCC5)C3CC4)CCc2c2c1cccc2

704 3787925 OC1(CCC2C3C(C4C(CC3)=CCCC4)CCC12C)C#C

705 3851247 O=C1NC(CCCN)C(=O)NC(CC(C)C)C(=O)NC(Cc2cccc2)C(=O)N2C(CCC2)C(=O)NC(C(C)C)C(=O)NC(CCCN)C(=O)NC(CC(C)C)C(=O)NC(Cc2cccc2)C(=O)N2C(CCC2)C(=O)NC1C(C)C

706 3915039 O1C2C(OC3OC(CC(=O)C13O)C)C(O)C([NH2+]C)C(O)C2[NH2+]C

707 3996620 O=C1c2ccc(nc2N(C=C1C(=O)[O-])CC)C

708 4320774 O(CCCC)c1cc(ccc1N)C(OCC[NH+](CC)CC)=O

709 4359763 O1C23C4(OC1)C1C(C56C(C(CCC5OC)(CN(C26)CC)COC)C3O)CC(C1OC)C(OC)C4

710 4479096 O(CC(O)C[NH2+]C(C)(C)C)c1c2c(ccc1)C(=O)CCC2

711 4486617 O1C2C3N(C(CC(OC(=O)C(=C)c4cccc4)C3)C12)C

712 4580358 O1C(CC)C(O)(C2OC(NC(C2C)C(CC(O)(C)C(OC2OC(CC(N(C)C)C2O)C)C(C)C(OC2OC(C)C(O)C(OC)(C2)C)C(C)C1=O)C)COCCOC)C

713 4636599 S1C2N(C(C(=O)[O-])C1(C)C)C(=O)C2NC(=O)c1c2c(ccc1OCC)cccc2

714 4677798 O1C(CC(O)CC1=O)CCC1C2C(=CC(CC2OC(=O)C(CC)(C)C)C)C=CC1C

715 4739621 S(=O)([O-])(=Nc1nc(cc(n1)C)C)c1ccc(N)cc1

716 5053503 P(O)(=O)([O-])C(P(O)(=O)[O-])(O)C

717 5171637 O1C23C(CCC4C1(O)C(OC(=O)c1cc(OC)c(OC)cc1)CCC24C)C1(O)C(C2C(C(O)C1O)C(O)(C1N(CC(CC1)C)C2)C)C3

718 5191579 OC(C(CO)(C)C)C(=O)NCCC(=O)[O-]

719 5231296 Clc1cc(cc(Cl)c1N)C(O)C[NH2+]C(C)(C)C

720 5280442 O1c2c(C(=O)C=C1c1ccc(OC)cc1)c(O)cc(O)c2

721 5280443 O1c2c(C(=O)C=C1c1ccc(O)cc1)c(O)cc(O)c2

722 5280445 O1c2c(C(=O)C=C1c1cc(O)c(O)cc1)c(O)cc(O)c2

723 5280953 O(C)c1cc2[nH]c3c(c2cc1)ccnc3C

724 5281404 [nH]1c2c(c3c1cccc3)ccnc2C

725 5281672 O1c2c(C(=O)C(O)=C1c1cc(O)c(O)c(O)c1)c(O)cc(O)c2

726 5312135 S1(=O)(=O)N(C)\C(=C(/O)\Nc2ncccc2)\C(=O)c2c1cccc2

727 5312154 s1c2c(S(=O)(=O)N(C)\C(=C(/O)\Nc3ncccc3)\C2=O)cc1

728 5351819 OC12C(CC3C(C1=O)=C(O)c1c(cccc1O)C3(O)C)C(N(C)C)C(=O)/C(=C(\O)/NCN1CCCC1)/C2=O

729 5353527 OC1C\C(=C/C=C\2/C3CCC(C(CCCC(C)C)C)C3(CCC/2)C)\C(CC1)=C

730 5353656 O=C1C=C2NC=3C(=C2C=C1)CCNC=3C

731 5353779 OC12C(CC3C(C1=O)=C(O)c1c(C3)c(N(C)C)ccc1O)C(N(C)C)C(=O)/C(=C(\O)/N)/C2=O

732 5353864 O1C2C(CC\C(=C/CCC3(OC23)C)\C)C(=C)C1=O

733 5353990 OC12C(CC3C(C1=O)=C(O)c1c(cccc1O)C3(O)C)C(N(C)C)C(=O)/C(

=C(\O)/N)/C2=O  
734 5360959 S1(=O)(=O)N(C)\C(=C(/O)\Nc2noc(c2)C)\C(=O)c2c1cccc2  
735 5367858 Clc1c2c(C(O)=C3C(CC4C(O)(C(=O)\C(=C(/O)\N)\C(=O)C4N(C)C)C3=O)C2O)c(O)cc1  
736 5367871 Clc1c2c(C(O)=C3C(CC4C(O)(C(=O)\C(=C(/O)\N)\C(=O)C4N(C)C)C3=O)C2(O)C)c(O)cc1  
737 5368888 O1CC=2C3N(CCC3OC(=O)/C(/CC(=C)C(O)(C)C1=O)=C/C)CC=2  
738 5378180 O1C(C(O)C(O)C(O)C1CO)c1c2OC(=CC(=O)c2c(O)cc1O)c1ccc(O)c  
c1  
739 5458656 Brc1ccc(cc1)\C(=C/CN(C)C)\c1cccnc1  
740 5462906 O1CC=2C3N(CCC3OC(=O)/C(/CC(C)C(O)(CO)C1=O)=C/C)CC=2  
741 5473483 O1c2c(N=C3C1=CC(N(CC)CC)C=C3)c(cc(O)c2O)C(=O)N  
742 5474986 O1C(C)C(OC2OC(C)C(OC(=O)CC(C)C)C(O)(C2)C)C(N(C)C)C(O)C1OC1C(OC)C(OC(=O)C)CC(OC(C\C=C\C=C/C(O)C(CC1CC=O)C)C)=O  
743 5791942 Fc1ccc(cc1)-c1c2c(n(C(C)C)c1\C=C\C(O)CC(O)CC(=O)[O-])cccc2  
744 5820787 O(C(=O)C)C/1CC2(C(CC(O)C3C2(CCC2C(C)C(O)CCC23C)C)\C\1=C(\CCC=C(C)C)/C(=O)[O-])C  
745 6420076 N12C(=Nc3c(C1)cccc3)CCCC2  
746 6914279 s1ccc(C)c1\C=C/C1=NCCCN1C  
747 6914280 s1cccc1\C=C/C1=NCCCN1C  
748 6914283 Oc1cc(ccc1)\C=C/C1=NCCCN1C  
749 6921821 Clc1ccc(cc1)C(=O)c1n(C)c(cc1C)CC(=O)[O-]

750 9604989 S(=O)(=O)([O-])c1cc(S(=O)(=O)[O-])c2c(c1N)C(=O)/C(=N/Nc1ccc(cc1OC)-c1cc(OC)c(N\N=C\3/C=Cc4c(c(N)c(S(=O)(=O)[O-])cc4S(=O)(=O)[O-])C/3=O)cc1)/C=C2

751 10099444 O1C(=O)C(=CC1C)CCCCCCCCCCCC(O)C1OC(CC1)C(O)CCCCCCCCCCCC

752 14506494 FC12C(C3CC(C)C(OC(=O)C)(C(=O)CO)C3(CC1O)C)CCC1=CC(=O)C=CC12C

753 16219826 OC1CC(=O)C(C\C=C/C/C(C)C(O)=O)C1\C=C\C(O)CCCCC

754 24848418 O(C)c1cc2[NH2+]C3=C(CCN=C3C)c2cc1

---



### Appendix IV:

#### 5-HT<sub>2B</sub> External Set Compounds

<b>Cp.</b>	<b>PubChem_</b>	<b>SMILES</b>
<b>ID</b>	<b>CID</b>	
755	896	<chem>O(C)c1cc2c([nH]cc2CCNC(=O)C)cc1</chem>
756	3194	<chem>[Se]1N(C(=O)c2c1cccc2)c1cccc1</chem>
757	3410	<chem>O(C)c1ccc(cc1)CC(NCC(O)c1cc(NC=O)c(O)cc1)C</chem>
758	4452	<chem>Clc1cc(C(=O)NC2CCN(Cc3cccc3)C2C)c(OC)cc1NC</chem>
759	5454	<chem>S1c2c(cc(S(=O)(=O)N(C)C)cc2)C(c2c1cccc2)=CCCN1CCN(CC1)C</chem>
760	5574	<chem>S1c2c(N(c3c1cccc3)CC(CN(C)C)C)cccc2</chem>
761	12454	<chem>Clc1cc2c(Sc3c(cccc3)C2=CCCN2CCN(CC2)CCO)cc1</chem>
762	60854	<chem>Clc1cc2NC(=O)Cc2cc1CCN1CCN(CC1)c1nsc2c1cccc2</chem>
763	62875	<chem>Clc1cc2N(c3c(Sc2cc1)cccc3)CCCNC</chem>
764	122295	<chem>Clc1c2c(CCN(CC2)C)c(OCC=C(C)C)cc1</chem>
765	123836	<chem>O(C)c1cc2C3N(CC(CC(C)C)C(O)C3)CCc2cc1OC</chem>
766	152151	<chem>S(Oc1cc2c([nH]cc2CCN)cc1)(O)(=O)=O</chem>
767	5361110	<chem>Clc1cc(ccc1)C1=NCCN=C2N(NC(=C12)C)C</chem>
768	9543513	<chem>O=C(N1CCC(CC1)CCCCNC(=O)CCc1ccnc1)c1cccc1</chem>
769	9952054	<chem>Ic1ccc(cc1)C1CC2N(C(CC2)C1C(OC)=O)C</chem>
770	24901748	<chem>O(C)c1ccc(cc1)C(n1ccnc1)COCCc1ccc(OC)cc1</chem>

### Appendix V:

Predictions on 5-HT<sub>2B</sub> Test Set Compounds for a Simple Model Generated Using the  
Number of Nitrogen Atoms, Number of Hydrophobic Groups, And LogP.

<b>Cp. ID</b>	<b>Model</b>	<b>Actual Class</b>	<b>Predicted Class</b>
1	Simple	Active	Inactive
5	Simple	Active	Active
11	Simple	Active	Inactive
15	Simple	Active	Inactive
18	Simple	Active	Inactive
21	Simple	Active	Inactive
26	Simple	Active	Inactive
27	Simple	Active	Inactive
28	Simple	Active	Active
33	Simple	Active	Inactive
37	Simple	Active	Active
44	Simple	Active	Active
49	Simple	Active	Inactive
59	Simple	Active	Inactive
60	Simple	Active	Inactive
64	Simple	Active	Active
65	Simple	Active	Inactive
72	Simple	Active	Inactive

73	Simple	Active	Active
74	Simple	Active	Inactive
82	Simple	Active	Active
84	Simple	Active	Inactive
85	Simple	Active	Inactive
86	Simple	Active	Inactive
106	Simple	Active	Inactive
107	Simple	Active	Active
124	Simple	Active	Inactive
129	Simple	Active	Inactive
134	Simple	Active	Active
136	Simple	Active	Inactive
151	Simple	Inactive	Inactive
158	Simple	Inactive	Inactive
163	Simple	Inactive	Inactive
168	Simple	Inactive	Inactive
170	Simple	Inactive	Inactive
179	Simple	Inactive	Inactive
188	Simple	Inactive	Active
193	Simple	Inactive	Inactive
195	Simple	Inactive	Inactive
201	Simple	Inactive	Inactive
203	Simple	Inactive	Inactive

208	Simple	Inactive	Active
210	Simple	Inactive	Active
218	Simple	Inactive	Inactive
221	Simple	Inactive	Active
223	Simple	Inactive	Inactive
232	Simple	Inactive	Inactive
233	Simple	Inactive	Inactive
238	Simple	Inactive	Inactive
250	Simple	Inactive	Inactive
251	Simple	Inactive	Inactive
254	Simple	Inactive	Inactive
261	Simple	Inactive	Inactive
266	Simple	Inactive	Inactive
267	Simple	Inactive	Inactive
268	Simple	Inactive	Inactive
270	Simple	Inactive	Inactive
274	Simple	Inactive	Inactive
283	Simple	Inactive	Active
287	Simple	Inactive	Active
292	Simple	Inactive	Inactive
293	Simple	Inactive	Inactive
296	Simple	Inactive	Inactive
300	Simple	Inactive	Inactive

304	Simple	Inactive	Active
311	Simple	Inactive	No Prediction
327	Simple	Inactive	Inactive
345	Simple	Inactive	Inactive
346	Simple	Inactive	Active
354	Simple	Inactive	Inactive
359	Simple	Inactive	Inactive
364	Simple	Inactive	No Prediction
369	Simple	Inactive	Inactive
370	Simple	Inactive	Active
374	Simple	Inactive	Inactive
375	Simple	Inactive	Inactive
380	Simple	Inactive	Inactive
381	Simple	Inactive	Inactive
385	Simple	Inactive	Inactive
399	Simple	Inactive	Inactive
408	Simple	Inactive	Inactive
413	Simple	Inactive	Inactive
414	Simple	Inactive	Active
427	Simple	Inactive	Inactive
431	Simple	Inactive	Inactive
432	Simple	Inactive	Inactive
440	Simple	Inactive	Inactive

441	Simple	Inactive	Inactive
442	Simple	Inactive	Inactive
443	Simple	Inactive	Inactive
450	Simple	Inactive	No Prediction
455	Simple	Inactive	Active
457	Simple	Inactive	Active
473	Simple	Inactive	Inactive
478	Simple	Inactive	Inactive
480	Simple	Inactive	Inactive
488	Simple	Inactive	Inactive
513	Simple	Inactive	Inactive
520	Simple	Inactive	Inactive
523	Simple	Inactive	Active
526	Simple	Inactive	Inactive
532	Simple	Inactive	Inactive
538	Simple	Inactive	Inactive
548	Simple	Inactive	Inactive
555	Simple	Inactive	Inactive
556	Simple	Inactive	Inactive
562	Simple	Inactive	Inactive
566	Simple	Inactive	Inactive
568	Simple	Inactive	Inactive
571	Simple	Inactive	Active

576	Simple	Inactive	Inactive
577	Simple	Inactive	Inactive
583	Simple	Inactive	Inactive
584	Simple	Inactive	Inactive
604	Simple	Inactive	Inactive
605	Simple	Inactive	Inactive
606	Simple	Inactive	Inactive
610	Simple	Inactive	Inactive
611	Simple	Inactive	Inactive
618	Simple	Inactive	Inactive
621	Simple	Inactive	Active
626	Simple	Inactive	Inactive
628	Simple	Inactive	Inactive
629	Simple	Inactive	Inactive
630	Simple	Inactive	Inactive
638	Simple	Inactive	Inactive
639	Simple	Inactive	Inactive
654	Simple	Inactive	Inactive
655	Simple	Inactive	Inactive
657	Simple	Inactive	Inactive
659	Simple	Inactive	Active
672	Simple	Inactive	Inactive
681	Simple	Inactive	Inactive

688	Simple	Inactive	Inactive
698	Simple	Inactive	Inactive
702	Simple	Inactive	Inactive
703	Simple	Inactive	Active
711	Simple	Inactive	Inactive
717	Simple	Inactive	Active
721	Simple	Inactive	Inactive
723	Simple	Inactive	Inactive
725	Simple	Inactive	Active
733	Simple	Inactive	Inactive
736	Simple	Inactive	Inactive
737	Simple	Inactive	No Prediction
741	Simple	Inactive	Active
742	Simple	Inactive	Inactive
751	Simple	Inactive	Inactive
753	Simple	Inactive	Inactive
754	Simple	Inactive	Inactive

---

**TP=9; FP= 20; TN=96; FN=21;  $CCR_{evs} = 0.55$**

---



## Appendix VI:

The Purity Data for Ten Virtual Screening Compounds Tested In 5-HT<sub>2B</sub> Study

Compd.* ID	PDSP ID	PubChem CID	Purity %	Method
<b>1</b>	14809	43922	99%	LC/MS Spectra
<b>2</b>	14807	71928	96%	LC/MS Spectra
<b>3</b>	14806	114709	98%	LC/MS Spectra
<b>4</b>	14814	3038495	100%	LC/MS Spectra
<b>5</b>	14821	3336	96%	LC/MS Spectra
<b>6</b>	14815	1715104	100%	LC/MS Spectra
<b>7</b>	27769	4140	98%	LC/MS Spectra
<b>8</b>	14805	195658	99%	LC/MS Spectra
<b>9</b>	13513	9909648	70%	LC/MS Spectra
<b>10</b>	13505	15940170	98%	LC/MS Spectra

\*Compound IDs as reported in chapter 4.

## **Appendix VII:**

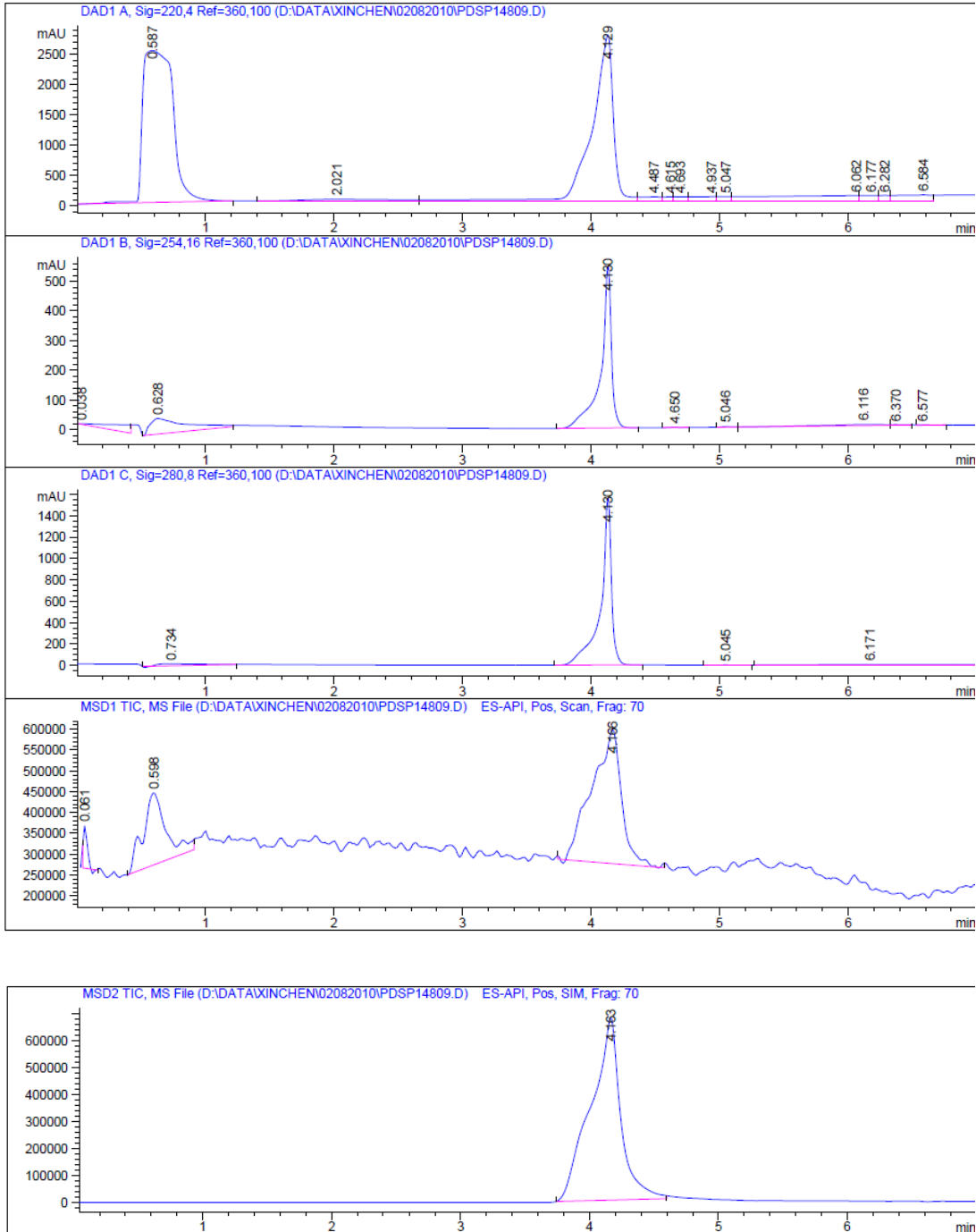
LC/MS Purity Spectra for Tested Virtual Screening Hits

# LC/MS purity spectra for compound 1.

```

Acq. Operator   : xin                               Seq. Line :    6
Acq. Instrument : Instrument 1                       Location  : PI-B-07
Injection Date  : 2/8/2010 3:45:15 PM                Inj       :    1
                                                    Inj Volume: 2 µl

Different Inj Volume from Sequence !   Actual Inj Volume : 20 µl
Acq. Method    : C:\CHEM32\1\METHODS\A_GENERAL METHOD LC
Last changed   : 1/7/2010 12:44:44 PM by xin
Analysis Method : C:\CHEM32\1\METHODS\A_GENERAL METHOD LCMS.M
Last changed   : 1/29/2010 3:48:53 PM by Feng
                (modified after loading)
    
```



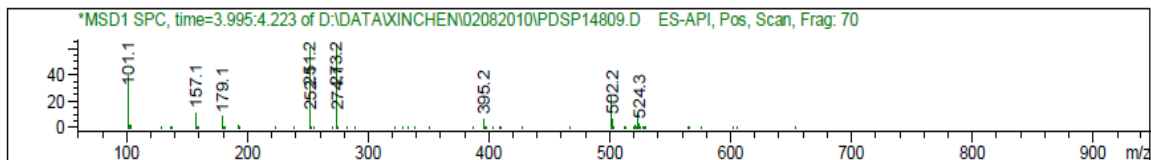
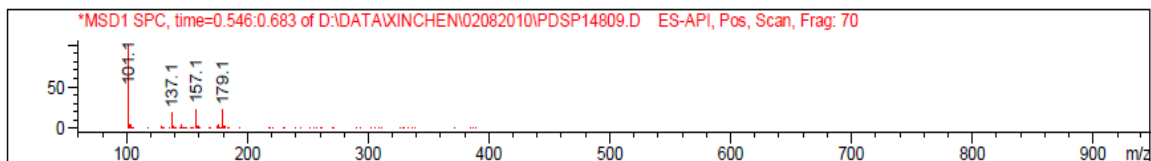
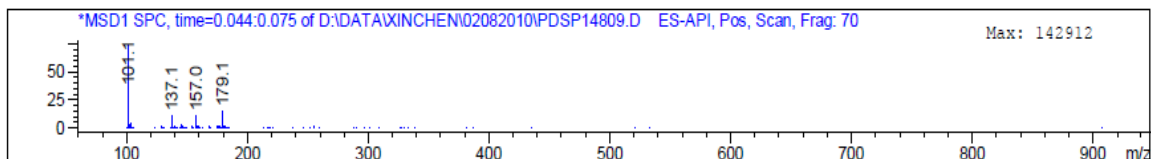
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70

Spectra averaged over upper half of peaks.

Noise Cutoff: 1000 counts.

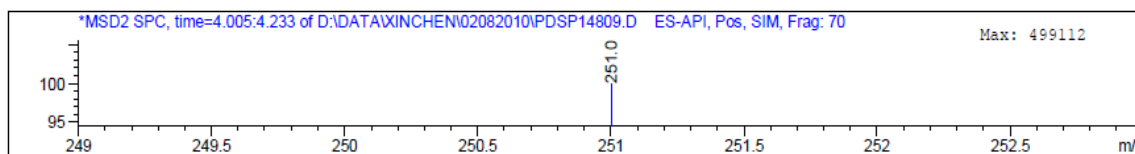
Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.061	207179	179.05 I
		157.20 I
		157.00 I
		137.05 I
		101.10 I
0.598	2139745	179.10 I
		157.10 I
		137.10 I
		101.10 I
4.166	5177026	523.25 I
		502.25 I
		501.30 I
		274.15 I
		273.20 I
		252.15 I
		251.20 I
		179.10 I
		157.10 I
		101.10 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
Spectra averaged over upper half of peaks.  
Noise Cutoff: 1000 counts.  
Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.163	11043291	251.00 I

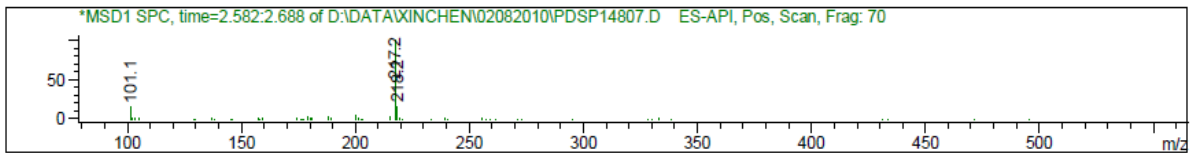
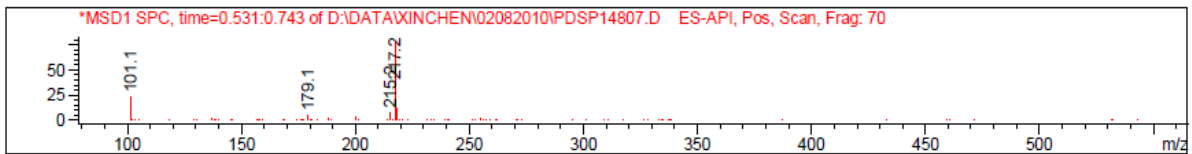
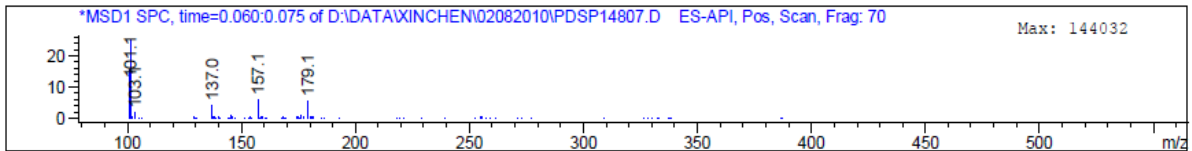


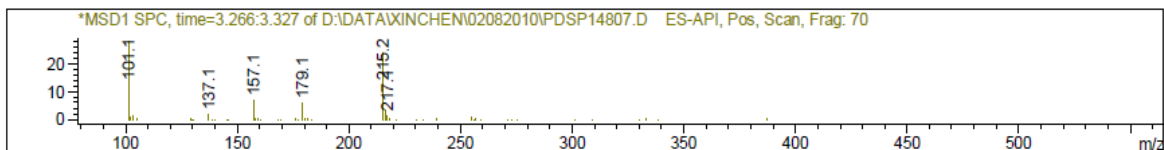
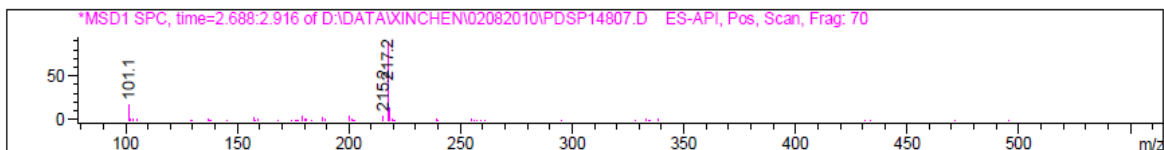
\*\*\* End of Report \*\*\*



MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.067	368918	179.05 I 157.05 I 137.00 I 101.10 I 100.80 I
0.601	10773873	218.20 I 217.20 I 101.10 I
2.635	3978179	218.20 I 217.20 I 101.10 I
2.743	7097169	218.20 I 217.20 I 101.10 I
3.298	786204	216.15 I 215.20 I 179.05 I 157.10 I 101.10 I





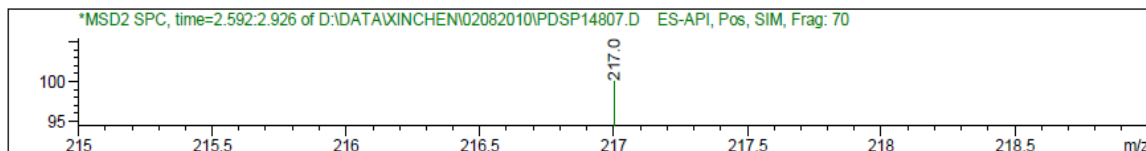
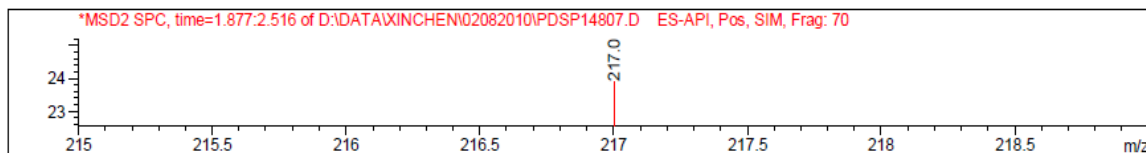
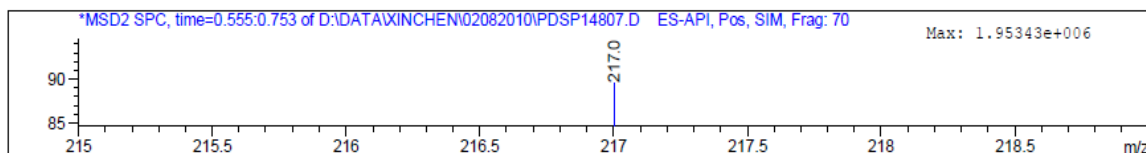
MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70

Spectra averaged over upper half of peaks.

Noise Cutoff: 1000 counts.

Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.608	30095968	217.00 I
2.340	14513706	217.00 I
2.659	51246800	217.00 I



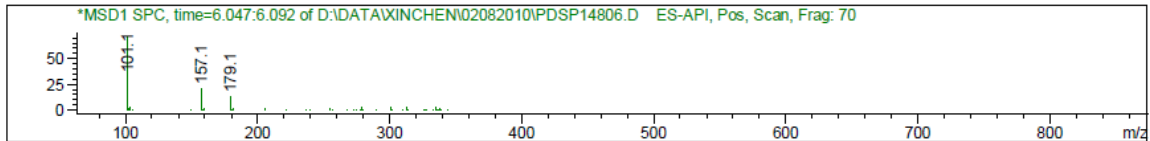
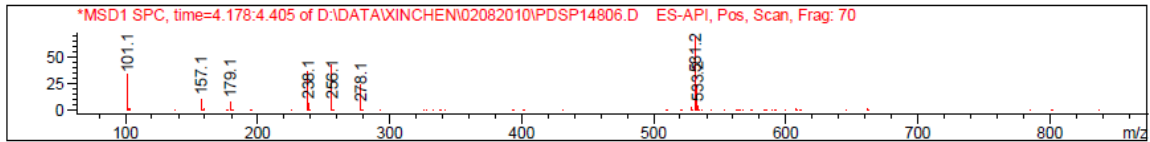
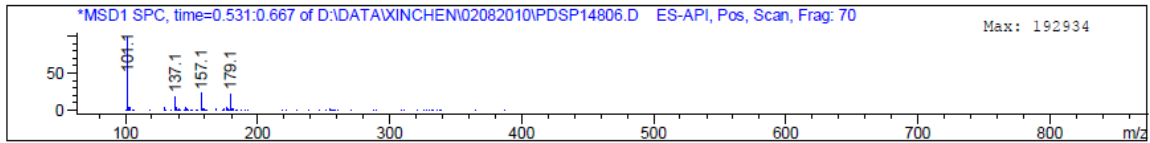
\*\*\* End of Report \*\*\*





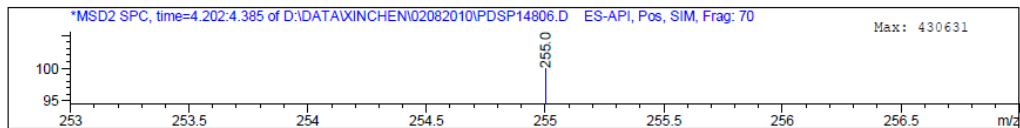
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.590	2207809	179.10 I 157.10 I 137.10 I 101.10 I
4.321	5158950	532.20 I 531.20 I 277.15 I 255.15 I 237.20 I 179.05 I 157.10 I 101.10 I
6.070	166676	179.10 I 157.10 I 101.10 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.321	7351510	255.00 I

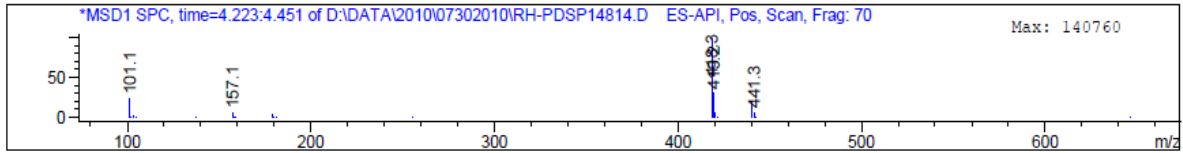


\*\*\* End of Report \*\*\*



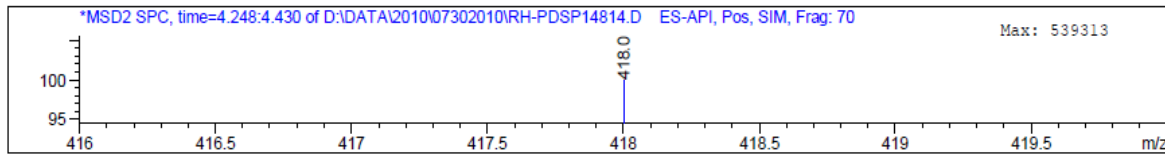
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
Spectra averaged over upper half of peaks.  
Noise Cutoff: 1000 counts.  
Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.309	4761690	440.25 I
		419.25 I
		418.30 I
		101.10 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
Spectra averaged over upper half of peaks.  
Noise Cutoff: 1000 counts.  
Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.285	8484095	418.00 I

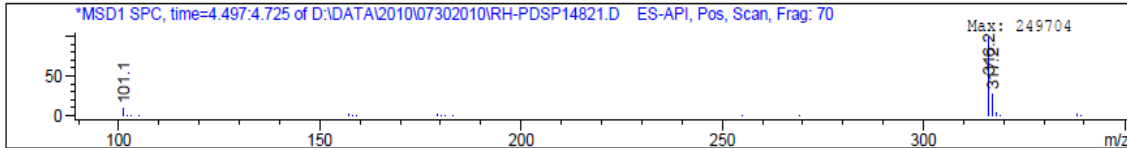


\*\*\* End of Report \*\*\*



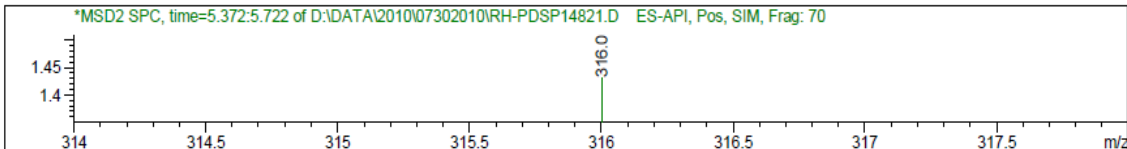
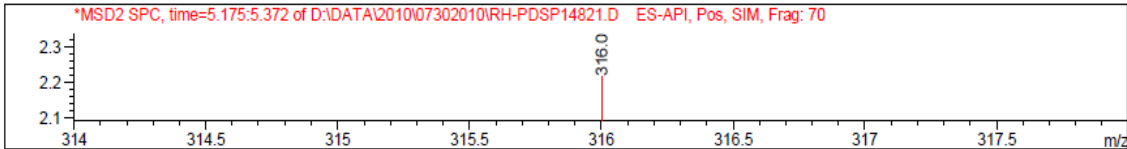
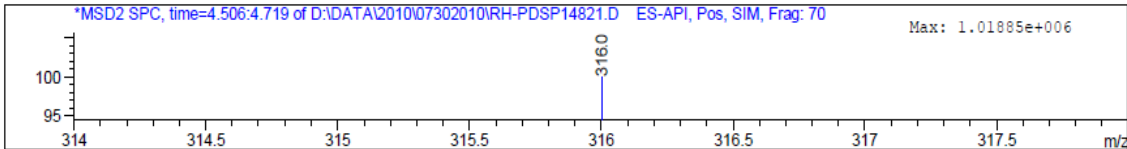
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.546	6187094	317.20 I 316.20 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.556	17161350	316.00 I
5.213	224484	316.00 I
5.405	262216	316.00 I

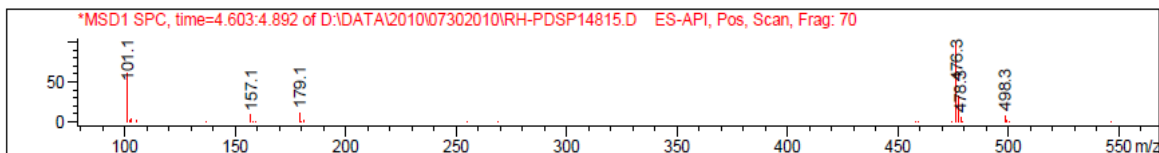
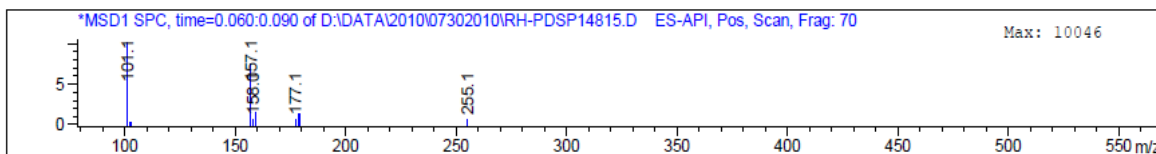


\*\*\* End of Report \*\*\*



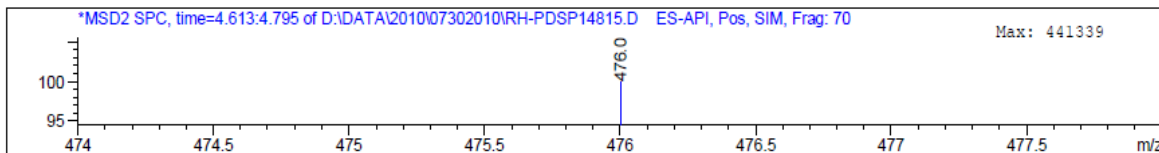
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.068	24197	179.15 I 178.90 I 159.10 I 157.05 I 156.80 I 101.10 I
4.662	4808172	477.30 I 476.30 I 179.10 I 101.10 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.661	6711315	476.00 I



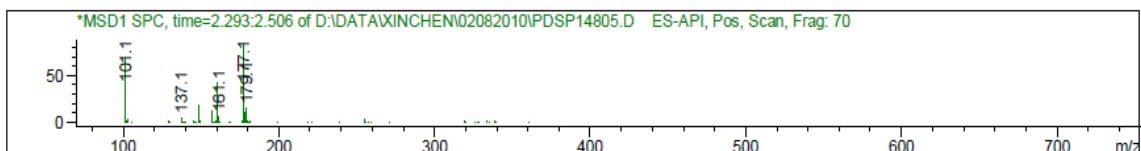
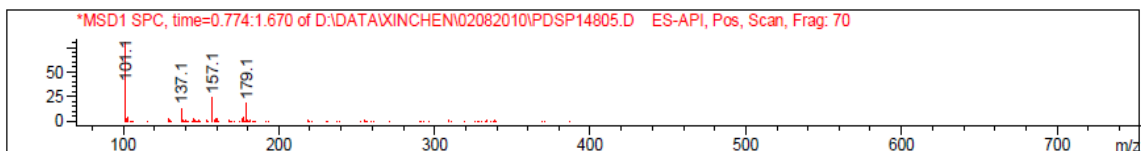
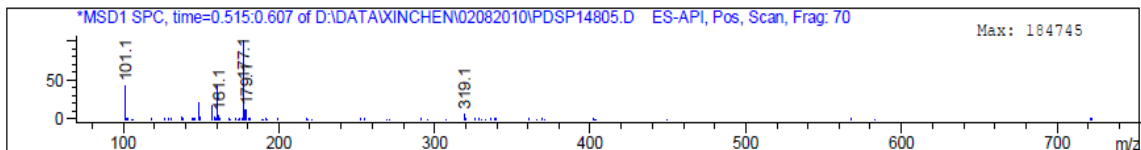
\*\*\* End of Report \*\*\*





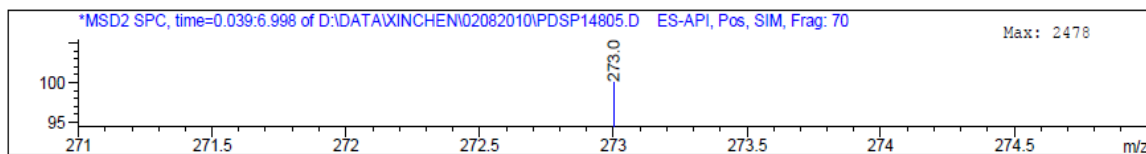
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
0.553	2679534	179.10 I
		178.15 I
		177.10 I
		160.20 I
		157.10 I
		148.20 I
		101.10 I
0.985	3490388	179.10 I
		157.10 I
		137.05 I
		101.10 I
2.346	3688703	179.05 I
		178.10 I
		177.15 I
		160.15 I
		157.10 I
		148.20 I
		101.10 I



MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
Spectra averaged over upper half of peaks.  
Noise Cutoff: 1000 counts.  
Reportable Ion Abundance: > 10%.

No peaks to report



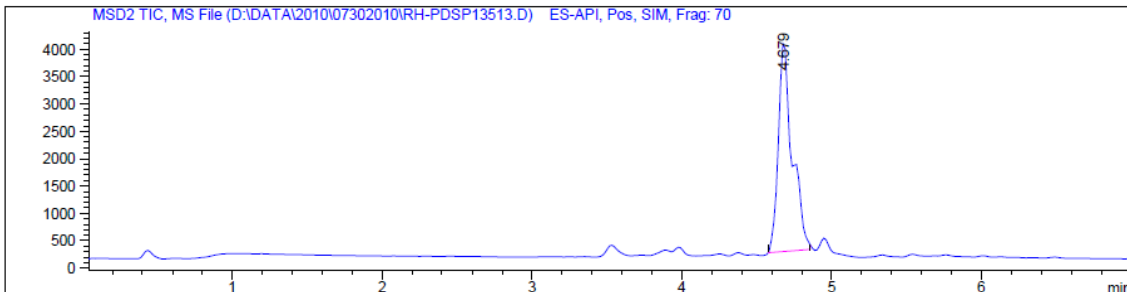
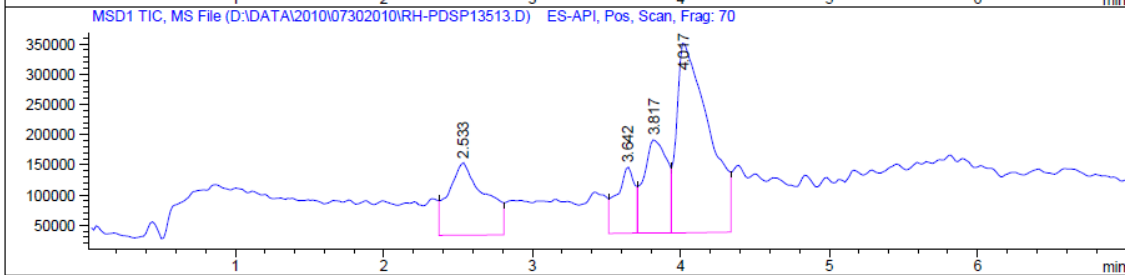
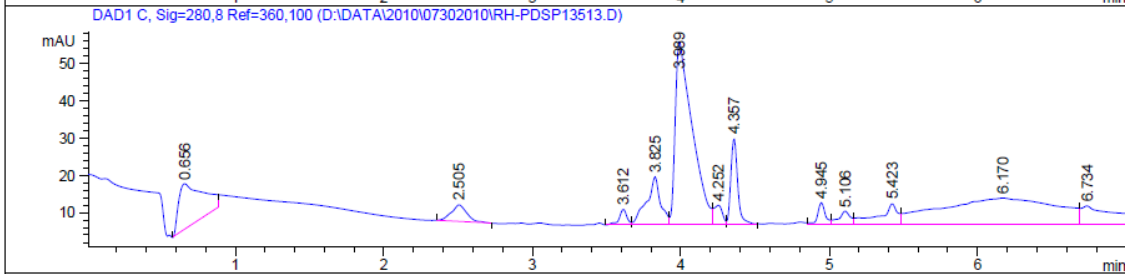
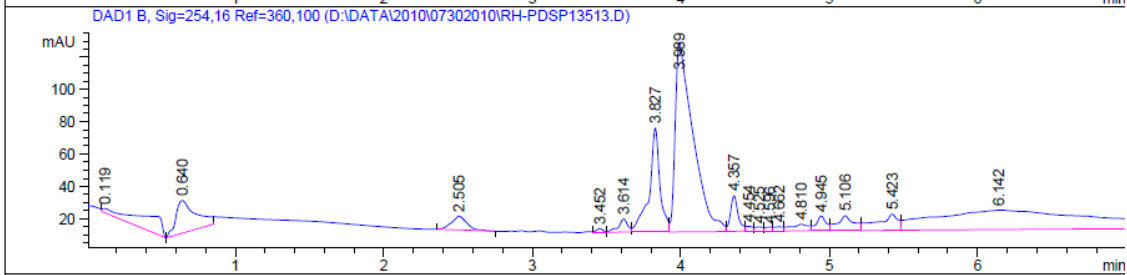
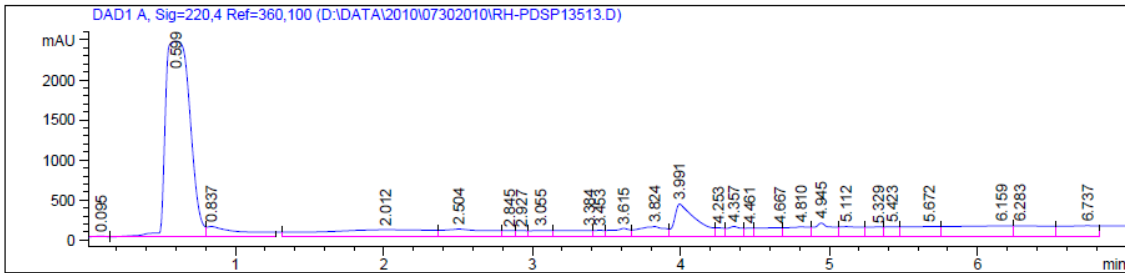
\*\*\* End of Report \*\*\*

# LC/MS purity spectra for compound 9.

```

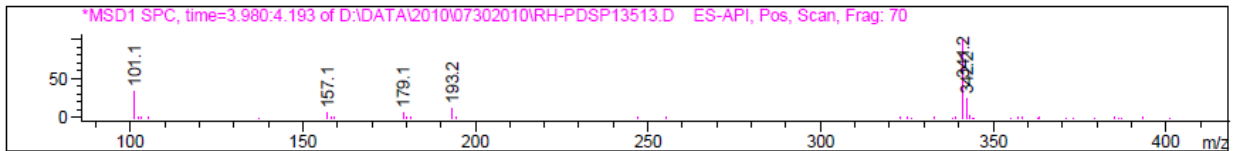
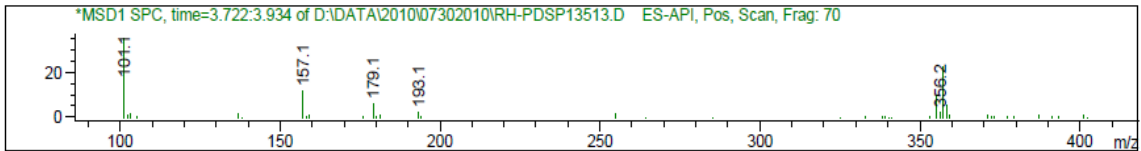
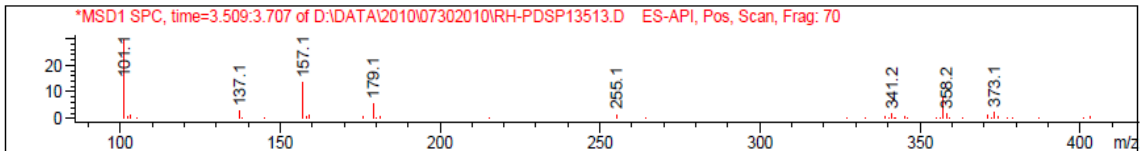
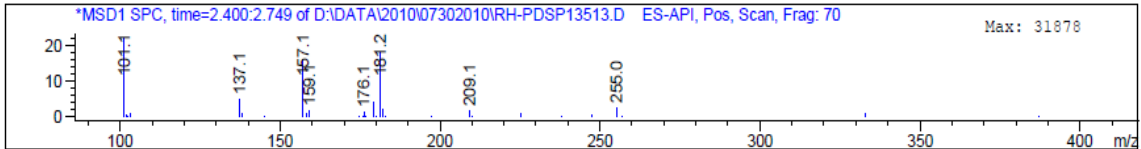
Acq. Operator   : Martin                               Seq. Line : 19
Acq. Instrument : Instrument 1                         Location  : P1-D-06
Injection Date  : 7/30/2010 6:55:55 PM                Inj       : 1
                                                    Inj Volume: 2 µl

Different Inj Volume from Sequence !   Actual Inj Volume : 8 µl
Acq. Method    : C:\CHEM32\1\METHODS\A_GENERAL METHOD LC
Last changed   : 4/14/2010 8:10:26 AM by xin
Analysis Method: C:\CHEM32\1\METHODS\A_NON-POLAR METHOD LCMS.M
Last changed   : 1/7/2010 12:47:56 PM by Martin
    
```



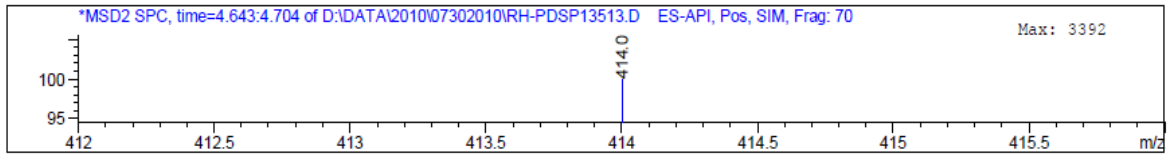
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70  
 Spectra averaged over upper half of peaks.  
 Noise Cutoff: 1000 counts.  
 Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
2.533	2075419	255.00 I
		181.20 I
		179.10 I
		157.10 I
		137.10 I
		101.10 I
		101.10 I
3.642	928127	357.20 I
		179.05 I
		157.10 I
		101.10 I
3.817	1677161	358.20 I
		357.20 I
		355.15 I
		179.10 I
		157.10 I
		101.10 I
		101.10 I
4.017	4750399	342.20 I
		341.20 I
		193.20 I
		101.10 I



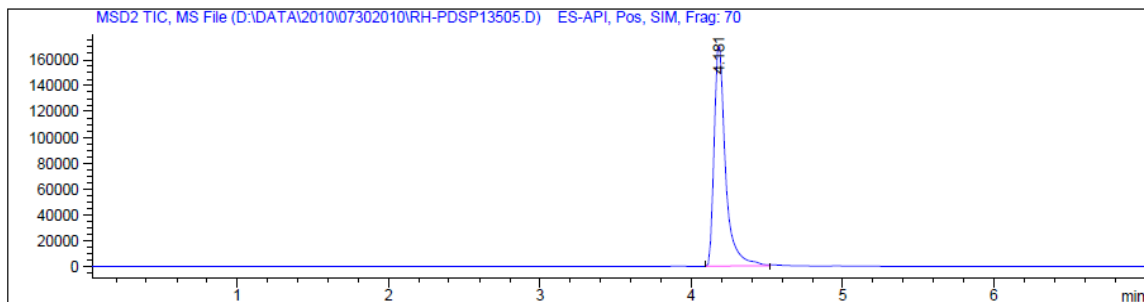
MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70  
Spectra averaged over upper half of peaks.  
Noise Cutoff: 1000 counts.  
Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.679	24564	414.00 I



\*\*\* End of Report \*\*\*





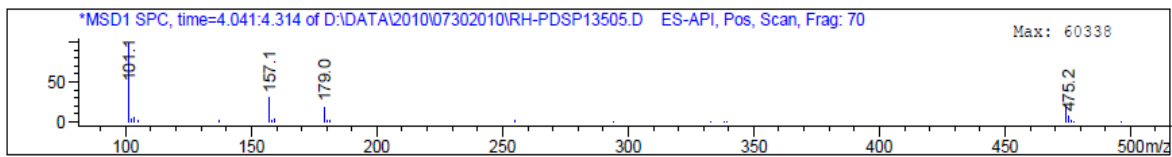
MS Signal: MSD1 TIC, MS File, ES-API, Pos, Scan, Frag: 70

Spectra averaged over upper half of peaks.

Noise Cutoff: 1000 counts.

Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.175	1547804	474.20 I
		179.00 I
		157.10 I
		101.10 I



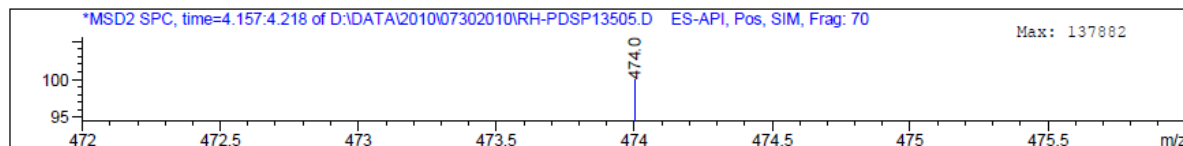
MS Signal: MSD2 TIC, MS File, ES-API, Pos, SIM, Frag: 70

Spectra averaged over upper half of peaks.

Noise Cutoff: 1000 counts.

Reportable Ion Abundance: > 10%.

Retention Time (MS)	MS Area	Mol. Weight or Ion
4.181	895445	474.00 I



\*\*\* End of Report \*\*\*



## REFERENCES

1. Daylight, World Drug Index (WDI). 2004.  
Ref Type: Generic
2. Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules, VLDB-94.
3. Alizadeh F, Goldfarb D. 2003. Second-order cone programming. *Mathematical Programming* 95(1):3-51
4. Altar CA, Vawter MP, Ginsberg SD. 2009. Target identification for CNS diseases by transcriptional profiling. *Neuropsychopharmacology* 34(1):18-54
5. Armbruster BN, Roth BL. 2005. Mining the receptorome. *J. Biol. Chem.* 280(7):5129-32
6. ASNN. ASNN. 2010.  
Ref Type: Computer Program
7. Asthana S, Brinton RD, Henderson VW, McEwen BS, Morrison JH, Schmidt PJ. 2009. Frontiers proposal. National Institute on Aging "bench to bedside: estrogen as a case study". *Age (Dordr. )* 31(3):199-210
8. Bajorath J. 2002. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1(11):882-94
9. Bajorath J. 2005. Molecular Similarity Methods and QSAR Models as Tools for Virtual Screening. In *Drug Discovery Handbook*, ed. SC Gad, 3:87-122. Hoboken, New Jersey: Wiley.
10. Baker NC, Hemminger BM. 2010. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *J. Biomed. Inform.*
11. Bakker BM, Michels PAM, Opperdoes FR, Westerhoff HV. 1999. What controls glycolysis in bloodstream form *Trypanosoma brucei*? *Journal of Biological Chemistry* 274(21):14551-9

12. Balaban AT. 1979. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta.*(53):355-75
13. Balaban AT. 1982. Highly Discriminating Distance-Based Topological Index. *Chemical Physics Letters* 89(5):399-404
14. Barrett T, Edgar R. 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.* 411:352-69
15. Becker OM. 2004. Structure based GPCR drug discovery: From the computer to the clinic. *Abstracts of Papers of the American Chemical Society* 227:U913
16. Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A et al. 2004. G protein-coupled receptors: In silico drug discovery in 3D. *Proceedings of the National Academy of Sciences of the United States of America* 101(31):11304-9
17. Benfenati E, Benigni R, DeMarini DM, Helma C, Kirkland D et al. 2009. Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-of-the-Art, and Perspectives. *Journal of Environmental Science and Health Part C- Environmental Carcinogenesis & Ecotoxicology Reviews* 27(2):57-90
18. Benmansour S, Piotrowski JP, Altamirano AV, Frazer A. 2009. Impact of Ovarian Hormones on the Modulation of the Serotonin Transporter by Fluvoxamine. *Neuropsychopharmacology* 34(3):555-64
19. Berger M, Gray JA, Roth BL. 2009. The Expanded Biology of Serotonin. *Annual Review of Medicine* 60:355-66
20. Bonchev D, Mekenyan O, Trinajstić N. 1981. Isomer Discrimination by Topological Information Approach. *Journal of Computational Chemistry* 2(2):127-48
21. Brady GP, Stouten PFW. 2000. Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design* 14(4):383-401
22. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P et al. 2001. Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics* 29(4):365-71

23. Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*, Wadsworth.
24. Brenner C. 2004. Chemical genomics in yeast. *Genome Biol.* 5(9):240
25. Burgun A, Bodenreider O. 2008. Accessing and integrating data and knowledge for biomedical research. *Yearb. Med. Inform.*:91-101
26. Burks C, Cassidy M, Cinkosky MJ, Cumella KE, Gilna P et al. 1991. GenBank. *Nucleic Acids Res.* 19 Suppl:2221-5
27. Burks C, Cinkosky MJ, Gilna P, Hayden JE, Abe Y et al. 1990. GenBank: current status and future directions. *Methods Enzymol.* 183:3-22
28. Campbell SJ, Gaulton A, Marshall J, Bichko D, Martin S et al. 2010. Visualizing the drug target landscape. *Drug Discov. Today* 15(1-2):3-15
29. Caruana R. 1997. Multitask learning. *Machine Learning* 28(1):41-75
30. Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, Lee PWN. 2002. Metabolic control analysis in drug discovery and disease. *Nature Biotechnology* 20(3):243-9
31. Cavalli A, Bolognesi ML, Minarini A, Rosini M, Tumiatti V et al. 2008. Multi-target-directed ligands to combat neurodegenerative diseases. *J. Med. Chem.* 51(3):347-72
32. Chang C-C, Lin C-J. LIBSVM. 2001.  
Ref Type: Computer Program
33. Chekmarev DS, Kholodovych V, Balakin KV, Ivanenkov Y, Ekins S, Welsh WJ. 2008. Shape signatures: New descriptors for predicting cardiotoxicity in silico. *Chemical Research in Toxicology* 21(6):1304-14
34. ChemAxon JChem. ChemAxon JChem. ChemAxon JChem . 2010.  
Ref Type: Electronic Citation

35. ChEMBL. ChEMBL. European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) . 2010.  
Ref Type: Electronic Citation
36. Chen B, Wild D, Guha R. 2009. PubChem as a source of polypharmacology. *J. Chem. Inf. Model.* 49(9):2044-55
37. Cheng J, Sun S, Tracy A, Hubbell E, Morris J et al. 2004. NetAffx gene ontology mining tool: A visual approach for microarray data analysis. *Bioinformatics* 20(9):1462-3
38. Connolly HM, Crary JL, Mcgoon MD, Hensrud DD, Edwards BS et al. 1997. Valvular heart disease associated with fenfluramine-phentermine. *New England Journal of Medicine* 337(9):581-8
39. Connor SC, Hansen MK, Corner A, Smith RF, Ryan TE. 2010. Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes. *Mol. Biosyst.* 6(5):909-21
40. Cortes C, Vapnik V. 1995. Support-Vector Networks. *Machine Learning* 20(3):273-97
41. Cristianini N, Shawe-Taylor J. 2000. *An introduction to support vector machines*, Cambridge, United Kingdom: Cambridge University Press.
42. Darvas F, Dorman G, Krajcsi P, Puskas LG, Kovari Z et al. 2004. Recent advances in chemical genomics. *Curr. Med. Chem.* 11(23):3119-45
43. Davis AP, Murphy CG, Saraceni-Richards CA, Rosenstein MC, Wiegers TC, Mattingly CJ. 2009. Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res.* 37(Database issue):D786-D792
44. Daylight. World Drug Index (WDI). 2004.  
Ref Type: Generic
45. de Cerqueira LP, Golbraikh A, Oloff S, Xiao Y, Tropsha A. 2006. Combinatorial QSAR modeling of P-glycoprotein substrates. *J. Chem. Inf. Model.* 46(3):1245-54

46. DeRisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(5338):680-6
47. Discovery Studio. Discovery Studio. Accelrys [2.1]. 2008. Accelrys.  
Ref Type: Electronic Citation
48. Doraiswamy PM, Krishnan KR, Oxman T, Jenkyn LR, Coffey DJ et al. 2003. Does antidepressant therapy improve cognition in elderly depressed patients? *J Gerontol. A Biol. Sci. Med. Sci.* 58(12):M1137-M1144
49. Dragon. Talete s.r.l.Dragon. [5.4.2006]. 2007. Milan, Italy.  
Ref Type: Computer Program
50. Droogmans S, Cosyns B, D'haenen H, Creeten E, Weytjens C et al. 2007. Possible association between 3,4-methylenedioxymethamphetamine abuse and valvular heart disease. *Am. J. Cardiol.* 100(9):1442-5
51. Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1):207-10
52. Ekins S, Crumb WJ, Sarazan RD, Wikel JH, Wrighton SA. 2002. Three-dimensional quantitative structure-activity relationship for inhibition of human ether-a-go-go-related gene potassium channel. *Journal of Pharmacology and Experimental Therapeutics* 301(2):427-34
53. Evers A, Klebe G. 2004. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *Journal of Medicinal Chemistry* 47(22):5381-92
54. Fan Y, Shi LM, Kohn KW, Pommier Y, Weinstein JN. 2001. Quantitative structure-antitumor activity relationships of camptothecin analogues: cluster analysis and genetic algorithm-based studies. *J. Med. Chem.* 44(20):3254-63
55. FDA. Evista Approved for Reducing Breast Cancer Risk. <http://www.fda.gov/> . 2007. U.S. Food and Drug Administration.  
Ref Type: Electronic Citation

56. Fitzgerald LW, Burn TC, Brown BS, Patterson JP, Corjay MH et al. 2000. Possible role of valvular serotonin 5-HT(2B) receptors in the cardiopathy associated with fenfluramine. *Mol. Pharmacol.* 57(1):75-81
57. Fourches D, Muratov E, Tropsha A. 2010. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem. Inf. Model.*
58. Frantz S. 2005. Drug discovery: playing dirty. *Nature* 437(7061):942-3
59. Freyhult E, Prusis P, Lapinsh M, Wikberg JE, Moulton V, Gustafsson MG. 2005. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. *Bmc Bioinformatics* 6
60. Frye CA. 2009. Progestogens influence cognitive processes in aging. *Future Medicinal Chemistry* 1(7):1215-31
61. Garg D, Gandhi T, Mohan CG. 2008. Exploring QSTR and toxicophore of hERG K<sup>+</sup> channel blockers using GFA and HypoGen techniques. *Journal of Molecular Graphics & Modelling* 26(6):966-76
62. Garman KS, Acharya CR, Edelman E, Grade M, Gaedcke J et al. 2008. A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities. *Proc. Natl. Acad. Sci. U S A* 105(49):19432-7
63. Geldenhuys WJ, Van der Schyf CJ. 2009. The serotonin 5-HT<sub>6</sub> receptor: a viable drug target for treating cognitive deficits in Alzheimer's disease. *Expert Rev. Neurother.* 9(7):1073-85
64. Girones X, Gallegos A, Carbo-Dorca R. 2000. Modeling antimalarial activity: application of Kinetic Energy Density Quantum Similarity Measures as descriptors in QSAR. *J. Chem. Inf. Comput. Sci.* 40(6):1400-7
65. Gloriam DE, Fredriksson R, Schioth HB. 2007. The G protein-coupled receptor subset of the rat genome. *Bmc Genomics* 8
66. Golbraikh A, Shen M, Tropsha A. 2002. Enrichment: A new estimator of classification accuracy of QSAR models. *Abstracts of Papers of the American Chemical Society* 223:U494-U495

67. Golbraikh A, Shen M, Xiao Z, Xiao YD, Lee KH, Tropsha A. 2003. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des* 17(2-4):241-53
68. Golbraikh A, Tropsha A. 2002a. Beware of q<sup>2</sup>! *J. Mol. Graph. Model.* 20(4):269-76
69. Golbraikh A, Tropsha A. 2002b. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* 5(4):231-43
70. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439):531-7
71. Hall LH, Kier LB. 1990. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quantitative Structure-Activity Relationships* 9(2):115-31
72. Hall LH, Mohny B, Kier LB. 1991a. The Electrotopological State - An Atom Index for Qsar. *Quantitative Structure-Activity Relationships* 10(1):43-51
73. Hall LH, Mohny B, Kier LB. 1991b. The Electrotopological State - Structure Information at the Atomic Level for Molecular Graphs. *Journal of Chemical Information and Computer Sciences* 31(1):76-82
74. Hamadeh HK, Todd M, Healy L, Meyer JT, Kwok AM et al. 2010. Application of genomics for identification of systemic toxicity triggers associated with VEGF-R inhibitors. *Chem. Res. Toxicol.* 23(6):1025-33
75. Hampton T. 2004. "Promiscuous" anticancer drugs that hit multiple targets may thwart resistance. *Jama-Journal of the American Medical Association* 292(4):419-22
76. Hart SA, Snyder MA, Smejkalova T, Woolley CS. 2007. Estrogen mobilizes a subset of estrogen receptor-alpha-immunoreactive vesicles in inhibitory presynaptic boutons in hippocampal CA1. *Journal of Neuroscience* 27(8):2102-11

77. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R et al. 2008. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. *Blood* 111(12):5654-62
78. Hata R, Masumura M, Akatsu H, Li F, Fujita H et al. 2001. Up-regulation of calcineurin Abeta mRNA in the Alzheimer's disease brain: assessment by cDNA microarray. *Biochem. Biophys. Res. Commun.* 284(2):310-6
79. Hawkins DM, Basak SC, Mills D. 2003. Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences* 43(2):579-86
80. Henderson VW. 2009. Estrogens, episodic memory, and Alzheimer's disease: a critical update. *Semin. Reprod. Med.* 27(3):283-93
81. Hieronymus H, Lamb J, Ross KN, Peng XP, Clement C et al. 2006. Gene expression signature-based chemical genomic prediction identifies a novel class of HSP90 pathway modulators. *Cancer Cell* 10(4):321-30
82. Holenz J, Pauwels PJ, Diaz JL, Merce R, Codony X, Buschmann H. 2006. Medicinal chemistry strategies to 5-HT(6) receptor ligands as potential cognitive enhancers and antiobesity agents. *Drug Discov. Today* 11(7-8):283-99
83. Hollander M, Wolfe D. 1999. *Nonparametric Statistical Methods*, pp. 178-185. New York: Wiley.
84. Holliday JD, Hu CY, Willett P. 2002. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screen.* 5(2):155-66
85. Hood L, Perlmutter RM. 2004. The impact of systems approaches on biological problems in drug discovery. *Nature Biotechnology* 22(10):1215-7
86. Hopkins AL. 2007. Network pharmacology. *Nature Biotechnology* 25(10):1110-1
87. Hopkins AL. 2008. Network pharmacology: the next paradigm in drug discovery. *Nature Chemical Biology* 4(11):682-90
88. Hopkins AL. 2009. Drug discovery: Predicting promiscuity. *Nature* 462(7270):167-8



89. Hopkins AL, Mason JS, Overington JP. 2006. Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology* 16(1):127-36
90. Horwood JM, Dufour F, Laroche S, Davis S. 2006. Signalling mechanisms mediated by the phosphoinositide 3-kinase/Akt cascade in synaptic plasticity and memory in the rat. *Eur. J Neurosci.* 23(12):3375-84
91. Hsieh JH, Wang XS, Teotico D, Golbraikh A, Tropsha A. 2008. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *Journal of Computer-Aided Molecular Design* 22(9):593-609
92. Huan J, Prins J, Wang W. 2003. Efficient Mining of Frequent Subgraph in the Presence of Isomorphism. *Proc. of the 3rd IEEE International Conference on Data Mining (ICDM)*:549-52
93. Huang XP, Setola V, Yadav PN, Allen JA, Rogan SC et al. 2009. Parallel functional activity profiling reveals valvulopathogens are potent 5-hydroxytryptamine(2B) receptor agonists: implications for drug safety assessment. *Molecular Pharmacology* 76(4):710-22
94. Hwang J, Zheng LT, Ock J, Lee MG, Kim SH et al. 2008. Inhibition of glial inflammatory activation and neurotoxicity by tricyclic antidepressants. *Neuropharmacology* 55(5):826-34
95. JChem. Standardizer. ChemAxon JChem (<http://www.chemaxon.com>) . 2009. Ref Type: Electronic Citation
96. Jeon KI, Xu X, Aizawa T, Lim JH, Jono H et al. 2010. Vinpocetine inhibits NF-kappaB-dependent inflammation via an IKK-dependent but PDE-independent mechanism. *Proc. Natl. Acad. Sci. U S A* 107(21):9795-800
97. Kandpal R, Saviola B, Felton J. 2009. The era of 'omics unlimited. *Biotechniques* 46(5):351-5
98. Keith CT, Borisy AA, Stockwell BR. 2005. Multicomponent therapeutics for networked systems. *Nat. Rev. Drug Discov.* 4(1):71-8

99. Kellogg GE, Kier LB, Gaillard P, Hall LH. 1996. E-state fields: Applications to 3D QSAR. *Journal of Computer-Aided Molecular Design* 10(6):513-20
100. Khalifa AE. 2003. Zuclopenthixol facilitates memory retrieval in rats: possible involvement of noradrenergic and serotonergic mechanisms. *Pharmacol. Biochem. Behav.* 75(4):755-62
101. Khashan R, Zheng WF, Huan J, Wang W, Tropsha A. 2005. Development of fragment-based chemical descriptors using novel frequent common subgraph mining approach and their application in QSAR modeling. *Abstracts of Papers of the American Chemical Society* 230:U1335-U1336
102. Kier LB. 1985. A Shape Index from Molecular Graphs. *Quantitative Structure-Activity Relationships* 4(3):109-16
103. Kier LB. 1987. Inclusion of Symmetry As A Shape Attribute in Kappa-Index Analysis. *Quantitative Structure-Activity Relationships* 6(1):8-12
104. Kier LB, Hall LH. 1976. *Molecular connectivity in chemistry and drug research*, New York: Academic Press.
105. Kier LB, Hall LH. 1986. *Molecular connectivity in structure-activity analysis*, New York: Wiley.
106. Kier LB, Hall LH. 1991. A Differential Molecular Connectivity Index. *Quantitative Structure-Activity Relationships* 10(2):134-40
107. Kier LB, Hall LH. 1999. *Molecular structure description: The electrotopological state*, New York: Academic Press.
108. Kiessling LL, Splain RA. 2010. Chemical approaches to glycobiology. *Annu. Rev. Biochem.* 79:619-53
109. Kinase SARfari. Kinase SARfari. The European Bioinformatics Institute (EMBL-EBI) . 2010.  
Ref Type: Electronic Citation

110. Kirkpatrick S, Gelatt CD, Vecchi MP. 1983. Optimization by Simulated Annealing. *Science* 220(4598):671-80
111. Kitchen DB, Decornez H, Furr JR, Bajorath J. 2004. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Reviews Drug Discovery* 3(11):935-49
112. Klabunde T, Giegerich C, Evers A. 2009. Sequence-Derived Three-Dimensional Pharmacophore Models for G-Protein-Coupled Receptors and Their Application in Virtual Screening. *Journal of Medicinal Chemistry* 52(9):2923-32
113. Klien LA. 1999. *Sensor and Data Fusion Concepts and Applications*, Bellingham, WA: SPIE.
114. Kohavi RA. 1995. Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection., pp. 1137-1143.
115. Kortagere S, Ekins S. 2010. Troubleshooting computational methods in drug discovery. *J. Pharmacol. Toxicol. Methods* 61(2):67-75
116. Kovatcheva A, Golbraikh A, Oloff S, Xiao YD, Zheng W et al. 2004. Combinatorial QSAR of ambergris fragrance compounds. *J. Chem. Inf. Comput. Sci.* 44(2):582-95
117. Kuhn M, Szklarczyk D, Franceschini A, Campillos M, von MC et al. 2009. STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.*
118. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. 2008. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research* 36:D684-D688
119. Lamb J. 2007. Innovation - The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer* 7(1):54-60
120. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC et al. 2006. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313(5795):1929-35

121. Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV et al. 2003. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell* 114(3):323-34
122. Lapinsh M, Prusis P, Lundstedt T, Wikberg JES. 2002. Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. *Molecular Pharmacology* 61(6):1465-75
123. Le-Niculescu H, McFarland MJ, Mamidipalli S, Ogden CA, Kuczynski R et al. 2007. Convergent Functional Genomics of bipolar disorder: from animal model pharmacogenomics to human genetics and biomarkers. *Neurosci. Biobehav. Rev.* 31(6):897-903
124. Ledoux VA, Smejkalova T, May RM, Cooke BM, Woolley CS. 2009. Estradiol Facilitates the Release of Neuropeptide Y to Suppress Hippocampus-Dependent Seizures. *Journal of Neuroscience* 29(5):1457-68
125. Levy RJ. 2006. Serotonin transporter mechanisms and cardiac disease. *Circulation* 113(1):2-4
126. Littleton-Kearney MT, Ostrowski NL, Cox DA, Rossberg MI, Hurn PD. 2002. Selective estrogen receptor modulators: tissue actions and potential for CNS protection. *CNS Drug Rev.* 8(3):309-30
127. Liu B, Hsu W, Ma Y. 1998. Integrating classification and association rule mining, pp. 80-68. New York: Fourth International conference on Knowledge Discovery and Data Mining (KDD-98, Plenary Presentation).
128. Liu B, Hsu W, Ma Y. 1999. Pruning and summarizing the discovered associations, San Diego, CA, USA.: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99, full paper).
129. Liu B, Hsu W, Ma Y. Classification Based on Association (CBA). [v2.1]. 2001. School of Computing, National University of Singapore.  
Ref Type: Computer Program
130. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J et al. 2003. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res.* 31(1):82-6

131. Marron JS. MATLAB software for smoothing, functional data analysis and distance weighted discrimination. 2002.  
Ref Type: Computer Program
132. Marron JS, Todd MJ, Ahn J. 2007. Distance-weighted discrimination. *Journal of the American Statistical Association* 102(480):1267-71
133. Mathworks. Matlab. 2010.  
Ref Type: Computer Program
134. MDL Ltd. MACCS. 1992. San Leandro, CA, MDL Ltd.  
Ref Type: Computer Program
135. Medina AE. 2010. Vinpocetine as a potent antiinflammatory agent. *Proc. Natl. Acad. Sci. U S A* 107(22):9921-2
136. Mencher SK, Wang LG. 2005. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin. Pharmacol.* 5(1):3
137. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH. 1953. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* 21:1087-92
138. MOE. MOE. Chemical Computing Group [2007.09 ]. 2008.  
Ref Type: Electronic Citation
139. Mohan CG, Gandhi T, Garg D, Shinde R. 2007. Computer-assisted methods in chemical toxicity prediction. *Mini-Reviews in Medicinal Chemistry* 7(5):499-507
140. MolconnZ. MolconnZ. <http://www.edusoft-lc.com/molconn/> . 2006.  
Ref Type: Electronic Citation
141. Morello KC, Wurz GT, DeGregorio MW. 2003. Pharmacokinetics of selective estrogen receptor modulators. *Clinical Pharmacokinetics* 42(4):361-72
142. Nebigil CG, Choi DS, Dierich A, Hickel P, Le Meur M et al. 2000a. Serotonin 2B receptor is required for heart development. *Proc. Natl. Acad. Sci. U. S A* 97(17):9508-13

143. Nebigil CG, Launay JM, Hickel P, Tournois C, Maroteaux L. 2000b. 5-hydroxytryptamine 2B receptor regulates cell-cycle progression: cross-talk with tyrosine kinase pathways. *Proc. Natl. Acad. Sci. U. S A* 97(6):2591-6
144. Newman-Tancredi A, Cussac D, Quentric Y, Touzard M, Verrielle L et al. 2002. Differential actions of antiparkinson agents at multiple classes of monoaminergic receptor. III. Agonist and antagonist properties at serotonin, 5-HT(1) and 5-HT(2), receptor subtypes. *J. Pharmacol. Exp. Ther.* 303(2):815-22
145. NIA.NIH. 2010. *Can Alzheimer's disease be prevented?* The National Institute on Aging (NIA),
146. Nicholson JK, Holmes E, Lindon JC, Wilson ID. 2004. The challenges of modeling mammalian biocomplexity. *Nat. Biotechnol.* 22(10):1268-74
147. Nislow C, Giaever G. 2003. "Chemogenomics: tools for protein families" and "Chemical genomics: chemical and biological integration". *Pharmacogenomics* 4(1):15-8
148. O'Neill K, Chen S, Diaz BR. 2004. Impact of the selective estrogen receptor modulator, tamoxifen, on neuronal outgrowth and survival following toxic insults associated with aging and Alzheimer's disease. *Exp. Neurol.* 188(2):268-78
149. Ogden CA, Rich ME, Schork NJ, Paulus MP, Geyer MA et al. 2004. Candidate genes, pathways and mechanisms for bipolar (manic-depressive) and related disorders: an expanded convergent functional genomics approach. *Mol. Psychiatry* 9(11):1007-29
150. Ogorevc J, Dovc P, Kunej T. 2010. Comparative Genomics Approach to Identify Candidate Genetic Loci for Male Fertility. *Reprod. Domest. Anim*
151. Olah M, Rad R, Ostopovici L, Bora A, Hadaruga N et al. 2007. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, ed. SL Schreiber, TM Kapoor, G Wess, 760-786. **New York: Wiley-VCH.**

152. Oloff S, Mailman RB, Tropsha A. 2005. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J Med. Chem.* 48(23):7322-32
153. Ong J, Kerr DI. 2005. Clinical potential of GABAB receptor modulators. *CNS Drug Rev.* 11(3):317-34
154. Ong J, Parker DA, Marino V, Kerr DI, Puspawati NM, Prager RH. 2005. 3-Chloro,4-methoxyfendiline is a potent GABA(B) receptor potentiator in rat neocortical slices. *Eur. J Pharmacol.* 507(1-3):35-42
155. Oprea T, Tropsha A. Target, Chemical and Bioactivity Databases – Integration is Key. *Drug Discov. Today* 3, 357-365. 2006.  
Ref Type: Journal (Full)
156. Palfreyman MG, Hook DJ, Klimczak LJ, Brockman JA, Evans DM, Altar CA. 2002. Novel directions in antipsychotic target identification using gene arrays. *Curr. Drug Targets CNS Neurol. Disord.* 1(2):227-38
157. Papa E, Pilutti P, Gramatica P. 2008. Prediction of PAH mutagenicity in human cells by QSAR classification. *Sar and Qsar in Environmental Research* 19(1-2):115-27
158. PDSP. PDSP Ki Database. <http://pdsp.med.unc.edu> . 2009.  
Ref Type: Electronic Citation
159. Peralta C, Wolf E, Alber H, Seppi K, Muller S et al. 2006. Valvular heart disease in Parkinson's disease vs. controls: An echocardiographic study. *Movement Disorders* 21(8):1109-13
160. Perou CM, Sorlie T, Eisen MB, van de RM, Jeffrey SS et al. 2000. Molecular portraits of human breast tumours. *Nature* 406(6797):747-52
161. Petak I, Schwab R, Orfi L, Kopper L, Keri G. 2010. Integrating molecular diagnostics into anticancer drug discovery. *Nat. Rev. Drug Discov.*
162. Peterson YK, Wang XS, Casey PJ, Tropsha A. 2009. Discovery of Geranylgeranyltransferase-I Inhibitors with Novel Scaffolds by the Means of

Quantitative Structure-Activity Relationship Modeling, Virtual Screening, and Experimental Validation. *Journal of Medicinal Chemistry* 52(14):4210-20

163. Petitjean M. 1992. Applications of the Radius Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical-Compounds. *Journal of Chemical Information and Computer Sciences* 32(4):331-7
164. Pezet S, Marchand F, D'Mello R, Grist J, Clark AK et al. 2008. Phosphatidylinositol 3-kinase is a key mediator of central sensitization in painful inflammatory conditions. *J Neurosci.* 28(16):4261-70
165. PFEIFER A, SCHRATTENHOLZ A, MUHS A. 2009. *USA Patent No. WO 2009/004038 A2*
166. Pinkerton JV, Henderson VW. 2005. Estrogen and cognition, with a focus on Alzheimer's disease. *Semin. Reprod. Med.* 23(2):172-9
167. Platt JR. 1947. Influence of neighbor bonds on additive bond properties in paraffins. *J. Chem. Phys.* 15:419-20
168. Polychronakos C. 2008. New applications of microarray data analysis: integrating genetics with 'Omics'. Organized by the Cambridge Healthtech Institute, 15-17 August 2007, Washington DC, USA. *Pharmacogenomics* 9(1):15-7
169. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M et al. 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415(6870):436-42
170. Pritchett AM, Morrison JF, Edwards WD, Schaff HV, Connolly HM, Espinosa RE. 2002. Valvular heart disease in patients taking pergolide. *Mayo Clinic Proceedings* 77(12):1280-6
171. PubChem. PubChem. NIH's Molecular Libraries Roadmap Initiative . 2009. Ref Type: Electronic Citation
172. Quinlan JR. C4.5: program for machine learning. 1992. Morgan Kaufmann. Ref Type: Computer Program



173. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A et al. 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19(1):45-50
174. Randic M. 1975. Characterization of Molecular Branching. *Journal of the American Chemical Society* 97(23):6609-15
175. Randic M, Basak SC. 2000. Construction of high-quality structure-property-activity regressions: the boiling points of sulfides. *J. Chem. Inf. Comput. Sci.* 40(4):899-905
176. Rapp SR, Espeland MA, Shumaker SA, Henderson VW, Brunner RL et al. 2003. Effect of estrogen plus progestin on global cognitive function in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA* 289(20):2663-72
177. Ricciarelli R, d'Abramo C, Massone S, Marinari U, Pronzato M, Tabaton M. 2004. Microarray analysis in Alzheimer's disease and normal aging. *IUBMB Life* 56(6):349-54
178. Riedel RF, Porrello A, Pontzer E, Chenette EJ, Hsu DS et al. 2008. A genomic approach to identify molecular pathways associated with chemotherapy resistance. *Mol. Cancer Ther.* 7(10):3141-9
179. Rosenbaum DM, Rasmussen SGF, Kobilka BK. 2009. The structure and function of G-protein-coupled receptors. *Nature* 459(7245):356-63
180. Roth BL. 2005. Receptor systems: Will mining the receptorome yield novel targets for pharmacotherapy? *Pharmacology & Therapeutics* 108(1):59-64
181. Roth BL. 2007. Drugs and valvular heart disease. *N. Engl. J. Med.* 356(1):6-9
182. Roth BL, Baner K, Westkaemper R, Siebert D, Rice KC et al. 2002. Salvinorin A: A potent naturally occurring nonnitrogenous kappa opioid selective agonist. *Proceedings of the National Academy of Sciences of the United States of America* 99(18):11934-9
183. Roth BL, Craigo SC, Choudhary MS, Uluer A, Monsma FJ et al. 1994. Binding of Typical and Atypical Antipsychotic Agents to 5-Hydroxytryptamine-6 and 5-

Hydroxytryptamine-7 Receptors. *Journal of Pharmacology and Experimental Therapeutics* 268(3):1403-10

184. Roth BL, Kroeze WK. 2006. Screening the receptorome yields validated molecular targets for drug discovery. *Curr. Pharm. Des* 12(14):1785-95
185. Roth BL, Lopez E, Patel S, Kroeze WK. 2000. The multiplicity of serotonin receptors: Uselessly diverse molecules or an embarrassment of riches? *Neuroscientist* 6(4):252-62
186. Roth BL, Sheffler DJ, Kroeze WK. 2004. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat. Rev. Drug Discov.* 3(4):353-9
187. Rothman RB, Baumann MH, Savage JE, Rauser L, McBride A et al. 2000. Evidence for possible involvement of 5-HT(2B) receptors in the cardiac valvulopathy associated with fenfluramine and other serotonergic medications. *Circulation* 102(23):2836-41
188. Ruiz P, Faroon O, Moudgal CJ, Hansen H, De Rosa CT, Mumtaz M. 2008. Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure-activity relationship (QSAR). *Toxicology Letters* 181(1):51-63
189. Salemme FR. 2003. Chemical genomics as an emerging paradigm for postgenomic drug discovery. *Pharmacogenomics* 4(3):257-67
190. Schade R, Andersohn F, Suissa S, Haverkamp W, Garbe E. 2007. Dopamine agonists and the risk of cardiac-valve regurgitation. *New England Journal of Medicine* 356(1):29-38
191. Schmidt PJ, Rubinow DR. 2009. Sex Hormones and Mood in the Perimenopause. *Glucocorticoids and Mood Clinical Manifestations, Risk Factors, and Molecular Mechanisms* 1179:70-85
192. Seierstad M, Agrafiotis DK. 2006. A QSAR model of hERG binding using a large, diverse, and internally consistent training set. *Chemical Biology & Drug Design* 67(4):284-96

193. Setlur SR, Mertz KD, Hoshida Y, Demichelis F, Lupien M et al. 2008. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J Natl. Cancer Inst.* 100(11):815-25
194. Setola V, Hufeisen SJ, Grande-Allen KJ, Vesely I, Glennon RA et al. 2003. 3,4-methylenedioxymethamphetamine (MDMA, "Ecstasy") induces fenfluramine-like proliferative actions on human cardiac valvular interstitial cells in vitro. *Mol. Pharmacol.* 63(6):1223-9
195. Setola V, Roth BL. 2005. Screening the receptorome reveals molecular targets responsible for drug-induced side effects: focus on 'fen-phen'. *Expert Opin. Drug Metab Toxicol.* 1(3):377-87
196. Setola V, Roth BL. 2008. The emergence of 5-HT<sub>2B</sub> receptors as targets to avoid in designing and refining pharmaceuticals. In *The Serotonin Receptors: From Pharmacology to Human Therapeutics*, ed. BL Roth, 13:419. Totowa, NJ: Human Press.
197. Shannon C, Weaver W. 1949. *In mathematical theory of communication.*, University of Illinois.
198. Shapiro DA, Renock S, Arrington E, Chiodo LA, Liu LX et al. 2003. Aripiprazole, a novel atypical antipsychotic drug with a unique and robust pharmacology. *Neuropsychopharmacology* 28(8):1400-11
199. Shen M, Beguin C, Golbraikh A, Stables JP, Kohn H, Tropsha A. 2004. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* 47(9):2356-64
200. Shen M, LeTiran A, Xiao Y, Golbraikh A, Kohn H, Tropsha A. 2002. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* 45(13):2811-23
201. Sheridan RP, Kearsley SK. 2002. Why do we need so many chemical similarity search methods? *Drug Discov. Today* 7(17):903-11
202. Shumaker SA, Legault C, Rapp SR, Thal L, Wallace RB et al. 2003. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in

postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA* 289(20):2651-62

203. Simon-Hettich B, Rothfuss A, Steger-Hartmann T. 2006. Use of computer-assisted prediction of toxic effects of chemical substances. *Toxicology* 224(1-2):156-62
204. Stahura FL, Bajorath J. 2004. Virtual screening methods that complement HTS. *Combinatorial Chemistry & High Throughput Screening* 7(4):259-69
205. Sukumar N, Krein M, Breneman CM. 2008. Bioinformatics and cheminformatics: Where do the twain meet? *Current Opinion in Drug Discovery & Development* 11(3):311-9
206. Swanson DR. 1990. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.* 78(1):29-37
207. Tang H, Wang XS, Huang XP, Roth BL, Butler KV et al. 2009. Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation. *Journal of Chemical Information and Modeling* 49(2):461-76
208. Tetko IV. 2002. Neural network studies. 4. Introduction to associative neural networks. *Journal of Chemical Information and Computer Sciences* 42(3):717-28
209. The connectivity map. The Connectivity Map Help. The Connectivity Map . 2010. Broad Institute.  
Ref Type: Electronic Citation
210. Todeschini R, Consonni V. *Handbook of molecular descriptors*. 2000. Weinheim, Germany, Wiley-VCH.  
Ref Type: Computer Program
211. Tropsha A. 2006. Predictive QSAR (Quantitative Structure Activity Relationships) Modeling. In *Comprehensive Medicinal Chemistry II*, ed. YC Martin, 7:113-126. Elsevier.

212. Tropsha A, Golbraikh A. 2007. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* 13:3494-504
213. Tropsha A. 2003. **Recent Trends in Quantitative Structure-Activity Relationships.** In *Burger's Medicinal Chemistry and Drug Discovery*, ed. Abraham D,49-77. New York: John Wiley & Sons.
214. Tropsha A. 2010. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* 29(6-7):476-88
215. Tropsha A, Pearlman RS. 2000. Computer-aided combinatorial chemistry and cheminformatics. *Pac. Symp. Biocomput.*(12):553-4
216. Tropsha A, Wang SX. 2006. QSAR modeling of GPCR ligands: methodologies and examples of applications. *Ernst. Schering. Found. Symp. Proc.*(2):49-73
217. van der Greef J, McBurney RN. 2005. Innovation - Rescuing drug discovery: in vivo systems pathology and systems pharmacology. *Nature Reviews Drug Discovery* 4(12):961-7
218. van 't V, Dai H, van de V, He YD, Hart AA et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530-6
219. Vangala S, Tonelli A. 2007. Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? *AAPS J* 9(3):E284-E297
220. Vapnik VN. 1995. *The Nature of Statistical Learning Theory.*, New York: Springer-Verlag.
221. Varnek A, Gaudin C, Marcou G, Baskin I, Pandey AK, Tetko IV. 2009. Inductive Transfer of Knowledge: Application of Multi-Task Learning and Feature Net Approaches to Model Tissue-Air Partition Coefficients. *Journal of Chemical Information and Modeling* 49(1):133-44
222. Varnek A, Tropsha A. 2008. *Chemoinformatics Approaches to Virtual Screening*, Cambridge, UK: RSCPublishing.

223. Venkatapathy R, Wang CY, Bruce RM, Moudgal C. 2009. Development of quantitative structure-activity relationship (QSAR) models to predict the carcinogenic potency of chemicals I. Alternative toxicity measures as an estimator of carcinogenic potency. *Toxicology and Applied Pharmacology* 234(2):209-21
224. Wagner BK, Clemons PA. 2009. Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling. *Curr. Opin. Chem. Biol.* 13(5-6):539-48
225. Wermuth CG. 2004. Multitargeted drugs: the end of the "one-target-one-disease" philosophy? *Drug Discov. Today* 9(19):826-7
226. Whittle M, Gillet VJ, Willett P. 2006a. Analysis of data fusion methods in virtual screening: Theoretical model. *Journal of Chemical Information and Modeling* 46(6):2193-205
227. Whittle M, Gillet VJ, Willett P, Loesel J. 2006b. Analysis of data fusion methods in virtual screening: Similarity and group fusion. *Journal of Chemical Information and Modeling* 46(6):2206-19
228. Wiener HJ. 1947. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* 69:17-20
229. Wishart DS. 2007. Human Metabolome Database: completing the 'human parts list'. *Pharmacogenomics* 8(7):683-6
230. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* 36:D901-D906
231. Wishart DS, Knox C, Guo AC, Eisner R, Young N et al. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* 37(Database issue):D603-D610
232. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M et al. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34:D668-D672

233. Wold S, Eriksson L. 1995. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design (Methods and Principles in Medicinal Chemistry, Vol 2)*, ed. Hvd Waterbeemd, 309-318. Weinheim (Germany): Wiley-VCH Verlag GmbH.
234. Woolley CS. 2007a. Acute effects of estrogen on neuronal physiology. *Annual Review of Pharmacology and Toxicology* 47:657-80
235. Woolley CS. 2007b. Estrogen and synapses in the hippocampus. *Faseb Journal* 21(5):A89
236. Woolley CS, Schwartzkroin PA. 1998. Hormonal effects on the brain. *Epilepsia* 39:S2-S8
237. Xiao Z, Varma S, Xiao YD, Tropsha A. 2004. Modeling of p38 mitogen-activated protein kinase inhibitors using the Catalyst HypoGen and k-nearest neighbor QSAR methods. *J. Mol. Graph. Model.* 23(2):129-38
238. Yaffe K, Krueger K, Cummings SR, Blackwell T, Henderson VW et al. 2005. Effect of raloxifene on prevention of dementia and cognitive impairment in older women: The multiple outcomes of raloxifene evaluation (MORE) randomized trial. *American Journal of Psychiatry* 162(4):683-90
239. Yamamoto M, Uesugi T. 2007. Dopamine agonists and valvular heart disease in patients with Parkinson's disease: evidence and mystery. *Journal of Neurology* 254:74-8
240. Yamamoto M, Uesugi T, Nakayama T. 2006. Dopamine agonists and cardiac valvulopathy in Parkinson disease - A case-control study. *Neurology* 67(7):1225-9
241. Yang R, Niepel M, Mitchison TK, Sorger PK. 2010. Dissecting variability in responses to cancer chemotherapy through systems pharmacology. *Clin. Pharmacol. Ther.* 88(1):34-8
242. Yoshida K, Niwa T. 2006. Quantitative structure-activity relationship studies on inhibition of HERG potassium channels. *Journal of Chemical Information and Modeling* 46(3):1371-8

243. Zanettini R, Antonini A, Gatto G, Gentile R, Tesei S, Pezzoli G. 2007. Valvular heart disease and the use of dopamine agonists for Parkinson's disease. *New England Journal of Medicine* 356(1):39-46
244. Zhang L, Yu J, Pan H, Hu P, Hao Y et al. 2007. Small molecule regulators of autophagy identified by an image-based high-throughput screen. *Proc. Natl. Acad. Sci. U S A* 104(48):19023-8
245. Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A. 2008a. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25(8):1902-14
246. Zhang ZY, Niu JF, Zhi X. 2008b. A QSAR model for predicting mutagenicity of nitronaphthalenes and methylnitronaphthalenes. *Bulletin of Environmental Contamination and Toxicology* 81(5):498-502
247. Zhao W, Wang J, Ho L, Ono K, Teplow DB, Pasinetti GM. 2009. Identification of antihypertensive drugs which inhibit amyloid-beta protein oligomerization. *J Alzheimers Dis.* 16(1):49-57
248. Zheng W, Tropsha A. 2000. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* 40(1):185-94
249. Zheng XF, Chan TF. 2002a. Chemical genomics in the global study of protein functions. *Drug Discov. Today* 7(3):197-205
250. Zheng XF, Chan TF. 2002b. Chemical genomics: a systematic approach in biological research and drug discovery. *Curr. Issues Mol. Biol.* 4(2):33-43
251. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E et al. 2008a. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena pyriformis*. *J. Chem. Inf. Model.* 48(4):766-84
252. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. 2008b. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics* 24(23):2798-800



253. Zimmer M, Ebert BL, Neil C, Brenner K, Papaioannou I et al. 2008. Small-molecule inhibitors of HIF-2 $\alpha$  translation link its 5'UTR iron-responsive element to oxygen sensing. *Mol. Cell* 32(6):838-48
  
254. Zimmer M, Lamb J, Ebert BL, Lynch M, Neil C et al. 2010. The connectivity map links iron regulatory protein-1-mediated inhibition of hypoxia-inducible factor-2 $\alpha$  translation to the anti-inflammatory 15-deoxy-delta12,14-prostaglandin J2. *Cancer Res.* 70(8):3071-9