Prioritizing Small Molecules for Drug Discovery or Chemical Safety Assessments using Ligand- and Structure-based Cheminformatics Approaches

Hao Tang

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 Department of Biochemistry and Biophysics

Chapel Hill
2011

Approved by

Advisor: Alexander Tropsha, Ph.D.

Reader: Matthew Redinbo, Ph.D.

Reader: Brian Kuhlman, Ph.D.

Reader: Andrew Lee, Ph.D.

Reader: Yufeng Liu, Ph.D.

ABSTRACT

Hao Tang : Prioritizing Small molecules for Drug Discovery or Chemical Safety
Assessments using Ligand- and Structure-based Cheminformatics Approaches
(Under the direction of Professor Alexander Tropsha)

Recent growth in the experimental data describing the effects of chemicals at the
molecular, cellular, and organism level has triggered the development of novel computational
approaches for the prediction of a chemical's effect on an organism. The studies described in
this dissertation research predict chemical activity at three levels of biological complexity:
binding of drugs to a single protein target, selective binding to a family of protein targets, and
systemic toxicity. Optimizing cheminformatics methods that examine diverse sources of
experimental data can lead to novel insight into the therapeutic use and toxicity of chemicals.

In the first study, a combinatorial Quantitative Structure-Activity Relationship
(QSAR) modeling workflow was successfully applied to the discovery of novel bioactive
compound against one specific protein target: histone deacetylase inhibitors (HDACIs). Four
candidate molecules were selected from the virtual screening hits to be tested experimentally,
and three of them were confirmed active against HDAC.

Next, a receptor-based protocol was established and applied to discover target-
selective ligands within a family of proteins. This protocol extended the concept of
protein/ligand interaction-guided pose selection by employing a binary classifier to
discriminate poses of interest from a calibration set. The resulting virtual screening tools
were applied for enriching beta2-adrenergic receptor (β2AR) ligands that are selective
against other subtypes in the βAR family (i.e. β1AR and β3AR). Moreover, some

computational 3D protein structures used in this study have exhibited comparative or even better performance in virtual screening than X-ray crystal structures of β2AR, and therefore computational tools that use these computational structures could complement tools utilizing experimental structures.

Finally, a two-step hierarchical QSAR modeling approach was developed to estimate *in vivo* toxicity effects of small molecules. Besides the chemical structural descriptors, the developed models utilized additional biological information from *in vitro* bioassays. The derived models were more accurate than traditional QSAR models utilizing chemical descriptors only. Moreover, retrospective analysis of the developed models helped to identify the most informative bioassays, suggesting potential applicability of this methodology in guiding future toxicity experiments.

These studies contribute to the development of computational strategies for comprehensive analysis of small molecules' biological properties, and have the potential to be integrated into existing methods for modern rational drug design and discovery.

DEDICATION

This thesis is dedicated to my parents, Miaorong Tang and Xiaobo Zhang, for their

support throughout my life. Also, I would like to dedicate this thesis to my husband,

Yuan Cheng, and my daughter, Claire Ziyun Cheng

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AD   Applicability Domain

ADME   Absorption, Distribution, Metabolism, Excretion

CCR   Correct Classification Rate

CV   Cross Validation

DFT   Density Functional Theory

DUD   Directory of Useful Decoys

EF   Enrichment Factor

HTS   High Throughput Screening

kNN   k-Nearest Neighbors

LBVS   Ligand-based Virtual Screening

LOO   Leave One Out

MCT   Maximal Charge Transfer

PDB   Protein Data Bank

QSAR   Quantitative Structure Activity Relationship

RF   Random Forest

RMSD Root Mean Square Deviation

ROC   Receiver Operating Characteristic

SBVS   Structure-based Virtual Screening

SE   Sensitivity

SP   Specificity

SVM   Support Vector Machines

TC   Tanimoto Coefficient

VS   Virtual Screening

Chapter 1. Introduction

1.1. Overview

The famous "central dogma" of molecular biology explains the information flow through macromolecules from DNA to RNA to proteins. Yet, many essential biological processes cannot be fulfilled with macromolecules alone. A variety of small molecules, such as transmitters, hormones, metabolites, and xenobiotics, are essential to understand the big picture of life and are arguably the "missing link" of the "central dogma" ([1] Figure1.1). Acquiring good knowledge of these small molecules and how they interact with biological systems could significantly enhance our abilities to characterize protein functions, decipher the signaling pathways, and direct the optimization of drug leads for treatments of diseases. In order to investigate the effects of small molecules, rapid growing disciplines including chemical biology, chemogenomics, and chemical genetics have evolved at the interface between chemistry and molecular biology, statistical modeling and computational sciences [2-4]. Despite the diversity and complexity of researches in these fields, the major themes are perhaps best described as to use small molecules as probes to study biological functions, with an ambitious goal to comprehensively explore all possible drug candidates, potential pharmaceutical relevant targets, and the entire drug-target interaction network [5,6].

With the increasing amount of data produced in this field, computational tools are recently developed to either help with interpretation of the data, or convert the information into knowledge and predictive tools that can be applied to guide future studies. Many of these computational tools are derived from conventional cheminformatics and bioinformatics methods

including, for instance, the analysis of small molecule similarity and structure-activity relationships, the study of ligand-target associations through molecular docking and scoring, or the construction of target associations through sequence/structure similarity and the ligand-target interaction network [7-9] (Figure 1.2). A comprehensive overview of computational tools developed in this field is beyond the scope of this chapter; rather, the focus here is to provide a brief summary of the cheminformatics and virtual screening tools that have been explored and evaluated in this dissertation, as well as highlight the contributions of this study.

1.2. Quantitative Structure-Activity Relationships Analysis

The major assumptions underlying any Quantitative structure-activity relationship (QSAR) analysis is that "similar molecules are most likely to obtain similar properties" (the Similarity Property Principle), which has shown to be generally valid. Classical QSAR models attempt to quantitatively relate structure variations of small molecules to changes in their biological properties, such as binding affinities and inhibition constants. Any QSAR practice can generally be represented by a mathematical formula: $P = \hat{k}(D_1, D_2, \ldots, D_n)$ where P is a biological property of interest for the molecules; D_1, $D_2 \ldots D_n$ are calculated structural descriptors that characterize molecules' physical-chemical properties, and $\hat{k}$ is some mathematical transformation to derive the property for the molecule from the descriptors [10]. The success of a QSAR modeling campaign thus depend on the robustness of the structural descriptors employed, as well as the statistical learning strategies applied to the construction of the structure-activity relationships. Recently, there is an increasing emphasis in the QSAR community to carry out rigorous model validation to afford robust and predictive models, which can be applied to virtual screening of external compounds with unknown activities [10-14].

Most of the traditional QSAR models assume a single mode of action for the tested small molecules, as well as a linear relationship between a biological activity and molecular descriptors. This may be a reasonable approximation for relatively small data sets. However, it is often not the case for the data from chemical biology, especially when the monitored biological functions can be phenomenological in nature, and a large set of structurally diverse compounds is considered. During the past decades, we have witnessed a great interest in using more complex machine learning techniques, such as the Bayesian Classifier [15], k nearest neighbors (kNN) [16], randomForest (RF)[17], and support vector machines (SVM) [18,19], to assist computational modeling in this field. As it is highly impossible to decide a priori as to which modeling technique could be most effective, it is usually recommended to use the combi-QSAR approach that explores all possible combinations of various descriptor types and statistical learning algorithms along with external model validation. Chapter 2 describes an example that employed the combi-QSAR approach and rigorous model validation to identify novel histone deacetylase inhibitors.

Despite the increasing complexity and diversity of statistical learning algorithms applied, it is difficult to develop predictive tools for chemical biology data using most traditional QSAR models that are based on chemical descriptors alone. There are two major challenges: the significant structure diversity of the data set on one hand, and the variety of structural features that can cause similar effect, on the other hand. These challenges are most prominent in the assessment of environmental chemicals' toxicological effects, where chemicals were designed for different reasons with various scaffolds, and can act through multiple mechanisms and hit various physiologically important targets to cause similar adverse effects [20,21]. On the other hand, the development of various *in vitro* toxicity testing methods, such as cell-based and cell-free HTS techniques, as well as toxicogenomic technologies, offered potential biological basis for

estimating the adverse effects of environmental chemicals [22-24]. Therefore, it is intriguing to develop novel QSAR modeling approaches that can combine traditional chemical descriptors with the knowledge extracted from the *in vitro* testing results. Indeed, our recent studies showed that it is possible to improve QSAR models' predictivity by including *in vitro* testing results as biological descriptors with traditional chemical descriptors [25,26]. However, this approach is not always effective, partially owing to the overwhelming influence of chemical descriptors when modeling with hybrid descriptors. Recently, we showed that it is also beneficial to utilize the correlation between rodent acute toxicity data (*in vivo* data) and cytotoxicity data (*in vitro* data) to enhance the performance of traditional QSAR models with chemical descriptors only [27,28]. In Chapter 5, we illustrated when it is difficult to build predictive models from traditional QSAR modeling approach using either chemical descriptors or hybrid descriptors (chemical plus biological descriptors), we can still manage to utilize the *in vitro* vs. *in vivo* correlations and a novel two-step hierarchical QSAR modeling workflow to construct models for three rat reproductive toxicity endpoints.

1.3. Molecular Docking

Receptor-ligand docking has become a fundamental component of modern drug discovery process. It consists of two distinct steps: docking and scoring. The docking step attempts to explore all possible conformations and orientations of a candidate ligand into the active site. Each solution is named a pose. The scoring step deals with determining the binding affinity of each pose, and ranks the ligands according to the predicted values. The goal is to find the most appropriate pose for each ligand and to indentify the ligands with the highest potential as drug candidates. Following the pioneering work form the first docking-based virtual screening

approaches, namely UCSF Dock, there are currently more than 60 docking programs (e.g. Dock, AutoDock, eHiTs, FlexX, Fred, Glide) and more than 30 scoring functions[9,29]. To date, structure-based virtual screening (SBVS) has been successfully applied in numerous studies and enabled well documented discovery of several approved drugs, e.g. dorzolamide [30,31] and relenza [32].

A docking study provides concomitantly the estimates on small molecule's binding pattern and affinity to the target macromolecule. When there is a set of related protein targets, docking studies would, ideally, provide good estimates on small molecule's binding selectivity across these targets (ligand profiling). Compared with ligand-centric methods that use small molecule information alone, docking represents a promising complementary approach that include 3D information about the target protein to predict compounds binding selectivity. Therefore, there is an increasing interest to use docking approaches in both retrospective and prospective studies of small molecules' binding selectivity[33,34]. Nevertheless, due to the inherent limitations of docking programs, this is never a trivial task.

Typically, the 3D structures of biomolecules obtained by X-ray crystallography and NMR spectroscopy are needed for the purpose of SBVS, whose performance is strongly affected by the quality of biomolecular structure, especially with respect to binding site description. When no experimentally determined structures are available, theoretical models based on either homology or *de novo* modeling approaches are employed instead [35-38]. In principle, the success of structure models is typically measured by how close the models could reproduce experimental structures, which implies that the latter are inherently more appropriate choice for SBVS applications. However, this may not always be true, especially when one takes into account of the fact that some of the computational models are actually manually refined with known medicinal chemistry data to reproduce conserved protein-ligand contacts. There have been some

discussions about the accuracy and applicability of theoretical models [38-44] in SBVS. The G-protein coupled receptors (GPCRs) serves as a good example where theoretical models have been used widely because of GPCRs' importance as targets for many drugs, on one hand, and the lack of experimental structures, on the other hand. Thus, the recent publications of the crystal structure of human beta-2 adrenergic GPCR (β2AR)[45-47] cleared the way for the validation of previous theoretical models, as well as provided critical guide for building structural models of other GPCRs. To assess the accuracy and applicability of structure models in SBVS, we have compared several beta-2 adrenergic receptor (β2AR) structural models *versus* the β2AR-T4L crystal structure in terms of both their global similarity and effectiveness of use in search for β2AR specific agonists and antagonists (Chapter 3).

Another critical issue for using docking as a ligand profiling approach is the inaccuracy of the predicted binding pose and affinities. While most docking program can reproduce experimentally resolved binding conformations, those poses are not always picked by the scoring functions [48]. The generated scores also do not correlate with the order of measured activities, thus most docking programs demand further visual inspections of the top-scored poses [49,50]. To overcome this problem, the scoring functions are constantly modified to enable more accurate prediction of the binding energy of a small molecule. The developed approaches, however, are often very computationally intensive, resulting in less pracitical approaches for large scale virutal screening. Alternative approaches are to use pre- or post-docking filters to eliminate unreasonable ligand poses or uninterested compounds. The examples of several approaches that allow pharmacophore constraints during docking runs include Glide [51], Gold [52], LIDAEUS[53], and FlexX [54]. Several other groups aim at selecting only ligand poses that possess known conserved protein-ligand contacts. Singh and colleagues [55,56] defined a series of protein-ligand interactions

6

properties to generated Structural interaction fingerprint (SIFt) at protein-ligand's binding site. Virtual screening studies using SIFt revealed superior performance as compared with traditional scoring functions. Graaf and Rognan et al. showed that a topological scoring function based on another ligand protein interaction fingerprint (IFPs) was able to selectively identify agonists or antagonists of the β2AR [57-59].

Our lab has developed a novel type of four-body statistical descriptor to effectively represent the protein-ligand interface: the PL/MCT-tess (Protein-Ligand atoms' pair wise Maximal Charge Transfer potential based on Delaunay Tessellation) descriptor. A recent study by Hsieh et al. in our group suggested that QSAR models developed based on PL/MCT-tess descriptors can effectively distinguish native-like docking poses from decoy poses, thus dramatically increased the virtual screening performance when combined with the traditional force-field based scoring functions[60,61]. Herein, we continue to use the PL/MCT-tess descriptors as an effective representation of the protein-ligand interface, and explored novel pose-filters that enable selective exclusion of irrelevant binding poses. The developed pose-filters have been further customized to learn significant contacts that relate to subtype specificity, and the derived filters were applied to search for selective ligands within the βAR sub-family (Chapter 4).

1.4. Thesis Outline

This dissertation has aimed to develop and validate computational approaches for the analysis of small molecules' toxicity and binding selectivity, both of which have high potential for application in the fast evolving fields of chemical biology and chemogenomics. A series of QSAR-based and docking-based computational tools have been developed and customized for this purpose.

Chapter 2 illustrates the power of combi-QSAR VS as a general approach for the identification of structurally novel bioactive compounds. Histone deacetylases inhibitors (HDACIs) have emerged as a new class of drugs for the treatment of human cancers and other diseases due to their effects on cell growth, differentiation, and apoptosis. To produce predictive QSAR models, a combi-QSAR approach that employ k Nearest Neighbor (kNN) and Support Vector Machines (SVM) QSAR modeling algorithms using both MolConnZ and MOE chemical descriptors have been employed. The validated QSAR models were used concurrently to screen large publically and commercially available compound databases totaling over 9.5 million molecules for novel HDAC inhibitors.

Chapter 3 and Chapter 4 contribute to the development and validation of a target-biased SBVS approach that can be further tailored to search target selective ligands. Selective small molecules provide an important library for probing biological functions in the field of chemical biology. To develop computational tools that are capable of discriminating selective ligands requires a paradigm shift from the previous single-target focus. The emphasis is on exploring interactions between individual ligand to a set of (related) protein targets. To perform ligand profiling based on SBVS requires a priori the 3D structures of all the protein targets, which may not always have experimentally solved structures available. In these situations, structure models via homology modeling or *de novo* design are used instead. This has been a long time debate whether SBVS using structure models can achieve reasoanble quality and applicability. Studies in Chapter 4 compare the SBVS applicaility of a set of well established β2AR theoretical models with that of the recently solved β2AR X-ray crystal structure. Employed both the carefully selected structural models and the recently solved β2AR X-ray crystal structure, Chapter 5 describes the development and validation of novel target-biased pose filters for selectively enrich

sub-type ligand within the βAR GPCR subfamily. The developed pose filters could not only help with ligands identification, but are also useful in finding potential binding residues that are essential for ligand selectivity.

In Chapter 5, traditional QSAR models, as well as the specialized QSAR models that incorporate biological information from *in vitro* testing data have been evaluated for their potential to forecast environmental chemicals' *in vivo* toxicity effects. Others' studies as well as our previous efforts showed that while it is possible to construct predictive QSAR models using chemical descriptors alone, it is usually not a simple task for most toxicity endpoints. We hypothesize that model performance could be improved by including additional information about how molecules' can perturb important signaling pathways and interact with the biological system. The available *in vitro* testing data could serve as biological probe to predict chemicals' systemic adverse effects. Therefore, we proposed to develop a novel two-step QSAR approach that incorporates the *in vitro* testing results to complement traditional chemical-descriptors-based QSAR models in predicting compounds' toxicity effects.

Figures and Tables
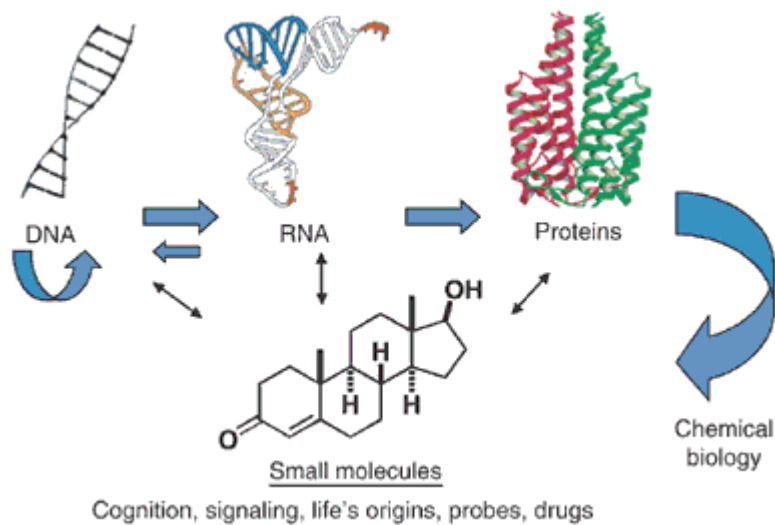
Figure 1.1. Illustration of the role of small molecules as viewed by a chemical biologist. (Adapted from [1])
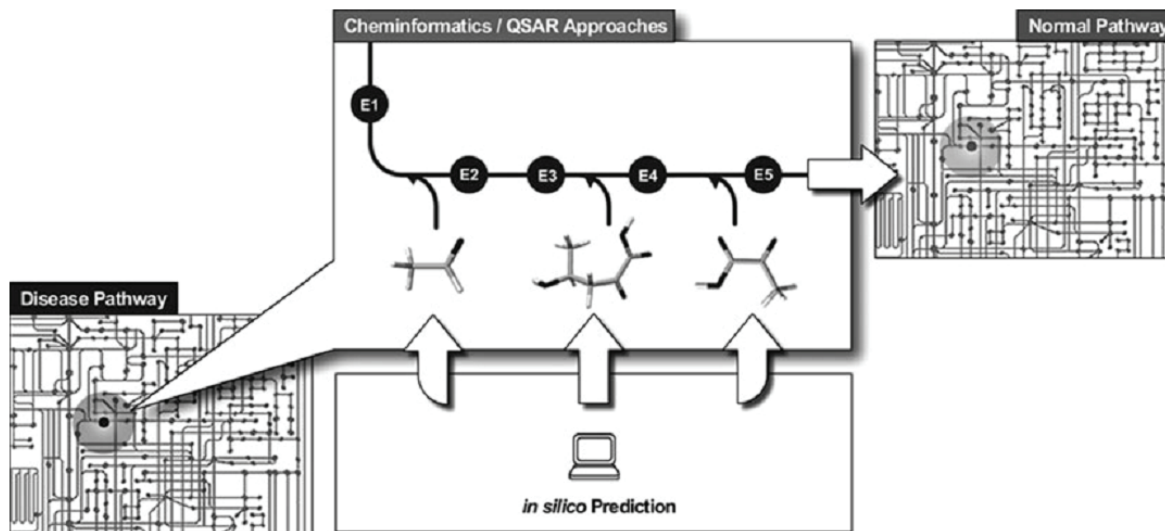
Figure 1.2. Contribution of cheminformatics to systems biology. Computational modeling helps with the prediction of compounds activity against single/multiple targets, while the PBPK models will provide estimation on compound distribution and accumulation in tissues. (Adapted from [62])

Chapter 2. Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation

(This chapter has been published as Tang, H., et al. J Chem Inf Model, 2009. 49(2): p. 461-76)

2.1. Introduction

The dynamic posttranslational modification of nucleosomal histones plays a critical role in transcriptional regulation. Hyperacetylation of core histones results in transcriptional activation, while hypoacetylation leads to expression repression[63]. This kind of regulation is considered to be the critical step in normal cell differentiation and chromatin condensation and is believed to be regulated by the balance between two groups of enzymes: histone deacetylases (HDACs) and histone acetyltransfereases (HATs) [64,65]. Inhibition of HDACs represents a novel approach to interfere with cell cycle regulation; therefore, it has a great therapeutic potential in the treatment of diseases of aberrant cellular proliferation [66]. It has been reported that hyperacetylation of histones and non-histone proteins induced by small molecule HDACs inhibitors (HDACI) leads to cell growth arrest, cellular differentiation and/or apoptosis of malignant cells [67-70]. For these reasons, HDACI has become a promising class of chemical agents for the treatment of cancer and other diseases associated with uncontrolled cell proliferation.

To date, a number of structurally distinct classes of HDACI have been reported, including hydroxamates, cyclic peptides, aliphatic acids and benzamides [71,72]. The natural product Trichostatin A (TSA) [73] is the most well-known member of the hydroxamates class;

this compound is considered to be a mimetic of the natural substrate, i.e., histone acetyl lysine side chain. Extensive structure-activity relationship (SAR) studies have been conducted for TSA and TSA-like compounds resulting in several potent HDACs inhibitors [74-77]. A TSA analog suberoylanilide hydroxamic acid (SAHA) [78] was recently approved by the FDA for the treatment of cutaneous T cell lymphoma (CTCL), stimulating further investigations of HDACI in the treatment of various diseases [79].

HDACs have been classified so far into four classes (Classes I-IV) depending on the sequence identity and domain organization. Among the Class I HDACs, HDACs 1,2 and 8 are primarily found in the nucleus, whereas HDAC 3 is found in the nucleus, cytoplasm and the membrane. In comparison, Class II HDACs subdivided into IIa (HDAC 4, 5, 7, 9) and IIb (HDAC 6, 10) are able to shuttle in and out of the nucleus depending on different signals. Class III HDACs include the SIRTs (sirtuins) or Sir2- related proteins; they are NAD-dependent [80] and are insensitive to TSA or other hydroxamate inhibitors. Class IV comprises of HDAC 11, based on a phylogenetic analysis and is the least characterized to date [81]. It has been considered important in recent years to develop class/subtype selective HDACI. Considering the number of pathways in which HDACs are involved, the HDACI that act exclusively on Class I or Class II enzymes are viewed as likely candidates as anticancer therapeutic agents.

The crystal structures of the histone deacetylase like protein (HDLP) both in the apo form and in complexes with TSA and SAHA were first published by Finnin et al. in 1999 [82]. Five years later, Somoza's group and Di Marco's group both solved the x-ray structures of another class I histone deacetylase, histone deacetylase subtype 8 (HDAC8) in complex with several small molecule HDACI [83,84]. The crystallographic structures revealed that both HDLP

and HDAC8 contain a conserved tunnel-like binding pocket with the polar active site at the bottom. In the x-ray structure of the HDLP/TSA complex, the long aliphatic chain of TSA (linker domain) spans the whole length of the tunnel-like pocket and the hydroxamic acid moiety interacts with the polar residues at the bottom of the pocket. The chelating atoms of hydroxamic acid coordinate zinc ion in a bidentate fashion, and form hydrogen bonds with the His-Asp diad [72,82,84]. At the other end of the aliphatic chain, the aromatic group of TSA (surface recognition domain) interacts with the hydrophobic rim of the pocket [72]. Thus, SAR studies have been typically focused on three regions of HDACI: the metal binding group, the linker domain, and the surface recognition domain [77].

Because of their potential clinical importance, HDACI have been a subject of several Quantitative Structure-Activity Relationship (QSAR) modeling studies [75,76,85]. The results of these studies are summarized in Table 2.1. Most of them focused on a series of hydroxamates and employed 3D QSAR modeling methods. This preference was partially due to the fact that a number of HDACs crystallographic structures have been solved in recent years and thus could be used for structural alignment of inhibitors to enable 3D QSAR modeling. The size of HDACI datasets varied among different reports, ranging from 19 to 124. The best reported models were characterized by leave-one-out cross-validation (LOO-CV) R^2 of 0.870 and R^2 of 0.987. For the test set, the R^2 was as high as 0.896. It should be pointed out that none of these earlier studies had employed an independent dataset for model validation, and none used models for virtual screening of chemical libraries to identify novel hits.

In the present study of HDACI, we have applied the modeling strategy that has been under development in our laboratory for several years [86]. The important feature of our approach is that it combines validated QSAR modeling of historic data and virtual screening

of available chemical libraries for the identification of novel active compounds, as illustrated in Figure 2.1. We have used experimental data for 59 histone deacetylase subtype 1 (HDAC1) inhibitors that were generated in one of our laboratories. All of the compounds in the dataset were hydroxamates but incorporated many novel chemical modifications in the three major domains, i.e., the hydroxamic acid, the linker domain and surface recognition domain. Our studies resulted in externally predictive QSAR models of HDAC1 inhibitors. Furthermore, the application of these models to virtual screening of a large (ca. 9.5 million) collection of commercially available chemical compounds identified several computational hits, and three of them were confirmed experimentally as novel active HDAC1 inhibitors.

2.2. Materials and Methods

2.2.1.   Datasets for Model Building and Validation.

59 compounds with known HDAC1 inhibition activities were employed for the QSAR study (cf. Appendix I). All data were generated in the laboratories of Dr. Kozikowski (chemistry) and Dr. Jung (biology) at the George Washington University and most of them were reported earlier [87-92]. The data for eight compounds, BC-2-87, BC-3-63, BC-3-70, BC-3-94, BC-4-93, BC-6-30, BC-6-33 and BC-6-34, are reported here for the first time. The half maximal (50%) inhibitory concentration of a substance (IC50) was measured on HDAC1 from HeLa cell extracts. It was then converted to the pIC50 scale (-logIC50), in which higher values indicate exponentially greater potency.

Two independent external validation sets of different nature were employed in the later phase of our modeling workflow (cf. Figure 2.1): one included 9 HDAC1 inhibitors randomly selected from the original set of 59 compounds, and another comprised 54 diverse HDAC1 inhibitors collected from two general reviews on HDACIs [72,74]. These external sets

14

have covered most chemical classes of known HDACI [77,93-97]. Other compounds discussed in the reviews were excluded either because their HDAC1 binding affinity data were not reported or they were duplicates of compounds included in the modeling set. The observed pIC50 values of 54 compounds ranged between 4.0 and 8.0 which are similar to the activity range observed for the 50 compounds used for model development.

2.2.2. Libraries for Virtual Screening.

The virtual screening was performed on our in-house collection of ca. 9,500,000 molecules, including the ZINC7.0 database of ca. 6,500,000 compounds[98], the World Drug Index (WDI) database of ca. 59,000 compounds [99], the ASINEX Synergy libraries (2006.04) of ca. 11,000 compounds [100], the InterBioScreen screening libraries (2007.03) of ca. 400,000 compounds [101], the Chemizon Progenitor databases (2006 v1.1) of ca. 3,300 compounds [102], and several other commercial databases. None of the compounds present in the modeling set were found in the screening libraries. MolConnZ4.09 (MZ4.09) descriptors were calculated for each compound in the databases and linearly normalized based on the maximum and minimum values of each descriptor type in the modeling dataset of 59 HDAC1 inhibitors.

2.2.3. Generation of MolConnZ Descriptors.

The MolConnZ4.05 (MZ4.05) software [103] affords the computation of a wide range of topological indices (descriptors) of molecular structure such as simple and valence path, cluster, path/cluster and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev-Trinajsti, information indices, counts of different vertices, counts of paths and edges between different kinds of vertices [104-111]. Overall, MZ4.05 produces more than 400 different descriptors. In this study,

only 262 chemically relevant descriptors were eventually used after removing those with zero value or zero variance. MZ4.05 descriptors were range-scaled because the absolute values of individual types could differ by orders of magnitude [16]. Therefore, range scaling prevents undesirable overweighting of descriptors with high ranges of values in calculating compound similarities as part of QSAR modeling procedure.

2.2.4. Generation of MOE Descriptors.

The MOE2006.08 software [112] generates both 2D and 3D descriptors. 2D molecular descriptors include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors [104,108,109,113-116]. 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors [117]. For model generation, we used 179 MOE descriptors with non-zero value and variance that were range-scaled.

2.2.5. Selection of Training and Test Sets.

The dataset was subdivided into multiple training/test set pairs using the Sphere Exclusion method developed in our laboratory [118,119]. By default, fifty different training/test set splits were initially tried using probe sphere radii defined by the minimum and maximum elements, Dmin and Dmax, of the distance matrix D between compound-vectors in the descriptor space and forty-two splits were ultimately accepted. The number of compounds in the test set was varied to achieve the largest possible size of the test set, while ensuring that the training set models were still able to accurately predict the binding affinity of the test set compounds.

2.2.6. kNN Regression Method

The k Nearest Neighbor (kNN) QSAR method used in this study employs the kNN pattern recognition principle [120] and variable selection method. In short, a subset of variables (descriptors) is selected randomly as a Hypothetical Descriptor Pharmacophore (HDP) [121]. The HDP is validated by LOO-CV, where each compound is eliminated from the training set and its HDAC1 inhibition activity is predicted as the weighted average of the activity(ies) of the k most similar molecules (k varies from 1 to 5). The weighted molecular similarity is represented by the modified Euclidean distance between compounds in HDP multidimensional space as shown in Equations 1 and 2. Essentially, the neighbor with the smaller distance from a compound is given a higher weight in calculating the predicted activity.

$$w_i = \frac{e^{-d_i}}{\sum_i e^{-d_i}} \qquad (1)$$

$$\tilde{y} = \sum w_i y_i \qquad (2)$$

where d_i is the Euclidean distance between the compounds i and its kth nearest neighbors; wi is the weight for the kth nearest neighbor; y_i is the experimentally measured activity value for the kth nearest neighbor; and $\tilde{y}$ is the predicted activity value.

Simulated annealing and Metropolis-like acceptance criteria were used to optimize the variables. Details of the kNN method implementation, including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere50. The statistical significance of the models were estimated by the LOO-CV q2 in the training set, a coefficient of determination R02 (Equation 3) and a linear fit predictive R^2 for both internal and external test sets.

$$q^2(R_o^2) = 1 - \frac{\sum(\tilde{y}_k - y_k)^2}{\sum(\bar{y} - y_k)^2} \qquad (3)$$

Here y_k and $\tilde{y}_k$ are the observed and predicted activities of compound k, respectively, and $\bar{y}$ is the average activity of all compounds. Model acceptability cutoffs were $q^2 > 0.60$ for training set and correlation coefficient $R^2 > 0.60$ for the internal test set[57.] All models that satisfied both criteria were applied to external validation sets.

2.2.7. SVM Regression Method

Support Vector Machines (SVM) was originally developed by Vapnik [122] as a general data modeling methodology where the training set error and the model complexity are incorporated into a special loss function and simultaneously minimized during model development. The importance of the prediction error versus the model complexity can be tuned during the optimization process, in order to generate models with reasonable complexity and avoid overfitting. SVM was later extended to afford the development of SVM regression models for datasets with non-integer variables.

We have implemented SVM for QSAR modeling as described earlier [123]. In brief, given a training set of pairs $(x_i, y_i), i = 1 \dots m$, where x_i is an array of descriptors of each compound and y_i is its biological activity (e.g., IC50 value), the sought correlation between structure and activity can be represented as $y_i = f(x_i)$. For simplicity, we define $f(x_i)$ as a linear function:

$$f(x_i) = \; < w_i, x_i > +b \qquad (4)$$

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

$$loss_{min} = \frac{\|w\|}{2} + C \sum_{i=1}^{m} \xi_i \qquad (5)$$

18

with the constraint:

$$|y - (w\phi(x_i) + b)| = \xi_i \qquad (6)$$

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

In many cases, the biological activity may contain small errors or the kernel function may not be capable of perfectly representing the training compounds in a simplified manner. In order to penalize against complex models, we have added a slack variable ε to the loss function [123] in addition to the penalty parameter C. It is a threshold of prediction error for any compound's activity before the algorithm is penalized for the poor prediction. Beyond this threshold the algorithm is penalized by the value of $\xi_i - \varepsilon$. When combining the SVM optimization process defined in Equation 7 with this slack variable, the following loss function is obtained:

$$loss_{min} = \frac{\|w\|}{2} + C \sum_{i=1}^{m} \begin{cases} 0 & if\ \xi_i \le \varepsilon \\ \xi_i - \varepsilon & if\ \xi_i > \varepsilon \end{cases} \qquad (7)$$

The nature of SVM regression requires one to specify the values of C and ε a priori since it is not known beforehand which values may work the best for one particular dataset; thus, a parameter tuning must be performed. The goal is to identify optimal values of C and ε in that the model can give the best prediction for the test set. For this study we have chosen to use a "grid-search" scheme on C and ε. It starts with randomly choosing a training/test set split of the dataset, conducting a grid-search using those compounds, then fine-tuning the complete dataset over the parameter value ranges that exhibited the best results. Our coarse

grid-search of C varied from 50 to 1000 with an increment of 80, and ε varied from 0 to 1.5 with an increment of 0.15. Once the best parameter ranges were found, a fine-tuned search was carried out to search values within 200 and 0.3 units for C and ε, with the steps of 5 and 0.05 respectively.

2.2.8. Applicability Domain

Ideally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, since kNN QSAR modeling predicts test set compound activities by interpolating those of the nearest neighbor compounds in the training set, a special applicability domain, or similarity threshold, should be introduced to avoid extreme model extrapolation by making predictions for compounds that are significantly dissimilar to members of the training set [16]. In order to measure similarity, each compound is represented by a point in the M-dimensional descriptor space (where M is the total number of descriptors selected in the descriptor pharmacophore) with the coordinates $(X_{i1}, X_{i2}, \cdots, X_{iM})$, where X_{id} are the values of individual descriptors for compound i. The similarity between any two molecules is characterized by the Euclidean distance between their representative points. The Euclidean distance between two points i and j in M-dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_k (X_{ik} - X_{jk})^2} \qquad (8)$$

Compounds with the smallest distance between them are considered to have the highest similarity. The distribution of pair wise compound similarity in the training set is analyzed to produce an applicability domain threshold, D_T, as follows:

$$D_T = \bar{y} + Z\sigma \qquad (9)$$

20

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

2.2.9. *External Validation and Y-Randomization Test*

It is critical to validate a QSAR model by assessing its prediction accuracy for an external set which was not used in model building. We have conducted extensive external validations on both kNN and SVM models using two external datasets as described above. In both cases, the prediction accuracy had to satisfy the following conditions:

$$R^2 > 0.60 \tag{10}$$

$$(R^2 - R_0^2)/R^2 < 0.10 \text{ and } 0.85 < k < 1.15 \tag{11}$$

where k is the slope of the regression lines (predicted versus observed activities) through the origin. The predictions were generated using consensus models and the model coverage for each external dataset was calculated as well (vide infra).

Our previous experience suggests that more accurate results are obtained by consensus, i.e., by averaging predictions from multiple QSAR models [123,124] Thus, the consensus QSAR prediction scheme was applied to all validation set compounds found within individual applicability domains of models used in consensus prediction. The

averaged predicted activity, the fraction of models that predict the activity and the variance of the prediction values have been calculated for each compound.

In addition to external validation, Y-randomization test was carried out to establish model robustness. The test consists of rebuilding models using shuffled activities of the training set and evaluation of such models' predictive accuracy in comparison with the original model. It is expected that models obtained for the training set with randomized activities should have significantly lower values of statistical parameters such as q^2, R^2, R_0^2, etc. for training and, especially, test sets. Therefore, if most QSAR models generated in the Y-randomization test exhibit relatively high values of the statistical parameters for both training and test sets, it implies that a reliable QSAR model cannot be obtained for the given dataset. This test was applied to all QSAR approaches in this study and was repeated twice for each division.

2.2.10. QSAR-based Virtual Screening

As illustrated in the workflow of Figure 2.1, the rigorously validated QSAR models were employed for virtual screening. A global applicability domain was applied first in the complete descriptor space in order to filter out compounds that differed in their structure from the modeling set compounds. All 59 known inhibitors were exploited as the probes during the calculation. During the consensus prediction, the results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction and the standard deviation of estimated means across all models was small. Furthermore, we restricted ourselves to the most conservative applicability domain for each model using $Z_{cutoff} = 0.5$.

2.2.11. Principle Component Analysis (PCA)

The PCA calculations were carried out using the entire set of MolconnZ4.05 descriptors calculated for all compounds in the modeling set, two external validation sets, and virtual screening hits. The purpose of these calculations was to provide a visual means to compare relative positioning of the three data sets plus hits in the chemistry (i.e., multidimensional descriptor) space. The programs in the kernlab package [125] of the latest version of R2.8.0 [126] were employed. Using PCA, the distribution of compounds in the original descriptor space could be visualized in a lower dimensional space, normally in the 3D space of the first three principal components.

2.2.12. Experimental Validation of Screening Hits

Recombinant HDACs were purchased from either BIOMOL International (Plymouth Meeting, PA) or PBS Bioscience (San Diego, CA). The inhibitor activity was determined using an HDAC Fluorimetric Assay/Drug Discovery Kit from BIOMOL International according to manufacturer's protocols. Briefly, reactions were set up in 96-well plates in a total of 50 μL HDACs assay buffer (50 mM Tris-HCl of pH 8.0, 137 mM NaCl, 2.7 mM KCl, 1 mM MgCl2) containing HDAC1 (or HDAC6), testing compounds, and HDACs substrate. Trichostatin A served as the positive control and the vehicle, 1% DMSO, was employed as the negative control. The reaction was initiated by the addition of HDACs substrate at room temperature and lasted for 30 minutes. The final concentration of HDACs substrate was around its apparent Km; For HDAC1, 50 μM of substrate was used and for HDAC6, 10 or 30 μM was used. The reaction was then stopped by adding 50 μL of Fluor de Lys (TM) Assay Developer and the mixture had been incubated for another 15 minutes at room temperature. The Assay Developer was added to stop the deacetylation reaction and produce fluorophore

from the deacetylated substrate. The fluorophore can be excited at 360 nm and emits light at 460 nm. The relative fluorescence is read by a FlexStation II plate reader (Molecular Devices, Sunnyvale, CA). Initial screening concentrations were 100 uM with samples with over 50% inhibition further tested in dose response assays. The raw data (relative fluorescence units) were plotted as a function of the molar concentration of test compounds (in logarithm) and fitted to the three-parameter logistic function to calculate pIC50 by Prism 5.0 (GraphPad Software, La Jolla, CA). Here the pIC50 is defined as the logarithm of molar concentration of test compound that inhibits the fluorescence production by 50%.

2.3. Results and Discussion

2.3.1. kNN QSAR Regression Modeling

The statistical results for the 10 best kNN QSAR models using MZ4.05 descriptors are summarized in Table 2.2. 1385 models, that is, ~50% of the total number of models generated, were accepted for they had both the LOO-CV q2 values for the training set and linear fit predictive R^2 values for the test set greater than 0.60. Seventy models with q^2/R^2 values exceeding 0.75/0.75 were retained for consensus prediction. As shown in Figure 2.2A, the most predictive model afforded q^2 value of 0.81 for 34 compounds and R^2 values of 0.80 for 16 compounds (RMSE = 0.38). For models built with MOE descriptors, the best q^2/R^2 values were as high as 0.70/0.76 (RMSE = 0.45, cf. Figure 2.2C). The statistics of the top 10 kNN/MOE models are summarized in Table 2.3. Similarly, thirteen models with q^2/R^2 values exceeding 0.70/0.70 were employed for consensus prediction. These results suggest that the intrinsic structure-binding affinity relationships exist for HDAC1 inhibitors that can be best described by kNN models using both independent descriptor sets.

To ensure that the models did not merely capture noise, Y-randomization test was carried out as described above. As expected, the best models using MZ4.05 descriptors and shuffled activities only produced training set models with q2 of less than 0.40 (data not shown). Also the best kNN/MOE models using randomized activity only yielded the q^2/R^2 values less than 0.40/0.40. These results confirmed that kNN models uncovered non-spurious correlations between both MolConnZ and MOE descriptors and compound inhibition activity.

2.3.2. SVM QSAR Regression Modeling

The statistical results for top 10 SVM QSAR models using MZ4.05 descriptors are summarized in Table 2.4. The best q^2, R^2, R_0^2, values are as high as 0.93, 0.87 and 0.62, respectively. Figure 2.2B shows the best predictive model with q^2 value of 0.94 for 34 compounds and R^2 values of 0.81 for 16 compounds (RMSE = 0.51). For this model, the optimum values of C and ε were found to be 200 and 0.30, respectively. The value of 0.30 is reasonable for ε, because it is common to observe a 0.30 log unit error in enzyme/inhibition assays. Seventeen models of SVM/MZ4.05 combination with q^2/R^2 values exceeding 0.70/0.70 were retained for consensus prediction. In comparison, the performance of SVM/MOE combination was much less satisfactory. The best R^2 and R_0^2 value were as low as 0.64 and 0.53, respectively. Meanwhile, the number of acceptable models was drastically small. Thus, we did not employ SVM/MOE models for consensus prediction because of their poor accuracy ($R_{test}^2 < 0.75$).

To ensure that our SVM QSAR modeling was based on non-spurious structure/activity relationship, the inhibition activities were randomly shuffled for the training set and all calculations were repeated following exactly same protocol. The best models using randomized data only produced a R^2 of 0.20 for the test set (data not shown),

suggesting that the high R^2 is not due to a chance correlation and our accepted SVM models were robust.

2.3.3. Model Validation using External Datasets

Both kNN and SVM QSAR models validated by test sets were used to predict the inhibition activity of two external validation sets (Tables 2.5 and 2.6). For consensus prediction we have employed 70 best kNN/MolConnZ models and 17 best SVM/MolConnZ models. For the external validation set 1, the data reported in Table 2.7A-C suggests that both kNN/MolConnZ and SVM/MolConnZ consensus models afforded reasonable results. Figure 2.3 shows the correlation between experimentally measured and calculated activities of the external validation set 1 using three types of consensus models. Among the three, kNN/MolConnZ consensus models showed the best performance, with the R^2 of 0.87, R_0^2 of 0.78 and RMSE of 0.59 for 8 compounds (BC-2-83 was found to be out of applicability domain of most kNN/MolConnZ models, cf. Table 2.7A). For 7 out of these 8 compounds, the predicted activities were within a reasonable range of 0.5 log unit. However, one compound corresponding to the black circle in Figure 2.3A was predicted with a large error (> 1.0 log unit). A possible explanation for this observation is that this compound is the only one that contains two metal binding groups but no aromatic group. The latter is known to be important for the inhibition activity as suggested by many SAR studies15. The SVM/MolConnZ models performed slightly worse than the kNN/MolConnZ models, despite the fact that the SVM/MolConnZ combination had better performance for both training and test sets. The R^2 and R_0^2 of consensus prediction by SVM/MolConnZ models was 0.71 and 0.68, respectively, for all 9 compounds (Figure 2.3B). Interestingly, kNN/MOE models showed much worse statistics for the external set 1: The R^2 was 0.60 but the R_0^2 was only

0.25, and the RMSE was as high as 0.84 for 9 compounds (Figure 2.3C). Though satisfying eq 10, the statistics is not acceptable for kNN/MOE combination because the value of $(R^2 - R_0^2)/R^2$ (cf. eq. 11) is too large. Thus, we did not apply this combination to external validation set 2 and the later virtual screening. These results demonstrate the critical need of external validation set for evaluating the model robustness as well as illustrate a known phenomenon that training set accuracy does not necessarily correlate with model performance for external data sets [127].

We have used the statistical index R_0^2 (eq 3) and RMSE to evaluate model robustness in addition to the correlation coefficient R^2. Traditionally, the latter is considered as a good indicator of predictive power of models. In fact, this coefficient reflects the similarity in relative ranking of compounds based on actual vs. the calculated activities rather than the accuracy of the activity prediction. On the other hand, R_0^2 directly compares the actual vs. predicted activities because it estimates the fitness of the data to the line with the intercept of zero and the slope of one. It thus gives a better measurement of how well the model predicts compounds' activities, which is why we advocated its use as an important model accuracy metric in our previous studies [128,129]. The above case of kNN/MOE consensus prediction illustrates the difference between R^2 and R_0^2, as underscored by eq 11. This suggests that R_0^2 and RMSE are also important indicators of model robustness especially when the size of the test set is small.

External validation set 2 contains HDAC1 inhibitors of different chemical scaffolds, therefore it can be considered as a real test of the predictability of QSAR models. Besides, it is fully independent from the 59 compounds of modeling set. Among all 54 inhibitors, 41 could be predicted by the majority of consensus models and the results are summarized in

Table 2.6. For both kNN/MolConnZ and SVM/MolConnZ models, 28 out of these 41 compounds had the errors of their predicted activities of less than 1.0 log unit. The RMSE was 0.86 for kNN/MolConnZ models, 0.92 for SVM/MolConnZ models and 0.86 for the consensus averaged value of all models combined. It was shown in our recent study that consensus models afford higher prediction accuracy for the external validation data sets with the highest space coverage as compared to individual constituent models [130]. The same pattern was observed in the present study as well. The RMSE of the consensus score is superior to constituent SVM/MolConnZ models and on par with constituent kNN/MolConnZ models. In addition, there is only one compound with a relatively large margin of error (> 1.5 log unit) when the consensus prediction is used. For individual constituent models, however, there are three compounds with similarly large errors of prediction with kNN/MolConnZ models and five compounds with SVM/MolConnZ models. Compounds 6e_AE, 17j_AE and 17d_AE are among those with a large margin of error (ca. 1.5 log unit). They could be analyzed to explore the reasons for QSAR prediction errors. It should be noted that both kNN and SVM methods converged on these three compounds and showed the similar trend of errors (cf. Table 2.6). It is feasible that these compounds could be the activity outliers because of experimental errors.

2.3.4. QSAR-based Virtual Screening

Based on the results of model validation in the previous section, only kNN/MolConnZ and SVM/MolConnZ approaches were used for virtual screening due to their good performances on both modeling set and two external validation sets. Therefore, 70 kNN/MolConnZ models and 17 SVM/MolConnZ models with defined applicability domains were applied concurrently towards virtual screening of our chemical libraries. Prior to the

consensus predictions, our initial filtering using global applicability domain of modeling set reduced the total number of compounds from ca. 9.5×10^6 to 3.2×10^3. The predicted activities from individual models were averaged to yield a consensus pIC50 value. Finally, 45 hits were selected to be of high predicted activities (6.68-7.43 for kNN/MolConnZ and 5.94-7.77 for SVM/MolConnZ) and structural uniqueness.

As expected, the predicted activities of HDAC1 inhibitors by two different types of models were not identical but differed by less than 1.0 log unit in most cases. For each of the kNN/MolConnZ and SVM/MolConnZ consensus hit, we searched published literature to find out if any of these compounds was reported independently as HDAC1 inhibitors. We found that compounds #34 and #40 have been indeed cited as potential HDAC1 inhibitors (cf. Table 2.8) [131,132]. Both compounds are structurally similar to SAHA which is a strong HDAC1 inhibitor included in the modeling dataset. Furthermore, compounds #2, #28 and #35 were reported to have anti-inflammatory activity that is commonly associated with HDAC1 inhibition, which may be viewed as indirect evidence in support of the prediction.

In general, as shown in Appendix II, most hit compounds contain long aliphatic chain that permits the chelating group to reach the bottom of the binding pocket and coordinate with the zinc ion. An aromatic group at the opposite end of the chelating group is supposed to enhance inhibition through hydrophobic interaction with the capping region of the active site. These are actually the common structural features known for HDAC1 inhibitors. Furthermore, many additional features are also found in the hit compounds, such as triple bonds (compounds #2, #28) and 3-bromo-4-hydroxy-phenyl group (compounds #11, #14), which exist in HDAC1 inhibitors such as Oxamflatin and Psammaplin A [72,93]. It should be pointed out that these functional groups were not present in the original modeling dataset, which

demonstrates the ability of QSAR-based virtual screening to uncover computational hits with novel chemical features. The existence of unsaturated bonds in the linker region between the chelating group and the cap region has been observed frequently among many screening hits. However, this feature is only found in TSA (that was included in the training set), which has the highest inhibitory activity (pIC50) of 8.46. Since this feature is not often seen in other known inhibitors, this observation should be additionally explored for lead optimization in future studies. The unsaturated bonds in the linker region likely restraint the conformational freedom of the long aliphatic chain, which could help decrease the unfavorable entropy change during the inhibitor binding.

In recent years, our group has explored the hit identification strategy that combines rigorously validated QSAR models and virtual screening [123,133-137]. It has been shown that our current workflow is capable of identifying potent compounds of novel chemical scaffolds as compared to modeling set compounds, especially in the cases of anticonvulsant agents [124] and D1 dopaminergic antagonists [123]. There are several aspects of our current protocol for QSAR based virtual screening that need to be highlighted. First, models built using variable selection approaches only include a subset of all descriptors, i.e., those identified as significant in the process of model optimization. This feature of individual models coupled with the applicability domain threshold could result in mis-annotation of some structurally diverse molecules in the virtual screening databases as inactives. Consensus prediction scheme provides a viable solution to this problem because each model has its own limitations but the ensemble of models covers much greater chemical feature space and consequently, could identify putatively active compounds of greater chemical diversity. Second, the dependent variable in the current dataset is the continuous value of inhibition potency.

During model building, all descriptors with constant values have been eliminated and only the descriptor types that are used in predictive QSAR models were retained. Obviously, descriptors with the same values for all compounds in the training set could not contribute to the QSAR model that always correlates changes in chemical structure to changes in biological activity. However, there is a possibility that some of these eliminated descriptors (that apparently describe chemical features common to all inhibitors) are essential for discriminating inhibitors from non-binders. Thus, if these descriptors are not considered in virtual screening there is a probability of identifying false positives. To circumvent this problem, we have applied global applicability domain in the preliminary screening step to filter out compounds that are generally structurally dissimilar from the modeling set compounds.

2.3.5. Experimental Validation

Four structurally diverse hits with moderate to high predicted activity were selected from the 45 consensus virtual screening hits for experimental validation taking into account commercial availability. To our satisfaction, compounds #2, #28 and #35 were confirmed to be μM inhibitors against HDAC1 (Figure 2.4 and Table 2.8). Among them, compound #28 showed the best inhibitory activity with pIC50 values of 6.00. The fourth compound, #47, did not inhibit HDAC1 at the concentration of 300 μM. However, interestingly enough this compound was later identified by us as a selective inhibitor for HDAC6, a class II HDACs enzyme. At the concentration of 30 μM, #47 inhibited about 42.6% of HDAC6 activity, while other three compounds (#2, #28 and #35) showed 105%, 101%, and 99% inhibition, respectively. Moreover, it is of notice that the chelating functional group in #47 is unique compared to other hits. This observation could be further explored for rational design of

class/subtype selective HDACI. Our current screening libraries include the WDI database, which contains approximately 59,000 approved or investigational drugs in the world. It has become a practical strategy to screen this database during the early phase of drug development. The hits identified in this library could be placed on the fast track and avoid the risk and length of preclinical/clinical studies. In our study, two hits that were submitted for experimental validation were actually identified from the WDI database. Compound #35 is Bufexamac, a marketed drug used for joint and muscular pain while the compound #47 is Roxatidine, a widely used competitive H2 receptor antagonist for the treatment of peptic ulcer. These two hits will enrich the candidates pool of HDACI and potentially facilitate the pipeline of drug development—a strategy known as repurposing [138].

2.4. Conclusions

We have employed a combinatorial QSAR approach to generate models for 59 chemically diverse compounds tested for their inhibitory activity against HDAC1. The SVM and kNN QSAR methods were used in combination with MolConnZ and MOE descriptors independently to identify the best approach with the highest external predictive power. Highly predictive QSAR models were generated with kNN/MolconnZ and SVM/MolconnZ approaches. Rigorously validated QSAR models were then used to screen our in-house database collection of a total of over 9.5 million compounds. This study resulted in 45 consensus hits that were predicted to be potent HDAC1 inhibitors. Two hit compounds that were not present in the original dataset were nevertheless reported recently as HDAC1 inhibitors [131,132]. Four hit compounds with interesting chemical features were purchased and experimentally validated. Three of them were confirmed to have inhibitory activities to HDAC1 (Class I HDACs) and the best activity obtained was IC50 of 1.00 μM. The fourth

compound was later identified to be a selective inhibitor to HDAC6, a Class II HDACs. Moreover, two of the confirmed hits are marketed drugs which could potentially expedite their development as anticancer drugs acting via HDAC1 inhibition. This study illustrates that validated QSAR models have the ability of identifying novel structurally diverse hits by the means of virtual screening. We believe that the technology described in this study could be used for data analysis and hypothesis generation in many computational drug discovery studies.

Tables and Figures
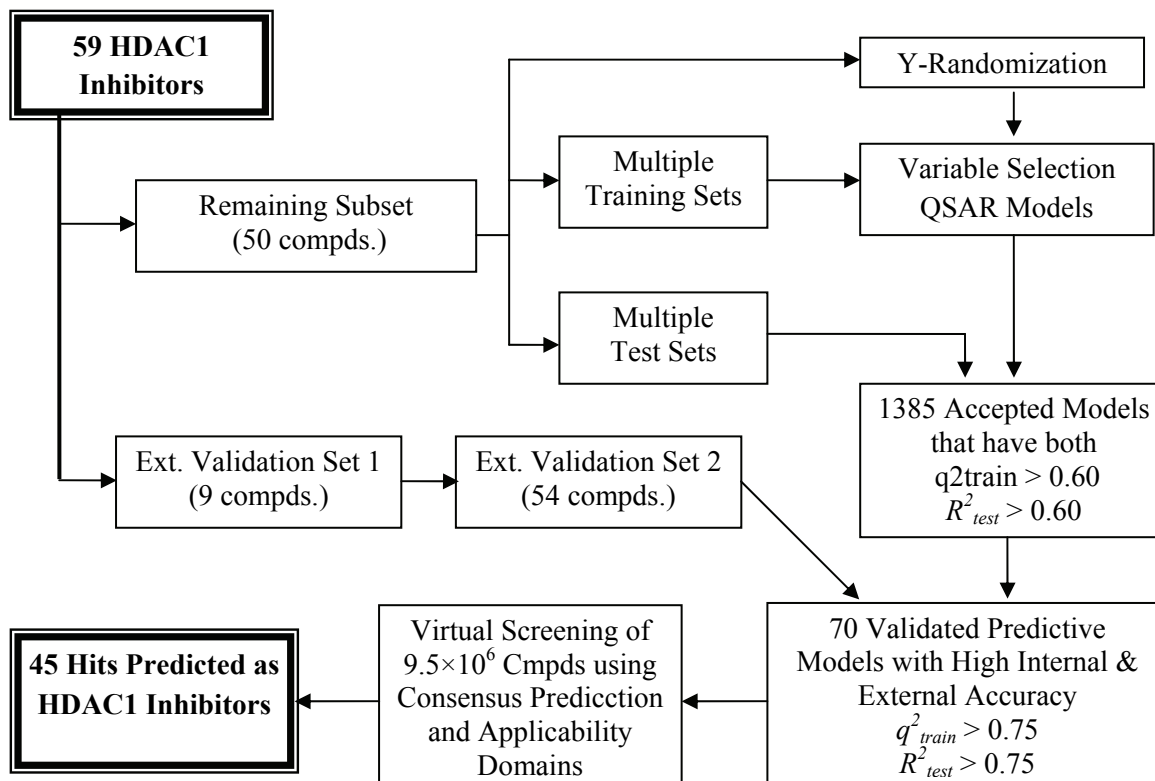
Figure 2.1 The workflow of QSAR model building, validation and virtual screening as applied to HDAC1 inhibitors. The specific data for kNN/MolConnZ modeling are used for illustration purpose.
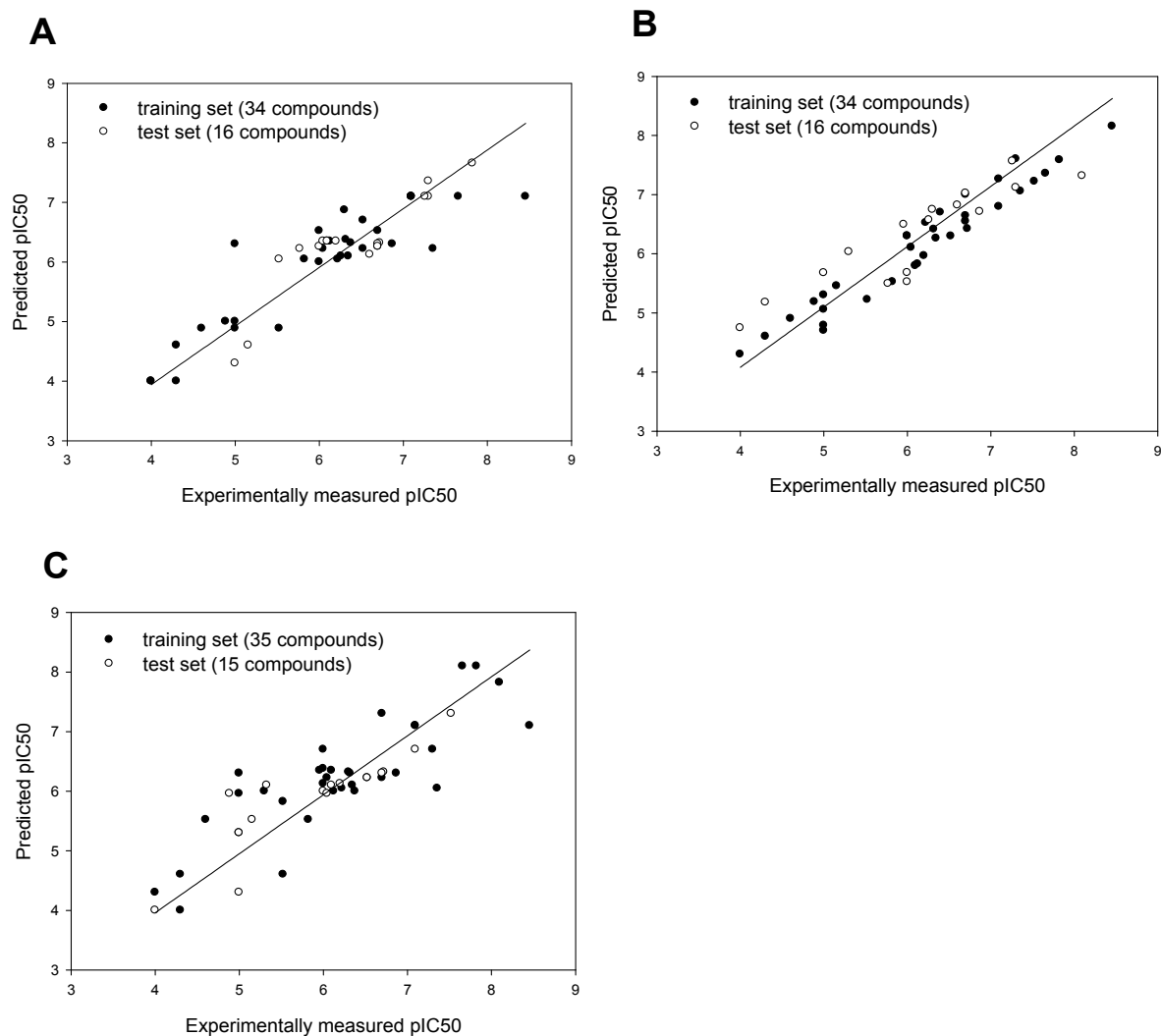
Figure 2.2 Comparison of actual vs. predicted inhibition efficiency (*pIC50*) values for the best QSAR model for each combination of statistical modeling approach and descriptor type. A. For kNN/MolConnZ method ($q^2 = 0.81$, $R^2 = 0.80$). The training set contains 34 compounds (dark circles) and test set contains 16 compounds (empty circles). B. For SVM/MolConnZ method ($q^2 = 0.94$, $R^2 = 0.81$). The training set contains 34 compounds (dark circles) and test set contains 16 compounds (empty circles). C. For kNN/MOE models ($q^2 = 0.70$, $R^2 = 0.76$). The training set contains 35 compounds (dark circles) and test set contains 15 compounds (empty circles).
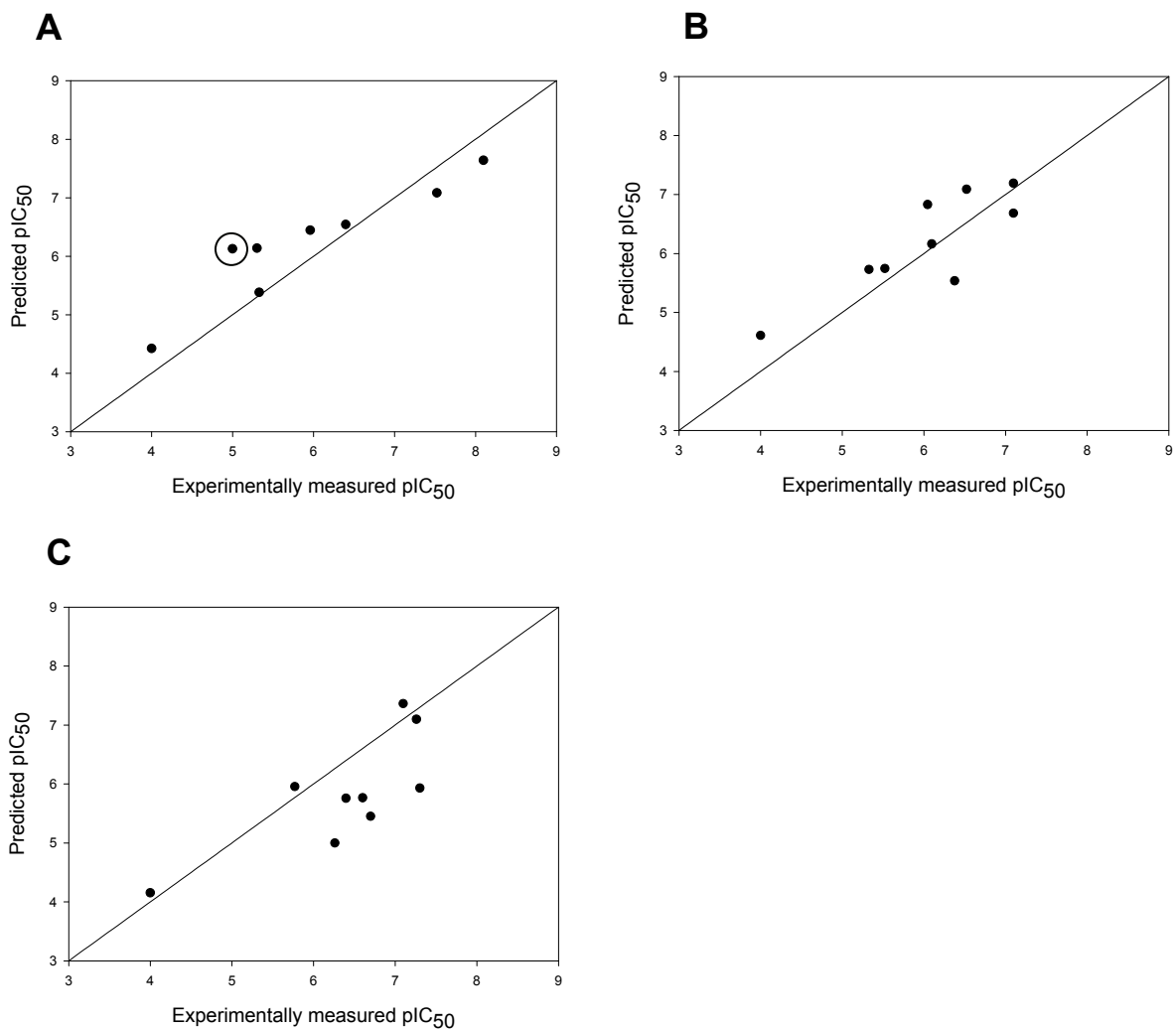
Figure 2.3. Comparison of actual *vs*. predicted inhibition efficiency (*pIC50*) values for the best QSAR model as applied to the external validation set 1. **A.** For the *k*NN/MolConnZ method ($R^2 = 0.87$, 8 compounds). The compound with the black circle is the possible structural outlier that has been discussed in the results. **B.** For the SVM/MolConnZ method ($R^2 = 0.71$, 9 compounds). **C.** For the *k*NN/MOE method ($R^2 = 0.60$, 9 compounds).
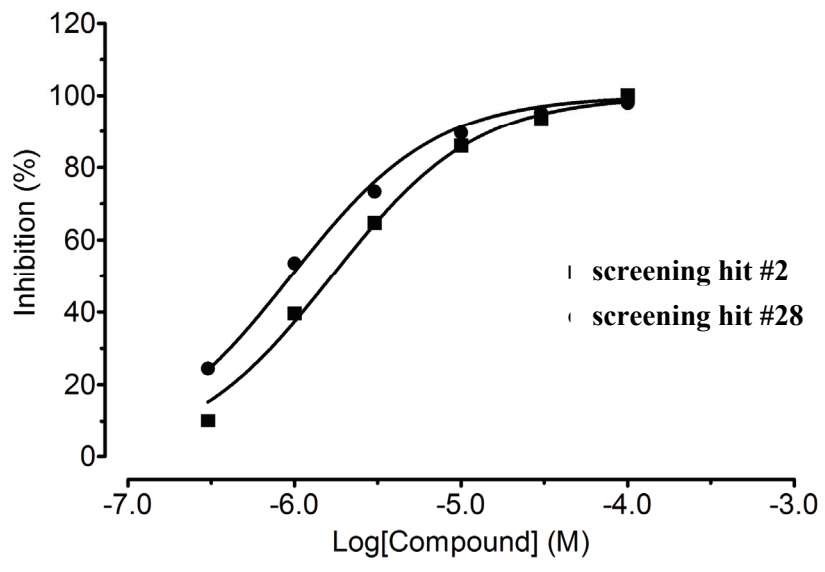
Figure 2.4. The full dose response curve for hit compounds #2 and #28 in human HDAC1 inhibition assay.

Table 2.1. Summary of Previous QSAR Studies of HDACs Inhibitors.

Group	Year	HDAC Types	Dataset Size	Chemical Class	Model Type	Model Validation
Chen et al.[139]	2008	HDAC1	Training set (30), Test set (25)	Mostly hydroxamates	3D chemical-feature-based QSAR pharmacophore model by Hypogen	Training set R^2=0.924, test set R^2=0.896
Kozikowski et al.[140]	2008	HDAC1-3, HDAC8; HDAC6, HDAC10;	20-23	Biphenyl or phenylthiazoles analogues bearing hydroxamates or mercaptoacetamides	Linear regression	Best model with training set R^2=0.943
Ragno et al.[141]	2008	Class II HDAC, Maize HDAC1-A, HDAC1-B	25	(aryloxopropenyl) pyrrolyl hydroxamates	GRIND[a] 3-D QSAR, PLS[b]	R^2/q^2 values of 0.96/0.81 and 0.98/0.85 for HDAC1-B and HDAC1-A Randomized variable selection
Juvale et al.[75]	2006	HDAC1	Training set (40), Test set (17)	Hydroxamates	CoMFA[c], CoMSIA[d]	Best CoMFA training set R^2=0.910, LOO-CVR^2=0.502, test set R^2=0.327 Best CoMSIA training set R^2=0.987, LOO-CV R^2=0.534, test set R^2=0.464 Y randomization
Ragno et al.[85]	2006	Maize HDAC2	101		GRID[e]/GOLPE[f], PLS	R^2, cross validated q^2, and cross-validated SDEP (RMSE) values of 0.94, 0.83, and 0.41
Guo et al.[142]	2005	HDAC1	Training set (25), Test set (4)	Indole amide analogues	CoMFA, CoMSIA	CoMFA R^2=0.982, LOO-CV R^2=0.601 CoMSIA R^2=0.954, LOO-CV R^2 =0.598
Wagh et al.[143]	2006	HDACs	Training set (39), Test (17)	Hydroxamic acid analogues	3D-QSAR by GFA[g]	Training set q^2=0.712, R^2=0.585
Wang et al.[131]	2004	PC-3 cell line	19	TSA and SAHA like hydroxamic acid	Stepwise multiple linear regression	LOO-CV R^2 = 0.870
Xie et al.[144]	2004	HDAC1, 4, 6	124	Hydroxamic acid, short chain	PLS,	Best model with R^2 of

| maize HDAC2 | fatty acid, cyclic tetrapeptides, cyclic peptides without the AOE moiety, benzamides | Binary classification | 0.76 , and LOO-CV R^2 of 0.73 LOO-CV accuracy for binary model is 0.92 |

[a]GRIND: Grid-independent Descriptors; [b]PLS: Partial Least Squares regression; [c]CoMFA: Comparative Molecular Field Analysis; [d]CoMSIA: Comparative Molecular Similarity Indices Analysis; [e]GRID: GRID force field; [f]GOLPE: Generating of Optimal Linear PLS Estimations; [g]GFA: Genetic Function Approximation

Table 2.2. The Statistics for Ten Best kNN Models for All Test Sets Using MolConnZ Descriptors.

Model No.	Training Set Size	Test Set Size	Descriptor No.	Nearest Neighbor No.	q^2 (Training Set)	R^2 (Test Set)	R_0^2 (Test Set)	RMSE (Test Set)
1	45	5	22	1	0.80	0.87	0.69	0.27
2	41	9	20	2	0.80	0.81	0.77	0.49
3	34	16	14	1	0.81	0.80	0.76	0.38
4	35	15	12	2	0.82	0.79	0.70	0.48
5	42	8	14	1	0.81	0.79	0.73	0.35
6	34	16	26	1	0.80	0.79	0.78	0.37
7	28	22	36	1	0.83	0.77	0.67	0.42
8	40	10	12	2	0.81	0.77	0.77	0.43
9	29	21	20	1	0.79	0.77	0.76	0.47
10	34	16	16	1	0.79	0.76	0.74	0.40

Table 2.3. The Statistics for Ten Best kNN Models for All Test Sets Using MOE Descriptors.

Model No.	Training Set Size	Test Set Size	Descriptor No.	Nearest Neighbor No.	q^2 (Training Set)	R^2 (Test Set)	R_0^2 (Test Set)	RMSE (Test Set)
1	35	15	14	1	0.70	0.76	0.76	0.45
2	35	15	12	1	0.73	0.75	0.75	0.46
3	35	15	30	1	0.73	0.75	0.75	0.47
4	32	18	18	1	0.81	0.74	0.63	0.49
5	35	15	16	1	0.72	0.74	0.74	0.48
6	35	15	18	1	0.71	0.74	0.73	0.48
7	35	15	24	1	0.72	0.73	0.73	0.48
8	31	19	22	1	0.70	0.72	0.69	0.53
9	35	15	14	1	0.77	0.71	0.70	0.51
10	28	22	12	1	0.79	0.70	0.70	0.51

Table 2.4. The Statistics for Ten Best SVM Models for All Test Sets Using MolConnZ Descriptors.

Model No.	Training Set Size	Test Set Size	C	E	q^2 (Training Set)	R^2 (Test Set)	R_0^2 (Test Set)	RMSE (Test Set)
1	37	13	200	0.40	0.93	0.87	0.62	0.36
2	37	13	200	0.50	0.91	0.86	0.66	0.34
3	37	13	200	0.35	0.94	0.85	0.59	0.38
4	37	13	200	0.55	0.90	0.85	0.67	0.34
5	37	13	200	0.60	0.89	0.84	0.68	0.33
6	34	16	200	0.30	0.94	0.81	0.76	0.51
7	35	15	200	0.30	0.95	0.72	0.66	0.49
8	39	11	200	0.30	0.94	0.72	0.71	0.51
9	29	21	200	0.30	0.96	0.71	0.66	0.49
10	35	15	200	0.35	0.94	0.71	0.66	0.49

Table 2.5. Summary of Combinatorial QSAR Modeling for the HDAC1 Inhibitor Modeling Set and the External Validation Set 1 (EV1).

Type of Descriptor	kNN				SVM			
	q^2 (CV)	R^2	R_0^2	RMSE	q^2 (CV)	R^2	R_0^2	RMSE
MZ4.05	0.80	0.87	0.69	0.27	0.93	0.87	0.62	0.36
		0.87(EV1)	0.78(EV1)	0.59(EV1)		0.71(EV1)	0.68(EV1)	0.52(EV1)
MOE2006.08	0.70	0.76	0.76	0.45	0.91	0.64	0.53	0.54
		0.60(EV1)	0.25(EV1)	0.84(EV1)		N/A	N/A	N/A

Table 2.6. Consensus Predictions of Inhibition Efficacy for the External Validation Set 2 by kNN/MolConnZ and SVM/MolConnZ Models.

Comp. ID*	SVM/MolConnZ			kNN/MolConnZ			Consens. Value	Exp. pIC_{50}
	Models Cov.	Pred. Value	Std. Dev.	Models Cov.	Pred. Value	Std. Dev.		
16c_AE	17/17	6.14	0.15	32/70	6.31	0.77	6.26	6.96
17a_AE	17/17	6.50	0.18	42/70	6.65	0.69	6.61	6.47
17b_AE	17/17	6.60	0.15	42/70	6.78	0.48	6.73	6.68
17d_AE	17/17	6.80	0.15	41/70	6.81	0.49	6.81	8.40
17e_AE	17/17	6.75	0.19	34/70	6.81	0.58	6.79	8.05
17f_AE	17/17	7.36	0.20	33/70	7.06	0.64	7.16	7.74
17g_AE	17/17	7.24	0.24	32/70	7.29	0.81	7.27	8.40
17h_AE	17/17	7.11	0.26	24/70	7.24	0.72	7.18	8.05
17i_AE	17/17	7.68	0.26	31/70	7.25	0.82	7.40	8.40
17j_AE	17/17	7.15	0.24	15/70	6.62	0.83	6.91	8.40
1g_AE	17/17	6.38	0.28	56/70	6.75	0.92	6.66	6.26
1h_AE	17/17	6.56	0.22	57/70	6.99	0.66	6.89	7.77
1i_AE	17/17	6.51	0.21	59/70	7.03	0.37	6.91	6.75
1j_AE	17/17	6.57	0.17	57/70	6.90	0.70	6.83	7.51
1k_AE	17/17	6.75	0.24	39/70	6.96	1.01	6.90	8.22
1l_AE	17/17	6.39	0.22	53/70	6.67	0.94	6.60	6.26
6d_AE	17/17	6.68	0.17	66/70	7.14	0.31	7.05	7.64
6e_AE	17/17	6.66	0.16	65/70	7.16	0.30	7.06	8.52
6f_AE	17/17	7.06	0.18	64/70	7.66	0.43	7.53	8.40
6g_AE	17/17	7.04	0.16	58/70	7.18	0.41	7.15	7.66
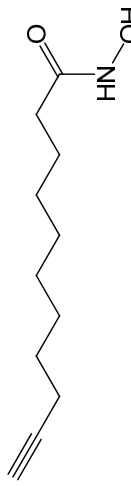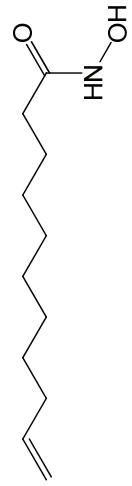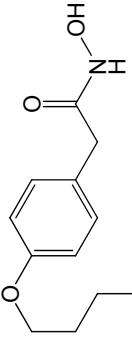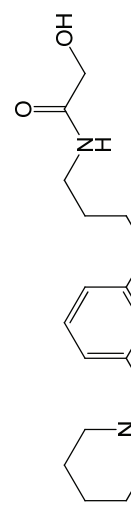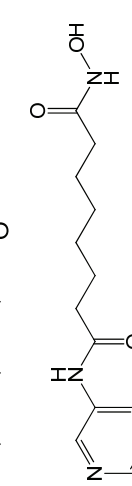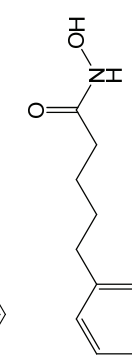
6h_AE	17/17	6.65	0.15	58/70	7.19	0.41	7.07	8.30
6i_AE	17/17	6.67	0.18	61/70	7.64	0.43	7.43	8.40
7c_AE	17/17	7.01	0.21	58/70	7.08	0.27	7.06	6.28
7d_AE	17/17	7.31	0.18	56/70	7.17	0.39	7.20	7.82
7e_AE	17/17	6.86	0.17	55/70	7.16	0.40	7.09	7.52
7f_AE	17/17	7.33	0.19	63/70	7.20	0.41	7.23	7.42
7g_AE	17/17	7.04	0.16	56/70	7.24	0.48	7.19	7.25
7h_AE	17/17	7.18	0.20	47/70	7.17	0.33	7.18	8.40
9a_AE	17/17	7.10	0.21	45/70	7.43	0.55	7.34	6.59
9b_AE	17/17	7.27	0.26	47/70	7.50	0.59	7.44	6.19
9d_AE	17/17	6.97	0.16	42/70	7.61	0.71	7.43	6.35
9e_AE	17/17	6.94	0.32	36/70	7.21	0.46	7.12	7.01
9g_AE	17/17	6.83	0.21	27/70	7.09	0.66	6.99	7.82
9h_AE	17/17	7.58	0.31	28/70	7.45	0.88	7.50	8.22
2_	17/17	6.50	0.23	62/70	7.28	0.57	7.11	7.36
3_	17/17	6.50	0.20	57/70	6.66	0.38	6.62	7.00
5_	17/17	7.11	0.16	48/70	7.36	0.45	7.30	8.70
6_	17/17	5.93	0.16	1/70	7.02	N/A	5.99	5.32
10_	17/17	6.14	0.15	32/70	6.31	0.77	6.26	6.96
12_	17/17	6.37	0.18	48/70	6.07	0.21	6.15	6.68
13_	17/17	7.44	0.29	70/70	7.12	0.14	7.18	6.70
RMSE		0.92			0.86		0.86	

*Only 41 out of the total of 54 compounds are listed. The other 13 compounds were found to be out of applicability domain of QSAR models.

Table 2.7Consensus Predictions for the External Validation Set 1 by A) kNN/MolConnZ B) SVM/MolConnZ, and C) kNN/MOE models

A.

Comp.ID	Experimentally measured	Total models used	Consensus prediction	std. of prediction
BC-4-86	5.33	70/70	5.38	0.65
BC-6-38	6.40	70/70	6.54	0.29
BC-2-48	4.00	70/70	4.42	0.25
YChdac045	7.52	67/70	7.08	0.22
Yc-II-88	8.10	70/70	7.64	0.35
BC-4-56	5.96	70/70	6.44	0.27
BC-4-4	5.30	69/70	6.14	0.30
BC-4-2	5.00	50/70	6.13	0.12
BC-2-83*	6.00	N/A	N/A	N/A
RMSE	0.59			
R^2	0.87			
R_0^2	0.78			

*This compound was found to be out of applicability domain of kNN/MolConnZ models.

B.

Comp.ID	Experimentally measured	Total models used	Consensus prediction	std. of prediction
Ag-b-57	7.10	17/17	7.19	0.20
AG-biph-38	7.10	17/17	6.68	0.14
AG-biph-40	5.52	17/17	5.74	0.10
BC-2-45	4.00	17/17	4.61	0.13
BC-3-22	6.38	17/17	5.54	0.19
BC-4-84	6.10	17/17	6.16	0.14
BC-4-86	5.33	17/17	5.73	0.13
BC-5-44	6.05	17/17	6.83	0.19
YC-03065	6.52	17/17	7.09	0.15
RMSE	0.52			
R^2	0.71			
R_0^2	0.68			

C.

Comp.ID	Experimentally measured	Total models used	Consensus prediction	std. of prediction
AG-biph-08	7.10	13/13	7.36	0.27
Yc-II-84	7.26	13/13	7.10	0.00
BC-1-30-2	4.00	13/13	4.15	0.38
BC-4-54	6.26	13/13	5.00	0.41
BC-4-55	6.70	13/13	5.45	0.42
BC-4-93	5.77	13/13	5.96	0.45
BC-6-26	7.30	13/13	5.93	0.67
BC-6-38	6.40	12/13	5.76	0.57
BC-6-40	6.60	13/13	5.77	0.54
RMSE	0.84			
R^2	0.60			
R_0^2	0.25			

Table 2.8. Consensus Predictions of Inhibition Efficacy for the Confirmed Screening Hits identified by kNN/MolConnZ and SVM/MolConnZ Models.

Mol.	Compd. #	CAS #	kNN/MolConnZ			SVM/MolConnZ			Consens. Value	HDAC1. pIC50	HDAC6 Inhibition*
			Model Cov.	Pred. Value	Std. Dev.	Model Cov.	Pred. Value	Std. Dev.			
	2	123041-56-5	37/70	7.00	0.46	17/17	6.28	0.30	6.77	5.90	105%
	28	16791-35-8	36/70	7.04	0.53	17/17	6.49	0.29	6.86	6.00	101%
	35	2438-72-4	46/70	7.15	0.46	17/17	6.67	0.26	7.02	4.00	99%
	47	78273-80-0	36/70	6.68	0.67	17/17	5.94	0.19	6.44	<4.00	42.6%
	34	382180-17-8	70/70	7.12	0.35	17/17	7.60	0.27	7.21	6.70[132]	
	40	114918-01-3	42/70	7.12	0.47	17/17	6.59	0.27	6.97	7.62[131]	
RMSE			1.59			1.05			1.50		

*The inhibition assay against HDAC6 at the single concentration of 30 µM.

Chapter 3. Comparative Studies of the Structural Models versus the X-ray Crystal Structures of Human β2 Adrenergic GPCR in terms of applications to virtual screening

(This chapter has been submitted to *Proc Natl Acad Sci)*

3.1. Introduction

Structure based drug discovery (SBDD) has become a major strategy in identifying novel leads for important biological targets. SBDD enabled well documented discovery of several approved drugs, e.g. dorzolamide and imatinib. Typically, the 3D structures of biomolecules obtained by the means of X-ray crystallography and NMR spectroscopy are needed for the purpose of virtual screening (VS), whose performance is strongly affected by the quality of biomolecular structure, especially with respect to binding site description. When no experimentally determined structures are available, theoretical models based on either homology or *de novo* modeling approaches are employed instead [35-38]. However, there were some debates about the accuracy and applicability of theoretical models [38-44] in SBDD. In principle, the success of homology modeling is typically measured by how close the models could reproduce experimental structures, which implies that the latter are inherently more appropriate choice for SBDD applications.

G-protein coupled receptors (GPCRs) is a protein family where theoretical models have been used widely because of GPCRs' importance as targets for many drugs, on one hand, and the lack of experimental structures, until recently, on the other hand. The experimental effort to characterize the 3D structure of GPCRs were seriously hindered by membrane related issues [176]. A large number of theoretical models have been employed in

the past decades [177-180] for VS often yielding reasonable results [181-184]. However, it was fully expected that the availability of the experimental structure of any GPCR would substantially enhance the efficacy of SBDD efforts. Thus, the recent publications of the crystal structure of human beta-2 adrenergic GPCR (β2AR)[45-47] cleared the way for the validation of previous theoretical models, as well as provided critical data for building homology based models of other GPCRs as the most obvious structural template.

It has been shown that human β2AR features a structurally conserved rhodopsin-like 7TM core, but there exist novel structural features that had not been identified previously. It remained unclear as to whether these structural divergences would affect the outcome of VS studies. Dissimilarity of computational models relative to the crystal structure should lead to relatively poorer performance of the former in docking and scoring of known ligands; however, this general expectation should not necessarily be regarded as a law. One should take into account that some of the computational models are actually manually refined with known medicinal chemistry data and therefore, there is at least a possibility that theoretical models may be even more suitable for drug discovery by VS than the crystal structure.

In this study, we have addressed this, both scientifically and pragmatically, important question directly. We have compared the x-ray structure of β2AR vs. several previously built theoretical models in terms of their respective ability to recover known β2AR ligands (both agonists and antagonists) from a large external compound library in VS experiments. None of these models were generated in our group to ensure objective and unbiased comparisons. Furthermore, although our group has developed both scoring functions [185] and virtual screening protocols [186], for the same reasons we restricted ourselves to using several popular commercial docking and scoring tools developed elsewhere. Thus, by design, this study

lacked any user biases concerning preferred homology models or most familiar computational tools to emulate the situation that is most commonly faced by the majority of molecular modeling practitioners both in academia and industry.

There have been previous studies on comparing homology models among themselves or even with the crystal structure but to the best of our knowledge nobody asked the question poised herein as directly as we did. For instance, Bissantz et al [40] employed three human GPCRs models derived from rhodopsin for virtual screening using multiple docking programs and scoring functions. Their work proved that homology models are suitable for VS but there was no comparison to the crystal structure. After the crystal structure of human β2AR was published, Costanzi reported a pioneering study [187] where carazolol was docked both into two rhodopsin-based homology models of human β2AR as well as into the X-ray structure. The models afforded high accuracy of the docking poses, especially after incorporating the biochemical data to adjust the orientation of the binding pocket residues. However, this study was limited to the pose prediction only using in-house models. Most recently, Fan et al [39] reported that for 27 out of the 38 protein targets, the consensus enrichment for multiple homology models was better than or comparable to that of both the holo- and the apo- X-ray structures. However, that study was focused on soluble protein targets and used a single homology model building tool that employed x-ray characterized structural templates. In contrast, all models employed herein were built before the β2AR crystal structure became available as possible template.

We have carried out a systematic study on a large collection of published human β2AR theoretical models and evaluated their structural accuracies and virtual screening performances in comparison with two crystal structures, i.e., 2RH1 (released by the RCSB

Protein Data Bank (PDB) on Oct. 30, 2007) and the latest, 3D4S (released by RCSB PDB on Jun. 17, 2008). Two other structures, 2R4R and 2R4S, were not employed because they came from the same source as 2RH1 but with lower resolutions. As shown in Table 3.1, we collected eight independently published theoretical models of human β2AR including both apo and holo structures. Both agonist and antagonist bound models were included to account for any structural features associated with functional activity.

Surprisingly, we found that some of the theoretical models displayed better VS performances than the x-ray structure. This study by no means undermines the extreme significance of the x-ray structure of β2AR as well as other GPCRs [45,188-192] in understanding the intricate details of GPCR structure in relation to its function nor in the significance of x-ray structures for SBDD. Nevertheless, it most certainly testifies to the importance of intelligent homology modeling approaches especially those incorporating comprehensive medicinal chemistry knowledge of receptor ligands for structure based virtual screening.

3.2. Methods
3.2.1. Structural Similarity Analysis.

In addition to 2RH1, the crystal structure of bovine rhodopsin (PDB ID: 1U19) chain B was also included in the analysis since it was used as major template for all β2AR homology models. The structural similarity was assessed in three aspects, i.e. the accuracy of the boundary definitions for each transmembrane (TM) helixes, the backbone root-mean-square-deviation (RMSD) for TM regions, and the Cα RMSD of the binding pocket residues. The numbering of amino acids followed the conventions set by Weinstein et al [193]. The highly conserved residues embedded in each TM region were used as anchors for the alignment. Each theoretical model was structurally aligned against 2RH1 or 1U19 by

individual TM helix as well as the whole TM bundle. The RMSDs were calculated using the entire lengths of the corresponding segments of 2RH1.

3.2.2. Molecular Docking Calculation.

To compile the screening database, we used 58 known antagonists of human β2AR reported in DrugBank [194] and GLIDA[195] databases as seeds (cf. Table 3.3 of Supporting Information). All antagonists have sub-micromolar potency and can be found in additional external databases, such as PDSP Ki [196], PubChem (PubChem, 2007) and KEGG [197]. A few of them are β2AR specific (e.g. butoxamine and aminoflisopolol) while others can act on both β1AR and β2AR. For comparison, thirteen agonists of human β2AR found in the DrugBank database were used in virtual screening experiments as well. In order to establish a reasonable decoy dataset for virtual screening, the World Drug Index (WDI) database [98] (version 2004) was used since most its entries are drug-like compounds. The original collection of 59,000 molecules was first cleaned by removing metals, salts and fragments, then filtered to eliminate unqualified compounds according to Lipinski's rule of five and later extensions of this rule [198]. The remaining collection of ca. 38,000 compounds was further reduced to a diverse subset of 374 compounds using MOE2007.09. In addition, 12 binding decoys with similar chemical scaffolds but poor binding affinity (Ki > 10μM) were selected from the PDSP Ki database and merged into the WDI diverse subset. In the end, 58 human β2AR antagonists and 13 agonists were seeded amongst 386 decoys separately to constitute two different screening databases.

We employed three popular docking programs, i.e. Glide4.01, AutoDock4.0 and eHiTS6.2, to evaluate the screening performance of structural models. The calculations by Glide version 4.01 [199] was carried out using Schrodinger Suite 2007. The targeted protein and

theoretical models were prepared through Protein Preparation module with the default setting and assigned with the OPLS 2001 force field atom types and partial charges. The screening databases were prepared within the LigPrep module and the ionization states of each molecule were calculated as to be compatible with the pH value of 7.0±2.0. All molecules were subjected to energy minimization with MMFFs force field before the docking computation. For x-ray structure and holo models, the center of the grid box was selected as the center of bound ligands. For apo models, their binding pockets were first aligned to that of 2RH1 and the center of co-crystallized carazolol was chosen. The proper size of the enclosing box was not set to be fixed but determined by the extent of the bound ligand. The Glide SP scoring function was used to rank the docking poses and the top-ranked poses for each database molecule were saved for post-docking analysis.

We prepared the targeted protein and docking parameters for AutoDock version 4.0 [200,201] using the AutoDockTools graphic interface. Explicit hydrogen atoms were added to the receptor structures while atom types and partial charges were assigned to generate the pdbqt receptor files. The database molecules were prepared using the 'prepare_ligand4.py' script to merge non-polar hydrogen atoms and define flexible torsions. The center and dimension of the enclosing boxes were defined to include the whole binding pockets, similar to those in the Glide docking. The genetic algorithm were employed during the docking with a start population size of 150 individuals and 20 runs combined with a maximum number of 12,500,000 energy evaluation for each molecule. Other parameters for genetic algorithm were kept by the default value.

The eHiTS version 6.2 [202] was used through the CheVi user interface. Protein preparations, such as protonation state determination for residues, hydrogen atoms addition

and partial charge assignment, were actually not needed since eHiTS' docking and scoring are based on the prior training data of its knowledge base of 97 protein families. The grid box was assigned automatically using the bound ligand's SDF file as the CLIP file. We employed the default settings for eHiTS docking and ranked the database molecule based on its lowest eHiTS score.

In addition, we added eight more scoring functions of different types by rescoring the top-ranked poses generated by Glide4.01. The multiple scoring functions in Sybyl8.0 [203] CScore module and OpenEye [204] FRED 2.2.4 were applied, including Chemscore, D_score, Gold_score, PMF, Chemgauss3, PLP, Screenscore and Shapegauss. The consensus scores were also used for the above scoring functions through the rank by rank strategy.

3.2.3. Assessment of Virtual Screening Performance

To measure the efficiency of virtual screening we used the following conventional parameters: the enrichment factor and the receiver operating characteristic (ROC) curve that characterizes the ability if a method to recover known ligands among the top-scored screening molecules. The enrichment factor follows the most popular definition as to how many more seed compounds (i.e., known ligands) were found within a defined "early recognition" fraction of the ranked list relative to a random distribution:

$$EF = H_{scr}/H_{tot} \times D_{tot}/D_{scr} \qquad (1)$$

where H_{scr} is the number of target-specific seeds recovered at a specific % level of the database; H_{tot} is the total number of seeds for the target; D_{scr} is the number of compounds screened at a specific % level of the database; D_{tot} is the total number of compounds of the database. The ROC curve is generated by plotting the sensitivity (Se) vs. $(1 - $ specificity (Sp)$)$ for a binary classifier system as its discrimination threshold is varied. In

the case of virtual screening for recovering the ith known active from the inactive decoys, the Se_i and Sp_i are defined as follows:

$$Se = H_{scr}/H_{tot} \qquad\qquad (2)$$

$$Sp = \frac{(D_{tot}-H_{tot})-(D_{scr}-H_{scr})}{D_{tot}-H_{tot}} \qquad (3)$$

The area under the ROC curve (AUC) is the metric that is widely accepted for assessing the likelihood that a screening method assigns a higher rank to known actives than to inactive compounds. The AUC values at a specific percentage of the ranked database are calculated from the following equation:

$$AUC = \sum_{i=1}^{n}[Se_i(Sp_{i+1} - Sp_i)] \qquad (4)$$

Here n is the total number of known actives in the screening database. One additional parameter, the yield, is also employed as the percentage of true hits retrieved by the virtual screening method:

$$Yield = H_{scr}/H_{tot} \times 100 \qquad\qquad (5)$$

Cluster Analysis of Binding Profiles.

To closely evaluate the key receptor/ligand interaction patterns, we employed the LigX module [205] in MOE2007.09 to analyze the crystal structure of β2AR/carazolol complex and the docking poses generated by Glide4.01. For each antagonist, the top-ranked docking pose with the highest score was selected. Two major types of interactions that contribute to protein/ligand binding affinities were considered, i.e. hydrogen bonds (donor or acceptor) and non-bonded weak interactions. The score to assess the hydrogen bond is based on a scale of 0 to 100% that indicates the probability of being a geometrically perfect hydrogen bond while the score for non-bonded weak interaction is the pair wise distance between residue and ligand atoms. In our studies, we took the default parameters in which 4.5 angstrom is the

cutoff for weak interactions and 2.5 angstrom is the closest distance between any residue/ligand atom pairs. The original score was normalized; thus, the values of the modified scores were between 0 and 1, which is proportional to the interaction intensities. To better visualize the binding patterns of docked poses for each theoretical models and crystal structures, the LigX scores were transformed into heat maps and clustered using the R statistical package [206]. We applied the hierarchical clustering with the Ward linkage algorithm; thus, the patterns of interaction between 58 human β2AR antagonists and residues in the active sites of three different structural models would be expected to be similar if the respective clusters are similar (cf. Figure 3.3).

In addition, we have exploited the Protein Ligand Interaction Fingerprints (PLIF, also available in MOE2007.09) for the same purpose (cf. Figure 3.6 of Supporting Information). PLIF can identify and score major protein/ligand interactions, including hydrogen acceptor from side chain, hydrogen donor to side chain, hydrogen acceptor from backbone, hydrogen donor to backbone, ionic attractions and surface contacts. For each docking pose, the PLIF fingerprints ranging from 30 to 50 bits were generated. The relative frequencies of each identified fingerprint can be then used to produce fingerprint significance chart, which is based on the hypothesis that 'if the bit is set, then the compound is active'.

3.3. Results and discussion

3.3.1. Comparison of Theoretical Models and the X-ray Structure of β2AR for Their Overall Structures and Protein Segments That are Critical for Ligand Binding

Prior to VS experiments, we analyzed the similarity between theoretical models and two x-ray structures, i.e. 2RH1 and 1U19 deposited to the Protein Data Bank [207]. All theoretical models used in this study are listed in Table 3.1; each model was aligned against

54

the x-ray structures to evaluate relative definitions of transmembrane (TM) helices, their conformations, and relative orientation. The x-ray structure of bovine rhodopsin (1U19) was also included in addition to β2AR because the former had been used as a common template for GPCR homology modeling. In doing so, we were interested to explore if the failure to predict the structural conservation and/or divergence from the template structure may cause poor VS results for the theoretical models.

Our initial efforts focused on evaluating the accuracy of TM helical boundaries as defined by models vs. x-ray structures. For this purpose we have employed MOE 2007.09 software [208] to annotate the secondary structural elements in the 3D structures. We found (cf. Table 3.3 of Supporting Information) that the MOE module assigned boundaries either at exact positions or only one amino acid apart to over 90% of the TM segments of 2RH1 and 1U19 in comparison to those in the PDB header. Thus, the MOE software was deemed reliable in identifying the helical boundaries; the results of applying MOE to the six theoretical models and two crystal structures (2RH1 and 1U19) are summarized in Figure 3.1a. From the alignment of the eight structures, it can be seen that the apo models (AM1-AM3) perform better than the holo models (CM1-CM3) in terms of accuracy of TM assignment. The location and length of the TM helixes for all three AM models are consistently close to those in the rhodopsin structure (1U19), with the only exception that TM6 and TM7 in the AM1 model are shorter than the corresponding helices in the crystal structures. This observation can be easily rationalized since AM1 and AM2 models are solely based on homology modeling whereas AM3 is a hybrid model developed with a combination of both threading and *ab initio* methods. In all cases, the crystal structures of bovine rhodopsin were used as a template for model building. Furthermore, we have

concluded that individual TM helixes were very close in terms of helical length and relative orientation when compared to crystal structures of bovine rhodopsin and human β2AR [45]. Thus, not surprisingly, given the methods used for model building, the secondary structural elements for all three AM models were found to be assigned very accurately as compared to their homologous experimental structures.

The accuracies of TM helix boundaries assignments for CM models were less satisfactory. In general, seven TM helices in all three CM models were shorter than expected with the largest disagreements located at TM1, TM4 and TM5 (Figure 3.1a). The CM3 model gave the largest deviation in terms of the percentage of correctly defined TM helical boundaries. Moreover, it had the shortest lengths for individual TM helices. For instance, it was eight residues shorter for TM1, fourteen for TM3 and eleven for both TM5 and TM6 in comparison to the β2AR crystal structure. One possible explanation is that all three CMs are *de novo* models, generated without any template structure. Both the Lybrand (CM1, 2) and the Goddard (CM3, 4) groups employed the standard alpha helix as a starting point and calculated the intrinsic tilt/kink and relative orientation of the TM helical bundle purely based on the physical considerations. If the rhodopsin structure is not employed as a reference, the secondary structure assignments could be affected by many factors, such as the type of phospholipids used in the MD simulations employed as part of model refinement in studies by the Goddard group [177,209].

It should be noted that the engineered modification of the wild type protein using a segment of T4 lysozyme to replace most residues of IL3 introduced an artifact in the crystal structure (2RH1) of human β2AR. This modification led to altering the boundaries of IL3, thus affecting the correct locations of both the TM5 terminus and the start of the TM6. The

accurate definition of these two boundaries is less important in the comparison of structural similarities between theoretical models and the crystal structure. However, the accuracy of predicting the TM4 terminus and the beginning of TM5 is critical considering the functional roles of EL2 in both rhodopsin and human β2AR [210,211]. As can be seen from Figure 3.1a and Table 3.4 of the Supporting Information among the three AM models, AM3 has the highest accuracy (1 residue error) for the segment between TM4 and TM5 (EL2) followed by AM2 and AM1. CM3 model is comparable to AM2 while CM1 and CM2 had much larger errors with respect to TM helix assignments.

The seven TM helices of each theoretical model were superimposed onto respective helices of β2AR as well as the rhodopsin structure, and the backbone pair wise RMSD of individual respective TM helices was calculated (Figures 3.1b, 3.1c and Table 3.4 of Supporting Information). As expected, the homology models (AMs) are generally more similar to the rhodopsin structure than to the β2AR structure. The RMSDs of most helices in AM1-3 range from 0.30 Å to 1.00 Å as compared to bovine rhodopsin where the RMSDs are as big as 1.60 Å to 3.80 Å when aligned against the human β2AR. For the whole TMs bundle, the RMSDs are 1.15 Å to 1.88 Å with respect to rhodopsin and 2.25 Å to 3.19 Å with respect to β2AR. In comparison, the *de novo* models (CMs) deviate more significantly from both crystal structures. The RMSDs of most helixes in CM1/CM2 are in the same range of 2.00 Å - 5.50 Å when aligned against both rhodopsin and β2AR. For CM3, the RMSD increases to 2.60 Å and 7.00 Å, respectively. Similarly, the RMSDs of TMs bundle for CM1 are 3.20 Å when aligned against the rhodopsin and 3.59 Å for β2AR. For CM3, the corresponding RMSDs are 3.83 Å and 4.41 Å (cf. Table 3.4 of Supporting Information).

Consistently, the AMs models were observed to be aligned well to the rhodopsin structure (cf. AM2 in Figure 3.1b as an illustrative example). They deviated from the β2AR structure with the noticeable shifts for TM1, TM3 and TM5, although the secondary structure assignment for these regions was relatively accurate. The large RMSD differences for these three TMHs were obviously due to the differences between rhodopsin and β2AR crystal structures (7TMs backbone RMSD of 2RH1 vs. 1U19 is 1.85 Å). It was indeed reported in the original publication on β2AR crystal structure [45] that there is a noticeable shift in TM1 of β2AR relative to bovine rhodopsin, primarily at the extracellular portion which tilts away from the TM bundle compared to bovine rhodopsin [45,46]. The long N-terminal fragment could not be observed in both crystal structures, but it could cause large flexibility/variation in the assignment of TM1 boundaries, especially in the upper helical region. TM5 has a proline-induced kink at conserved positions along the transmembrane segments, which is believed to be responsible for the structural rearrangements required for the GPCR activation [193,212-215]. The subtle difference in the activation status of the current β2AR structure (2RH1, bound to an inverse agonist carazolol) may lead to the structural diversification at the kink region, in terms of the amplitude of motion and rotation degree. Notably TM3 and TM5 form half of the binding pockets for the co-crystallized carazolol [45] (Figure 3.1b).

Unlike AM models, three CMs deviate from both bovine rhodopsin and β2AR in a similar way (cf. CM1 in Figure 3.1c as the representative case). A large discrepancy can be found at TM1, 4, 6 and 7 for CM1/CM2 and TM1, 3, 5 and 6 for CM3. Here the similar reasoning used in the analysis of AMs can be applied to TM1 because the N-terminal fragment was not considered as part of model building and optimization. Interestingly, the RMSD of CM1's TM4 is as large as 5.00 Å with respect to rhodopsin and 5.21 Å when

aligned against β2AR. It is surprising since TM4 seems the easiest one to model. Among all seven TMs, TM4 is the shortest and the most orthogonal to the plane of the phospholipid membrane. This observation indicates the limitation of computational protocols employed in developing the CM models in their ability to reproduce the conformation of this 'anchor' helix in the TM bundle. For TM5, 6 and 7, one of the common shared features is the proline-induced kink. The comparison between β2AR models and the crystal structure highlights the difficulty associated with the accurate modeling of this unique structural feature of GPCRs.

3.3.2.   Comparison of VS Performance for Theoretical Models and X-ray Structure of β2AR.

Figure 3.2 compares the virtual screening performances on discriminating 58 known β2AR antagonists against decoys by three docking methods. Both the enrichment factor plot (a,c,e) and ROC curves (b,d,f) were included for each method. The yield plot is shown in Figure 3.4 (a,c,e) of the Supporting Information since it essentially delivers the same information as the ROC curves. The detailed statistical parameters characterizing the VS performance, such as the maximum EF (EFmax), ROC AUC and the recently proposed Boltzmann-Enhanced Discrimination of ROC (BEDROC) [216,217], are summarized in Table 3.2. All four holo models as well as the β2AR structure were used in the VS study. The data for only two apo models (AM1 and AM3) are included in the plots because of these models' generally poor performance during virtual screening. In many cases, such as AM1 in Figures 3.2d and 3.2f, the ROC curve is close to the random expectation (the diagonal line). It is understandable because the side chain rotamers of binding pocket residues in the AM models had not been optimized in the way it was done for holo models.

Among all four CMs models, those from the Lybrand group (CM1, CM2) achieved better enrichment than the models from the Goddard group (CM3, CM4). In most cases, the

CM2 model yielded comparable results to the β2AR crystal structure. In the screening by Glide4.01, CM2 model gave higher EF at the very early phase (0% - 2% of ranked database, cf. Figure 3.2a and Table 3.2) After that, it remained up to one unit lower than the crystal structure until converging with the latter at the 22% of ranked database. In terms of ROC AUC, CM2 reached the value of 0.86, close to AUC of 0.88 for the crystal structure. Based on the BEDROC metric, CM2 was similar to 2RH1 when a weak emphasis or weight was put on early enrichment (tuning parameter α=20), while showing better performance when the emphasis increases (tuning parameter α=53.6 or 100).

We note that Glide4.01 gave better VS results in this study compared to AutoDock4.0 and eHiTS6.2. Thus, we placed more emphasis on the screening data/docking poses generated by Glide4.01. Nevertheless, the results obtained with both AutoDock4.0 and eHiTS6.2 highlighted the impressive performance of CM2. As demonstrated by both types of plots in Figures 3.2c-f and most criteria in Table 3.2, the CM2 afforded VS results superior to the crystal structure when using these two docking programs. Therefore, practically speaking, these results suggest that the use of crystal structure is not advantageous in terms of VS performance when the scoring function is not highly accurate. As mentioned above, CM3 and CM4 had poorer performance than CM1 and CM2 but were comparable to AM1 and AM3 models in this case.

The crystal structure of β2AR represents an inactive state of the receptor because it is bound to the inverse agonist, carazolol [45,218]. Thus, it may be considered unfair to compare the crystal structure of β2AR with theoretical models as applied to agonist screening, especially when the models were created to capture known data on agonists. However, for comparison purposes, we did explore the possible utility of 2RH1 for screening for agonists.

The screening results are summarized in Figures 3.5a,b and Table 3.5 of the Supporting Information. As expected, the CM2 model showed the best performance to enrich for thirteen β2AR agonists. With Glide4.01 method, the CM2 model could recover 100% of seed agonists at the 15% of ranked databases and its maximum EF could be as high as 36.09. Thus, it excelled over 2RH1 greatly in terms of these two parameters of VS performance. Taking into account the data for the antagonist virtual screening, we shall conclude that CM2 model demonstrated remarkable performance as a model of choice for virtual screening for both agonists and antagonists.

The possible explanation for the better performance of CM1/CM2 models is that Lybrand et al [178,179] exploited many site-directed mutagenesis data during the model optimization. The important receptor/ligand interactions had been turned into the distance restraints that were applied explicitly to specific atoms of both the receptor and its ligands during molecular dynamics simulations [178]. In comparison, CM3 and CM4 models from Goddard et al [177,209] did not employ such information; their models were built by optimizing the target/ligand interaction using physical force field. Obviously, the differences in the type of data utilized for theoretical model building and optimization can largely affect the accuracy of binding pocket modeling, and consequently, the model performance in virtual screening experiments. In order to evaluate the similarity between binding pockets of individual models, we superimposed Ca atoms of key residues inside the pocket with their counterparts in 2RH1. The binding pocket was defined by residues found within 4 Å of the co-crystallized carazolol. Carazolol was merged into the binding sites of all models as defined by the alignment. As shown in Figure 3.2a-g, the CM2 binding pocket (Figure 3.2b) is most similar to that of 2RH1 with respect to both the ligand pose and the position of

61

residues interacting with the ligand. The RMSD for its Cα atoms was 2.40Å while the one for CM1 was 2.33Å (cf. Table 3.4 of Supporting Information). These two models also reproduced the contacts of carazolol with residues Ser203[5.42] , Asn312[7.39] and Phe193 . For three AMs models, the RMSDs ranged from 3.39Å to 3.71Å. CM3 and CM4 models had the largest deviation (RMSD = 5.64Å), as can also be seen in Figures 3.2c and 3.2d.

Furthermore, the close inspection of the top-ranked docking poses of all seed antagonists showed that the interactions between the antagonists and the binding site of the CM2 model were largely in agreement with the site-directed mutagenesis data. The protonated nitrogen in most β2AR antagonists formed salt bridges with Asp113[3.32] and Asn312[7.39] ; the amide hydroxyl group formed hydrogen bonds with Ser203[5.42] , Ser204[5.43] or Ser207[5.46] . Another important interaction was formed between antagonists and Phe193 of EL2, i.e., the residue that was also found to interact with carazolol within the crystallographic structure of β2AR [45,46,178] . It should be pointed out that CM2/CM1 models include both extracellular and intracellular loops, whereas CM3/CM4 and AM1 models did not incorporate these regions [46,177].

To elucidate the molecular basis for dissimilar virtual screening performance of different CM2 and CM3 models in comparison with 2RH1 we have conducted the cluster analysis of the binding profiles of all 58 antagonists docked to the respective binding sites. Binding profiles reflected the strengths of interaction between antagonists and active site residues. Importantly, there were significant differences in the weak interaction patterns of CM2 (Figure 3.3b) and CM3 (Figure 3.3c) in comparison to 2RH1 (Figure 3.3a). The major clusters formed by CM2 in the region of conserved residues matched well to those found in 2RH1, suggesting a critical role of weak interactions between binding site and antagonists. In

comparison, clusters formed by CM3 were scarce and many key interactions were missing, especially for residues Trp286$^{6.48}$, Phe289$^{6.51}$, Phe290$^{6.52}$ and Asn293$^{6.55}$ of the binding pocket. This analysis indicates that the binding pocket of CM2 was well-organized and similar to that of 2RH1 whereas the CM3 binding pocket was formed by somewhat different residues, with the key residues found in the binding site of the x-ray structure were inaccessible to the bound antagonists. The cluster profile of hydrogen bonding pattern was less informative as some prominent patterns at 2RH1 such as the ones with Asn312$^{7.39}$ were absent at both CM2 and CM3. Notably, Ser204$^{5.43}$ and Ser207$^{5.46}$ of CM2 were found to be hydrogen bonded to ca. 15 antagonists, but the same pattern was not observed with either CM3 or 2RH1.

During the course of our studies, the Kobilka et al. experimentally characterized two new structures of β2AR, one in a nanobody-stabilized active state and another in complex with an irreversible agonist. Compare with the inactive state β2AR structure, the agonist-binding pockets have fairly subtle changes, with the major differences at the hydrogen bonding contacts with Ser203$^{5.42}$ and Ser207$^{5.46}$. To determine whether these minor changes will increase receptor's selectivity toward agonists, we carried out structure-based virtual screening studies using the nanobody-stabilized structure (PDB 3P0G), and compared its performances with the inactive state crystal structure as well as the collected theoretical structural models. As expected, the active state crystal structure showed better performances than the inactive state crystal structure in enriching the 13 agonists, and inferior performances in enriching the 58 antagonists (Figure 3.2 and supplementary Figure 3.5). With Glide4.01 method, the active state crystal structure could recover 100% of seed agonists at the 25% of ranked databases and its maximum EF could be as high as 30.69. On

the other hand, the performance of active state structure model CM2 showed relatively comparable performances to the active state crystal structure in terms of EF and AUC. (Table 3.2, supplementary Table 3.5) As mentioned above, the theoretical model CM2 captured the critical agonist-protein interaction at Ser207$^{5.46}$. This piece of evidence illustrated further that properly optimized theoretical models can provide a sensible picture for the active-state binding pocket, and can be employed for structure-based virtual screening of β2AR agonists.

To summarize our observations, we have established that theoretical models of GPCRs generated with knowledge-based approaches can achieve similar if not better VS performance as structural models based on x-ray crystallographic studies. This somewhat surprising observation is reassuring with respect to using carefully developed theoretical models of protein structures for SBDD.

3.4. Conclusions

In this study we have addressed the long-standing debate about the structural accuracy and applicability of theoretical models vs. x-ray structures of proteins for SBDD. We have carried out a systematic study on a large collection of historical human β2AR theoretical models and evaluated their structural accuracies and screening performances in comparison with two recent crystal structures. We have shown that there exists a discrepancy between global structural accuracies of β2AR theoretical models and their screening performances. In general, β2AR theoretical models differ largely from the crystal structure in terms of TMHs definition and global packing while many can achieve the same performance in virtual screening and as demonstrated elsewhere [187], pose predictions. Our analysis indicates that the binding pockets of models showing the best performance are well-organized and they also align well to active sites in the crystal structures. The key

interactions of residues in the active site with the bound antagonists were found to be preserved in models that were built and refined taking into account the site-directed mutagenesis and other experimental data. Our results emphasize that knowledge-based approaches result in structural models that can achieve the same or even better performance in virtual screening as those built with x-ray crystallographic data. At the same time, we must stress that our studies address very specific and pragmatic question concerning the use of protein models vs. experimental structures for virtual screening. They by no means undermine the critical importance of experimental structures for understanding protein structure-function relationships as well as the role that crystal structures serve as a critical reference for evaluating the accuracy of predicted protein/ligand interactions.

Tables and Figures
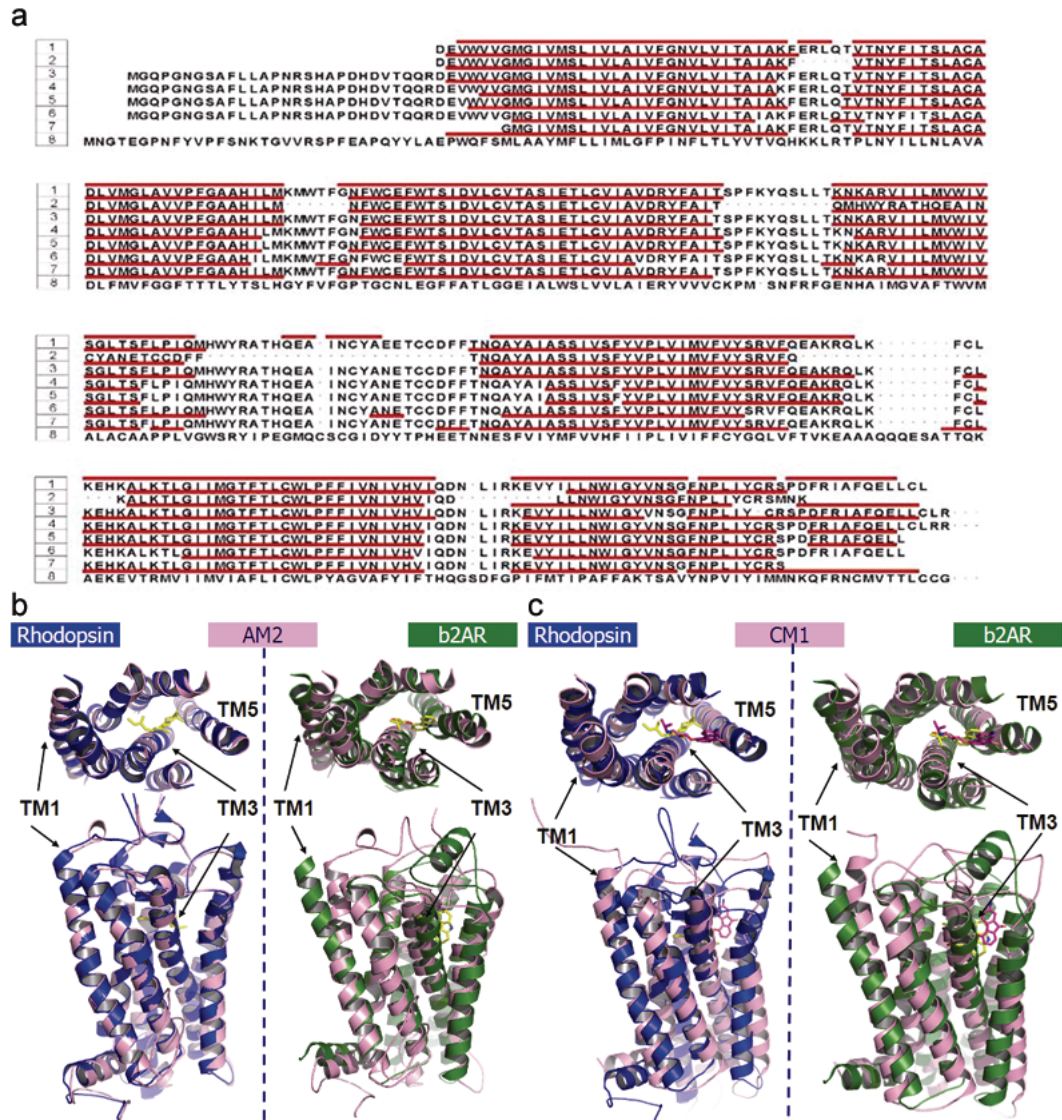
Figure 3.1. The structural similarity analysis of theoretical models in comparison with crystal structures. (**a**) The secondary structure assignment for TM segments of six theoretical models and two crystal structures (2RH1 and 1U19). The numbers and their corresponding structures are coded by 1:2RH1, 2:AM1, 3:AM2, 4:AM3, 5:CM1, 6:CM2, 7:CM3, 8:1U19. The remaining two models, i.e., AM4 and CM4, share similar backbone structures to CM3 with pair wise RMSD of TMs less than 0.4 Å. Therefore only CM3 is included in the sequence alignment plot. The red bars indicate the helical structure elements identified by MOE. (**b**) The structural superposition of the theoretical models AM2 (rendered in pink) to 2RH1 (rendered in dark green) and 1U19 (rendered in blue). Note that the most structurally divergent TM regions are indicated. (**c**) The structural superposition of the theoretical models CM1 (rendered in pink) to 2RH1 (rendered in dark green) and 1U19 (rendered in blue). Note that the most structurally divergent TM regions are indicated.

66
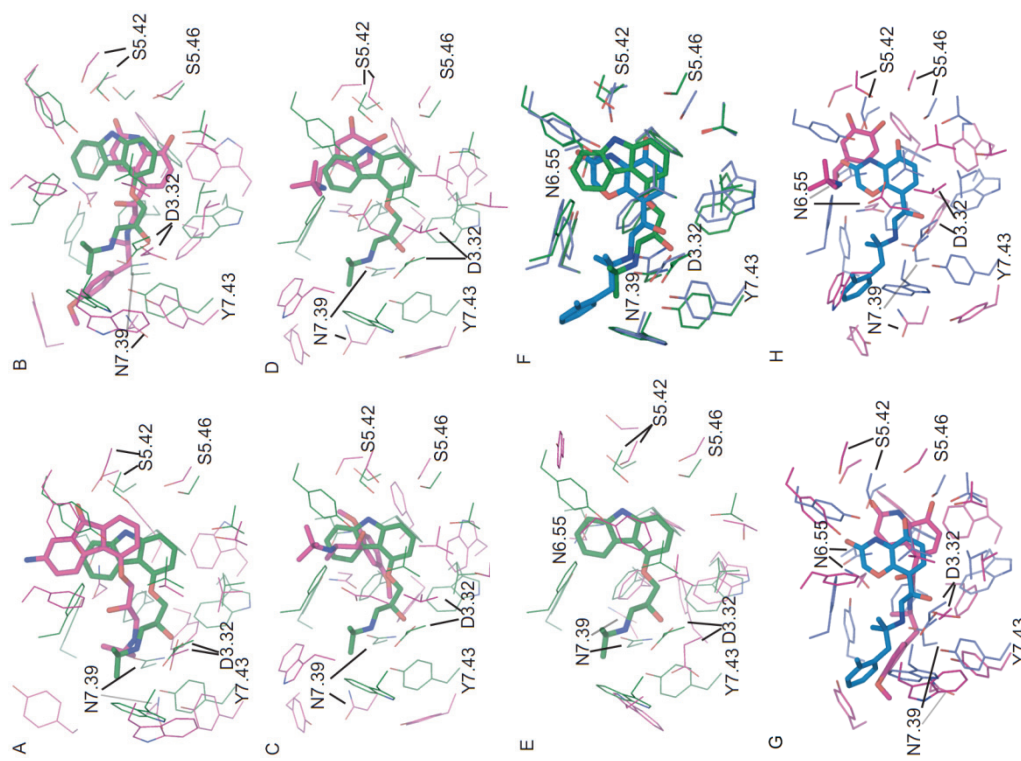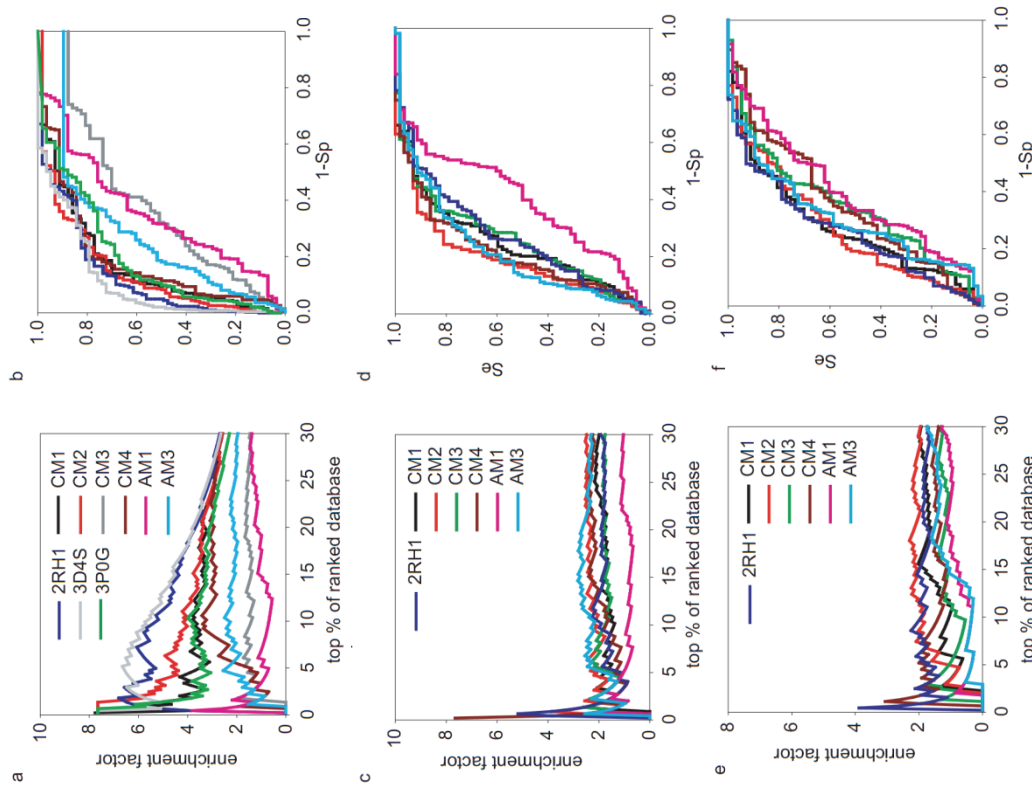
Figure 3.2. The docking performance of theoretical models in comparison with crystal structure 2RH1. (**Left panel, a-f**) The enrichments (**a, c, e**) and ROC curves (**b, d, f**) of fifty-eight known β2AR antagonists for the three docking methods. The annotations are (**a, b**) Glide4.01 (**c, d**) AutoDock4.0 and (**e, f**) eHiTS6.2. (**Right panel, A–G**) The comparison of the binding pockets for theoretical models versus the β2AR crystal structure 2RH1 (inactive state) and 3P0G (active state). (A:CM1, B:CM2, C:CM3, D:CM4, E:AM1) vs. the crystal structure 2RH1. (F:2RH1, G:CM2, H:CM4) vs. the crystal structure 3P0G. The active sites were superimposed by Cα atoms of key binding site residues of 2RH1 (W3.28, D3.32, V3.33, V3.36, T3.38, F5.32, Y5.38, S5.42, S5.43, S5.46, W6.48,

F6.51, F6.52, N6.55, Y7.35, N7.39, Y7.43). The crystal structures 2RH1, and 3P0G are colored in green and blue respectively, while the model structures are colored in pink.
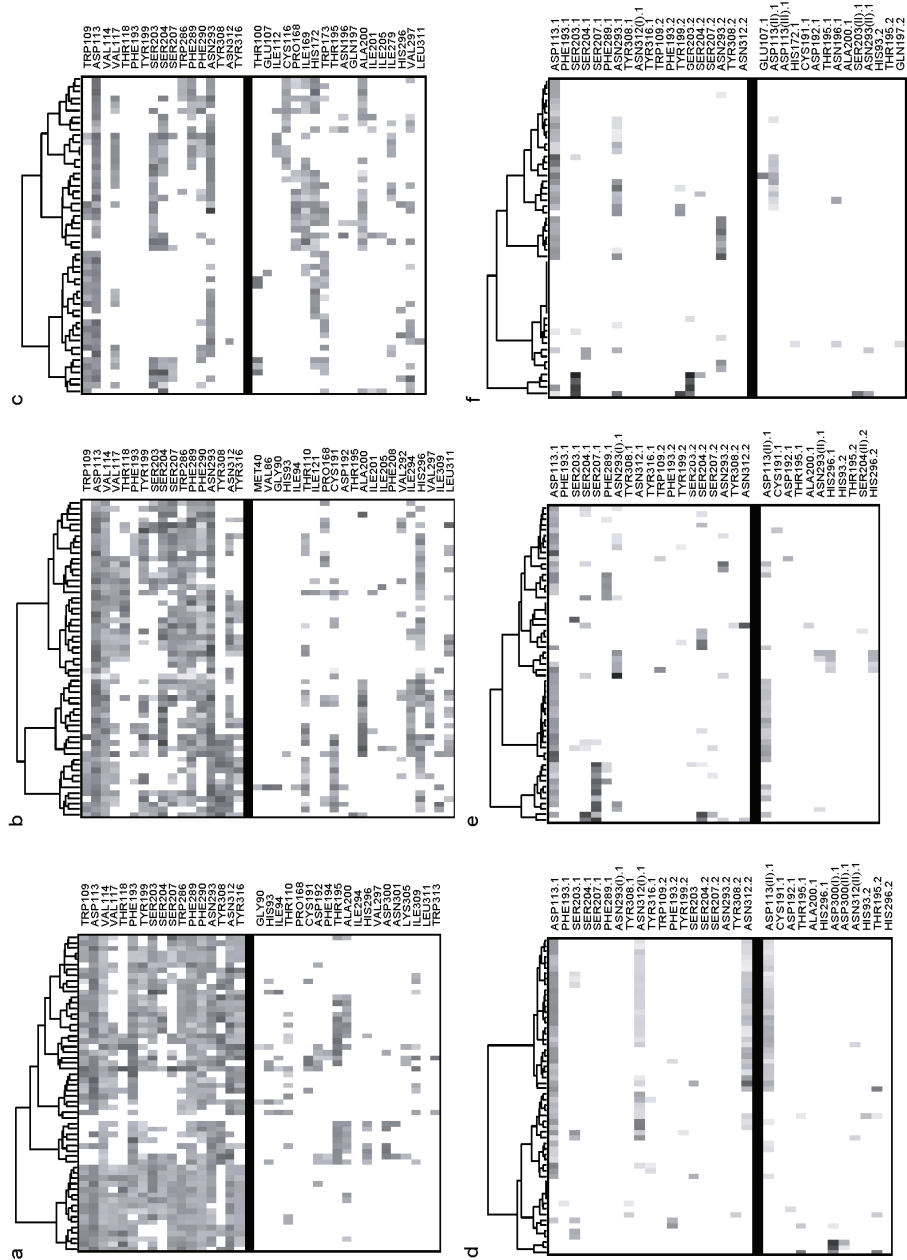
Figure 3.3. The cluster analysis of the antagonists binding profile. (**Upper panel, a–c**) The comparison of weak interaction profiles of fifty-eight antagonists of human β2AR with 2RH1 (**a**), CM2 (**b**) and CM3 (**c**). (**Lower panel, d–f**) The comparison of hydrogen bonding (HBond) profiles of fifty-eight antagonists of human β2AR with 2RH1 (**d**), CM2 (**e**) and CM3 (**f**). The weak and HBond interactions were identified/scored by LigX module in MOE2007.09 and marked as X.1 for HBond donors and X.2 for HBond acceptors. In the case that one residue forms two HBonds to the ligand, the interactions were labeled as X(I).X or

68

X(II).X, in which the X(I).X had the better score. The upper block in each map contains the binding pocket residues of 2RH1 while the lower block contains other interacting residues. Each point in the maps represents the identified interactions and was shaded by their respective scores, wherever darker indicates higher score and thus greater interaction strength. All the points in the map had been reorganized using hierarchical clustering by interaction scores.

Table 3.1. The synonym of eight human β2AR theoretical models employed in this study.

Apo model	Source	Holo model	Source
AM1	By **G. Vriend**, a homology model[219] based on the crystal structure of bovine rhodopsin	CM1	By **T. Lybrand**, a *de novo* model[178,179] bound by aminoflisopolol (β2AR antagonist)
AM2	By **A. Sali**, a homology model[220] based on the crystal structure of bovine rhodopsin	CM2	By **T. Lybrand**, a *de novo* model[178,179] bound by TA2005 (β2AR agonist)
AM3	By **J. Skolnick**, a hybrid model[180] combined threading and *ab initio* methods	CM3	By **W. Goddard**, a *de novo* model[177,209] bound by butoxamine (β2AR antagonist)
AM4	By **W. Goddard**, a *de novo* model[177] based on first principles methods	CM4	By **W. Goddard**, a *de novo* model[177,209] bound by salbutamol (β2AR agonist)

Table 3.2. The enrichment factors (EF) and yields of fifty-eight known β2AR antagonists by theoretical models versus two crystal structures.

Structure /Model	Docking method	EF at 1% db	EF at 5% db	EF at 25% db	% db to find 25% seeds	Yield at 25% db	EF$_{max}$	% db EF$_{max}$ occurred	ROC AUC	AUC at 25% db	BEDROC α=20	BEDROC α=53.6	BEDROC α=100
2RH1	Glide4.01	5.83	5.30	3.08	4.74	77.19	6.80	1.81	0.88	0.11	0.72	0.75	0.74
	AutoDock4.0	3.89	1.06	1.75	11.74	43.86	0.68	5.18	0.73	0.05			
	eHiTS6.2	1.94	1.77	1.82	13.54	45.61	3.89	0.45	0.74	0.05			
CM1	Glide4.01	5.83	3.53	2.87	6.77	71.93	7.77	1.35	0.83	0.08	0.50	0.52	0.58
	AutoDock4.0	0.00	1.77	2.17	13.77	54.39	2.59	1.35	0.76	0.04			
	eHiTS6.2	0.00	0.71	1.75	14.67	43.86	2.03	15.58	0.72	0.04			
CM2	Glide4.01	7.66	4.52	2.90	5.63	72.41	7.66	1.35	0.86	0.09	0.61	0.72	0.83
	AutoDock4.0	3.83	2.09	2.48	11.49	62.07	3.83	0.90	0.81	0.06			
	eHiTS6.2	0.00	1.04	2.14	12.16	53.45	2.30	18.02	0.73	0.05			
CM3	Glide4.01	0.00	1.41	1.54	17.61	38.60	1.94	0.90	0.61	0.04	0.20	0.15	0.11
	AutoDock4.0	1.94	1.41	1.82	14.90	45.61	2.59	1.35	0.75	0.05			
	eHiTS6.2	0.00	1.06	1.26	23.25	31.58	1.79	2.93	0.65	0.03			
CM4	Glide4.01	1.91	1.39	2.90	2.70	72.41	3.42	10.59	0.82	0.07	0.34	0.19	0.13
	AutoDock4.0	1.91	1.39	2.34	12.61	58.62	7.66	0.23	0.78	0.05			
	eHiTS6.2	1.91	1.74	1.52	16.89	37.93	3.06	1.13	0.65	0.04			
AM1	Glide4.01	1.91	0.70	1.31	20.27	32.76	3.83	0.45	0.65	0.03	0.14	0.17	0.22
	AutoDock4.0	1.91	1.04	1.03	22.75	25.86	1.91	0.90	0.61	0.03			
	eHiTS6.2	0.00	0.35	0.97	25.90	24.14	1.42	37.61	0.60	0.03			
AM3	Glide4.01	0.00	2.44	2.07	10.36	51.72	2.55	4.73	0.72	0.06	0.28	0.20	0.14
	AutoDock4.0	1.91	1.74	2.34	10.81	58.62	2.83	14.64	0.78	0.07			
	eHiTS6.2	0.00	0.35	1.45	17.57	36.21	1.81	32.43	0.69	0.03			
3D4S	Glide4.01	6.12	6.66	3.17	4.05	79.31	7.65	0.22	0.89	0.11	0.77	0.80	0.79
3P0G	Glide4.01	5.74	3.33	2.69	7.21	67.24	7.66	0.68	0.81	0.08	0.50	0.52	0.62

Supplementary Material

a

b

c
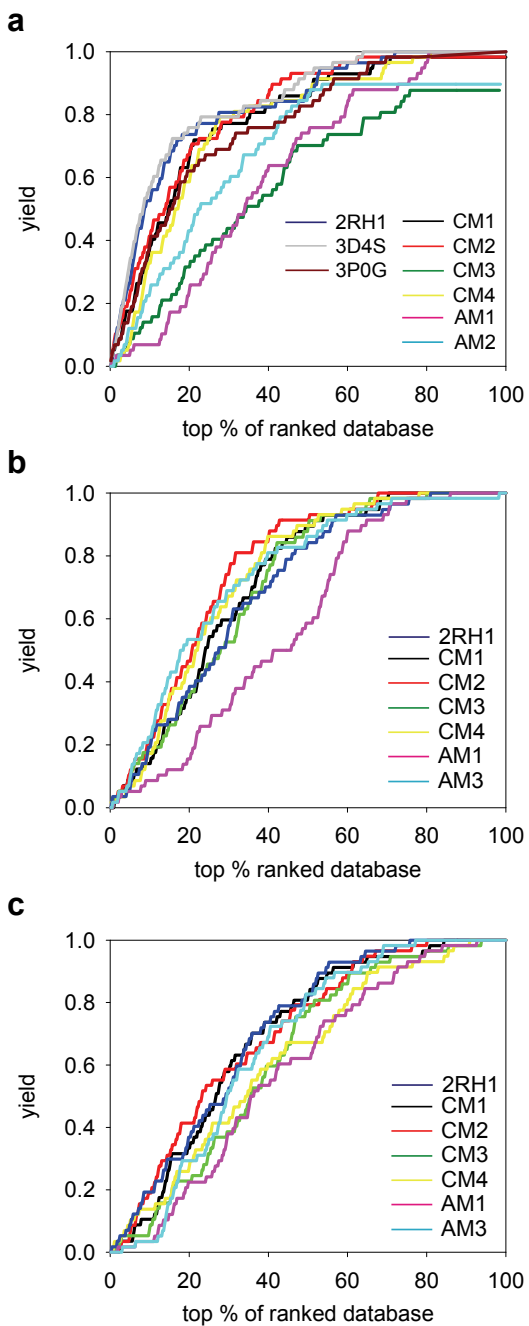
Figure 3.4. The yields of 58 known β2AR antagonists during the screening by three docking methods versus the crystal structure and six theoretical models. The annotations are (a) Glide4.01 (b) AutoDock4.0 and (c) eHiTS6.2.
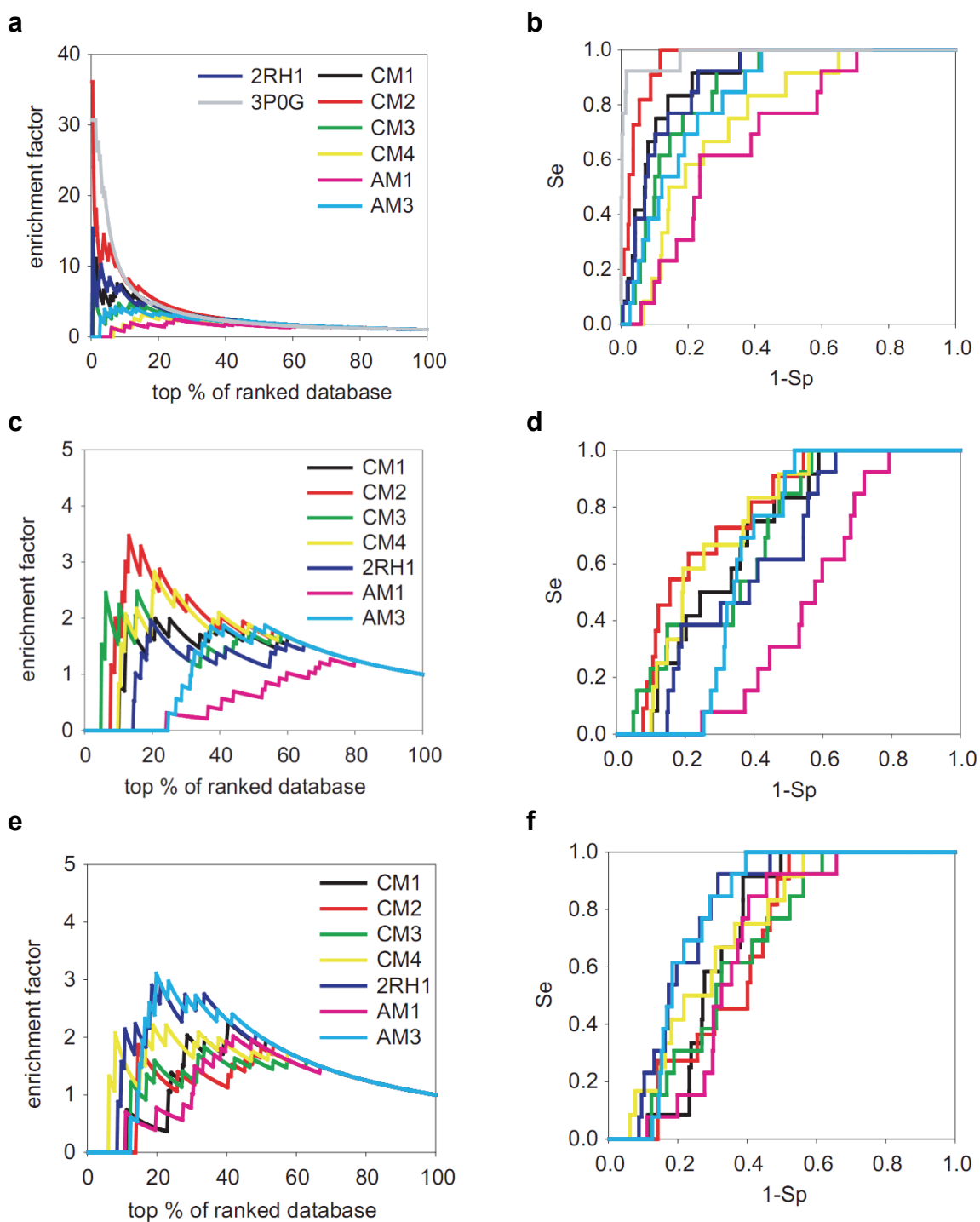
Figure 3.5.. The enrichments and ROC curves of 13 known β2AR agonists during the screening by three docking methods versus the crystal structure and six theoretical models. The annotations are (a) Glide4.01 (b) AutoDock4.0 and (c) eHiTS6.2.
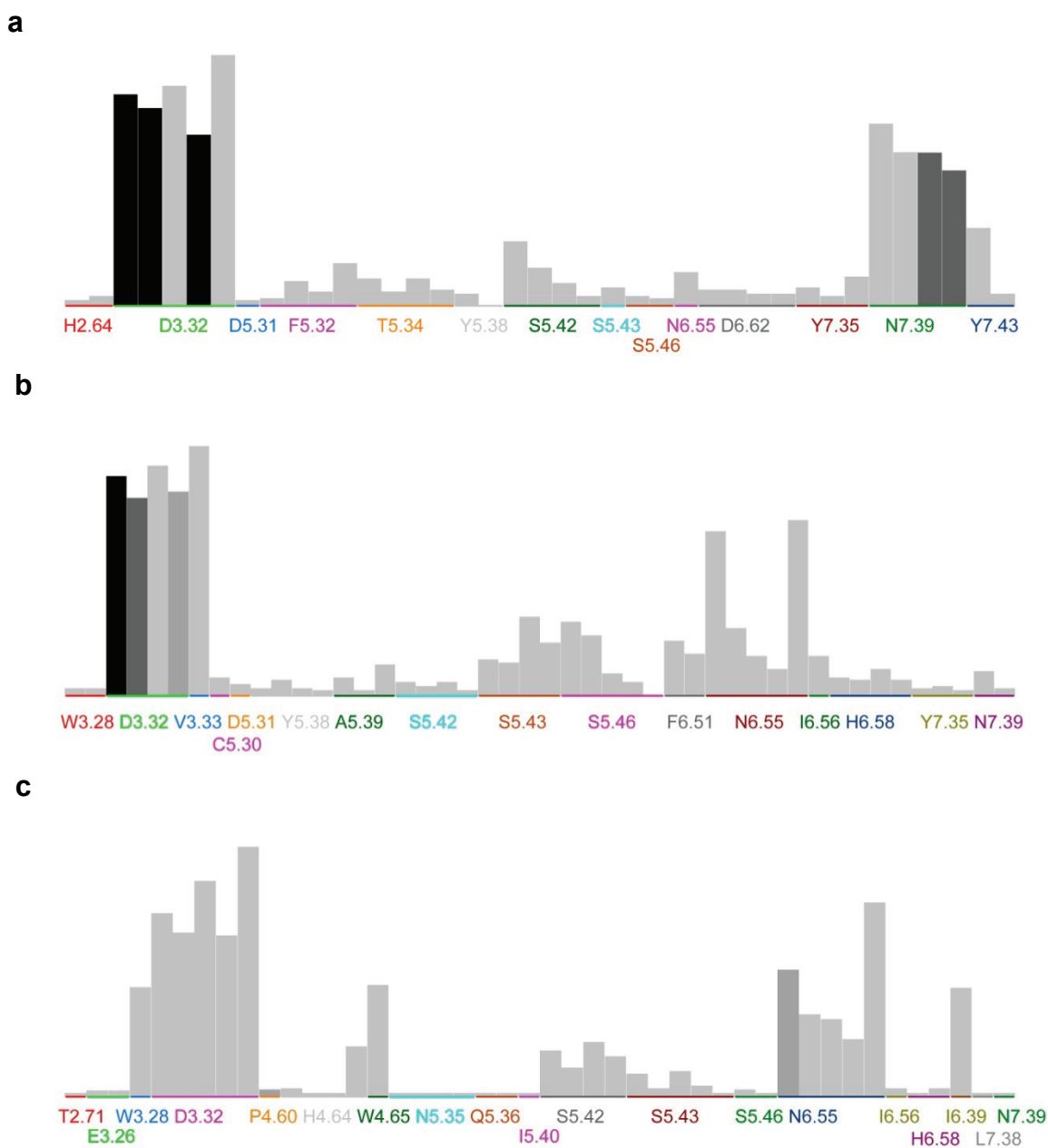
Figure 3.6. The significance chart for PLIF fingerprints generated from docking poses against (a) β2AR crystal structure, (b) CM2 model and (c) CM3 model. The higher the bar (individual fingerprint bit) is, more frequently this type of interaction occurs in the dataset and of higher probability it contributes to the activities. The shade of the color indicates the significance of the particular bit to the actives, which is based on the hypothesis that 'if the bit is set, then the compound is active'. The residues are randomly colored and several bars of the same residue indicate that they have different types of contacts.

Table 3.3. The boundary definition of seven TM helices in 1U19 and 2RH1 by PDB header and MOE annotation. The exact matches are marked in bold while the mismatches by two residues are shown in italic.

Crystal str.	Annotation	TMI		TMII		TMIII		TMIV		TMV		TMVI		TMVII	
		start	end	start	end	start	end	start	end	start	end	start	end	start	end
1U19	PDB header	1.28	1.60	**2.38**	**2.68**	3.20	3.55	4.38	4.62	5.30	5.61	6.24	6.60	7.31	7.57
	MOE annot.	1.29	1.59	**2.38**	**2.67**	3.21	3.54	4.39	4.62	5.31	5.62	6.25	6.59	7.32	7.55
2RH1	PDB header	*1.28*	**1.59**	**2.38**	**2.67**	3.22	**3.55**	**4.39**	4.63	**5.36**	**5.68**	**6.29**	**6.60**	**7.32**	**7.55**
	MOE annot.	*1.30*	**1.59**	**2.38**	**2.67**	3.21	**3.55**	**4.39**	4.62	**5.36**	**5.68**	**6.29**	**6.60**	**7.32**	**7.55**

Table 3.4. Comparisons for six β2AR theoretical models versus two crystal structures in terms of TM helical boundaries and RMSD. The bordering residues for each TM region are identified by MOE and annotated using a specialized numbering rule. The RMSDs for each model's TM were calculated using the true length of the corresponding segments of 2RH1 and 1U19 (in italic).

Segment	2RH1 / *1U19*	AM1 start/end	AM1 RMSD	AM2 start/end	AM2 RMSD	AM3 start/end	AM3 RMSD	CM1 start/end	CM1 RMSD	CM2 start/end	CM2 RMSD	CM3 start/end	CM3 RMSD
TMI	1.30/1.59 *1.29/1.59*	1.29/1.60	2.77Å *0.30Å*	1.29/1.59	3.02Å *0.30Å*	1.28/1.59	3.01Å *0.83Å*	1.31/1.59	3.18Å *2.93Å*	1.32/1.58	3.19Å *3.03Å*	1.35/1.56	5.12Å *6.90Å*
TMII	2.38/2.67 *2.38/2.67*	2.38/2.67	1.76Å *0.32Å*	2.38/2.67	1.59Å *0.37Å*	2.38/2.67	1.90Å *0.68Å*	2.37/2.65	3.19Å *2.40Å*	2.37/2.65	3.09Å *2.36Å*	2.45/2.64	4.38Å *4.31Å*
TMIII	3.21/3.55 *3.21/3.54*	3.22/3.55	4.32Å *0.34Å*	3.22/3.54	2.09Å *0.37Å*	3.23/3.54	1.67Å *1.08Å*	3.22/3.55	2.57Å *2.21Å*	3.23/3.55	2.56Å *2.25Å*	3.27/3.47	6.99Å *5.00Å*
TMIV	4.39/4.62 *4.39/4.62*	N/A	1.95Å *0.32Å*	4.39/4.61	1.94Å *0.49Å*	4.39/4.62	1.67Å *1.06Å*	4.40/4.57	5.21Å *5.00Å*	4.41/4.57	5.43Å *5.02Å*	4.38/4.68	3.47Å *4.60Å*
TMV	5.36/5.68 *5.31/5.62*	5.34/5.62	2.37Å *0.43Å*	5.35/5.62	3.48Å *0.38Å*	5.35/5.70	2.76Å *1.00Å*	5.41/5.67	2.17Å *2.04Å*	5.41/5.67	2.13Å *2.06Å*	5.37/5.58	6.85Å *3.59Å*
TMVI	6.29/6.60 *6.25/6.59*	6.33/6.60	1.67Å *0.32Å*	6.33/6.59	3.80Å *9.11Å*	6.27/6.60	1.96Å *0.80Å*	6.28/6.59	4.24Å *3.63Å*	6.28/6.59	4.21Å *3.69Å*	6.38/6.59	8.12Å *12.54Å*
TMVII	7.32/7.55 *7.32/7.55*	7.38/7.56	3.82Å *0.38Å*	7.32/7.51	1.65Å *1.01Å*	7.33/7.55	2.35Å *1.37Å*	7.32/7.55	4.08Å *3.34Å*	7.32/7.55	4.20Å *3.42Å*	7.34/7.55	3.93Å *2.64Å*
TMI-VII			2.47Å *1.15Å*		3.19Å *1.88Å*		2.25Å *1.38Å*		3.59Å *3.20Å*		N/A *N/A*		4.41Å *3.83Å*
ECLII		N/A	N/A *N/A*		13.43Å *1.32Å*		13.19Å *1.81Å*		7.00Å *13.72Å*		7.18Å *13.96Å*		19.00Å *19.70Å*
Binding pocket			3.71Å		3.39Å		3.58Å		2.33Å		2.40Å		5.64Å

Table 3.5. The enrichment factors (EF) and yields of 13 known β2AR agonists by six theoretical models versus the crystal structure 2RH1.

Structure /Model	Docking method	% db to find 25% agonists	EF at 25% db	Yield at 25% db	Max. EF	% db EF_{max} occurred	AUC	AUC at 25% db
2RH1	Glide4.01	3.01	3.38	84.62	15.35	0.50	0.90	0.15
	AutoDock4.0	18.55	1.53	38.46	1.97	19.55	0.63	0.03
	eHiTS6.2	13.78	2.46	61.54	2.92	21.05	0.79	0.06
CM1	Glide4.01	3.77	3.65	91.67	11.06	0.75	0.91	0.15
	AutoDock4.0	12.31	1.99	50.00	2.03	12.31	0.70	0.04
	eHiTS6.2	23.37	1.33	33.33	2.27	40.45	0.70	0.02
CM2	Glide4.01	1.51	3.97	100.00	36.09	0.25	0.96	0.20
	AutoDock4.0	11.08	2.53	63.64	3.47	13.10	0.77	0.08
	eHiTS6.2	14.61	1.08	27.27	1.87	53.40	0.66	0.03
CM3	Glide4.01	7.77	3.07	76.92	15.35	0.50	0.86	0.11
	AutoDock4.0	15.04	1.53	38.46	2.48	15.54	0.69	0.06
	eHiTS6.2	19.30	1.23	30.77	1.83	33.58	0.66	0.03
CM4	Glide4.01	12.31	2.32	58.33	3.26	15.33	0.75	0.07
	AutoDock4.0	26.13	2.32	58.33	2.83	20.60	0.74	0.05
	eHiTS6.2	15.33	1.99	50.00	2.21	18.84	0.72	0.05
AM1	Glide4.01	17.04	2.46	61.54	2.48	24.81	0.69	0.05
	AutoDock4.0	44.11	0.31	7.69	1.27	72.68	0.44	0.00
	eHiTS6.2	30.08	0.61	15.38	2.02	41.85	0.66	0.02
AM3	Glide4.01	7.27	3.07	76.92	4.23	7.27	0.83	0.10
	AutoDock4.0	31.33	0.31	7.69	1.87	41.10	0.64	0.00
	eHiTS6.2	15.79	2.76	69.23	3.11	19.80	0.78	0.05

Table 3.6. The rescoring results using Sybyl8.0 CScore module on the poses generated by Glide4.01. The best metrics under each scoring function is highlighted by bold/italic. *GS: Glide score; CS: Consensus score of the four scoring functions through rank by rank strategy; CH: Chemscore; D: D_score; G: G_score; PS: PMF_score.

	EF at 5% db						EF at 25% db						AUC					
	*GS	CS	CH	D	G	PS	GS	CS	CH	D	G	PS	GS	CS	CH	D	G	PS
RH	6.08	0.71	*1.41*	2.83	0.71	0.71	3.08	1.13	1.70	3.11	2.12	1.98	0.88	0.66	0.77	0.88	0.78	0.79
C1	3.72	1.06	1.06	1.41	*1.77*	0.71	2.80	1.84	*2.68*	2.47	1.48	2.26	0.83	0.71	*0.82*	0.82	0.72	*0.80*
C2	4.52	1.39	1.39	2.09	1.39	0.70	2.90	2.14	1.86	3.03	2.28	2.28	0.86	0.76	0.76	0.84	0.78	0.78
C3	1.41	1.06	0.00	0.35	0.35	*1.06*	1.54	0.92	0.28	1.20	0.64	1.27	0.61	0.54	0.50	0.57	0.51	0.63
C4	1.39	2.09	1.39	1.39	1.04	1.04	2.90	1.93	1.66	2.21	2.55	1.31	0.82	0.71	0.74	0.78	*0.82*	0.69
A1	0.70	1.39	1.39	0.70	0.70	0.35	1.31	1.38	1.38	1.72	1.38	1.45	0.65	0.66	0.72	0.74	0.70	0.70
A3	2.44	1.74	1.39	*3.13*	1.74	0.35	2.07	1.86	2.28	3.03	*2.76*	*3.03*	0.72	0.71	0.74	0.80	0.70	0.77
DS	*6.66*	*2.44*	1.39	*3.13*	1.04	1.04	*3.17*	*2.28*	1.86	*3.17*	2.21	2.07	*0.89*	*0.77*	0.77	*0.89*	0.78	0.79

Table 3.7. The rescoring results using OpenEye FRED2.2.4 scoring functions on the poses generated by Glide4.01. The best metrics under each scoring function is highlighted by bold/italic. *CS: Consensus score of five scoring functions through rank by rank strategy; CG: Chemgauss3; CH: Chemscore; SS: Screenscore; SG: Shapegauss.

	EF at 5% db						EF at 25% db						AUC					
	*CS	CG	CH	PLP	SS	SG	CS	CG	CH	PLP	SS	SG	CS	CG	CH	PLP	SS	SG
RH	3.18	*3.89*	*4.95*	2.47	*3.18*	*2.47*	2.90	3.11	3.25	*2.83*	2.47	1.41	0.85	0.89	0.91	*0.85*	0.81	0.35
C1	3.18	3.18	2.83	2.83	*3.18*	*2.47*	2.47	1.98	2.90	2.05	2.26	1.41	0.82	0.79	0.86	0.77	0.79	0.70
C2	3.48	2.78	3.13	2.78	3.13	1.74	*2.97*	*3.24*	2.90	2.69	*2.83*	1.52	0.84	0.85	0.87	0.79	0.80	0.73
C3	0.35	0.00	0.35	0.00	0.00	0.00	0.57	0.57	1.20	0.78	0.64	0.21	0.56	0.57	0.58	0.54	0.55	0.59
C4	1.74	2.44	1.74	1.39	2.09	1.74	2.34	2.90	2.55	2.07	1.86	1.24	0.82	0.85	0.82	0.75	0.75	0.69
A1	1.39	1.39	1.04	1.39	0.70	0.35	1.86	1.93	1.86	1.38	1.38	1.17	0.72	0.75	0.74	0.64	0.65	0.66
A3	*4.52*	1.74	3.13	*3.48*	3.13	1.39	2.76	2.69	3.17	1.59	1.72	*2.62*	0.78	0.75	0.81	0.71	0.71	*0.75*
DS	2.44	3.83	*5.57*	3.13	2.44	2.09	2.83	*3.24*	*3.45*	2.76	2.55	1.38	*0.86*	*0.89*	*0.92*	0.85	*0.83*	0.72

Table 3.8. The actual number of compounds used for calculating EF and AUC metrics.

Structure/Model	[a]Antagonists Num.	[b]Nonbinders Num.
2RH1	29	272
CM1	29	272
CM2	30	245
CM3	28	307
CM4	30	307
AM1	30	303
AM3	30	205
D-SS	30	374
F-SS	30	374

[a]The bound ligand for 2RH1 and CM1~CM4 were excluded from the calculation.
[b]A number of decoys were filtered by Glide before the refinement stage. They thus did not have a Glide score and were treated to be at the end of the ranked database.

Chapter 4. Development, Validation and Application of Target-specific Pose Filters in Structure-Based Virtual Screening of Subtype Selective Ligands

4.1. Introduction

Traditional drug discovery and development approaches were generally applied for searching novel active compounds and optimizing compounds' potency with a specific drug target in mind. Recent fast developments in proteomics and chemogenomics studies, however, have triggered a paradigm shift from this focus toward a more comprehensive analysis of the entire ligand space and target space and to explore all prospective drug-target interactions [3,8,221,222]. The role of small molecules are no longer limited to a inhibitor or active compound that can be promoted to drug candidates, but more as a probe to understand biological functions through perturbing cellular circuits and pathways. Despite the increasing complexity and diversity of research evolved from this new field, the major questions could be generally categorized into two themes, for an interested organic compound with certain observed phenotype, how to identify the responsible biological target (target fishing); and what is the selectivity profiles of a known organic compound across all relevant targets (ligand profiling)[4,5,33].

The success of computational approaches, including structure-based drug design and Quantitative Structure-Activity Relationship (QSAR) analysis, has been well recognized in the field of rational drug design and discovery[223]. However, the rapid growing research in the area of chemogenomics and chemical biology propose new requests for the development of computational methods. There are already a number of reviews discussing possible

computational strategies that can be applied in this field [8,33,62]. A main theme emerged from this area relates to prediction of small molecule's selectivity profile at different levels, for instances, within a target family, or subfamily. Several previous studies have explored the effectiveness of using ligand-based approaches such as two-dimensional fingerprints and Quantitative Structure-Selectivity Relationship (QSSR) models in "selectivity searching". It was shown that carefully designed fingerprints and QSSR models are capable of enriching subtype-selective ligands in small to medium size data sets [136,224,225]. Several other groups aimed at structure-based approaches by incorporating information from predicted protein-ligand contacts. Singh and colleagues [55,56] defined a series of protein-ligand interaction properties to generated Structural interaction fingerprint (SIFt) at protein-ligand's binding site. Virtual screening studies using SIFt revealed superior performance as compared with traditional scoring functions. Graaf and Rognan et al. showed that a topological scoring function based on their ligand protein interaction fingerprint (IFPs) was able to selectively identify agonists or antagonists of the beta2 adrenergic receptor [57-59]. In addition, there is an increasing interest in the computational drug discovery community to build target-specific scoring functions, which have a straightforward application in exploring ligand's binding profiles [60,61,226,227].

Our lab has developed a novel type of four-body descriptor to effectively represent the protein-ligand interface: the PL/MCT-tess (Protein-Ligand atoms' pair wise Maximal Charge Transfer potential based on Delaunay Tessellation) descriptor. Application of this descriptor to our routine QSAR modeling workflow resulted in predictive models for protein ligand binding affinity[61]. A recent study by Hsieh et al. in our group suggested that QSAR models developed based on multiple docking poses from a single cognate ligand can

effectively distinguish native-like docking poses from decoy poses, thus dramatically increased the virtual screening performance when combined with the traditional force-field based scoring functions[60,61]. Herein, we continue to use the PL/MCT-tess descriptors as an effective representation of the protein-ligand interface, and explored other pose-filters that explicitly account for small molecules' subtype binding selectivity. In order to do this, a reference PL/MCT-tess profile (normally from the X-ray protein-ligand complex) was initially used to partition a training set of known actives' docking poses into reference-similar versus dissimilar class. Then a binary classifier was developed that can distinguish reference-similar poses from all available poses of known nonselective/inactive compounds. By changing the reference ligand type (non-selective or subtype selective ligand) and the calibration pool included in target-specific pose filter (treat inactive compounds only as decoys or include also non-selective compounds as decoys), we hope to direct the proposed pose-filters to selectively enriching subtype-biased ligands.

As proof of concept, we first assessed the capabilities of the proposed pose filters to identify known active compounds using 13 subsets from the Directory of Useful Decoys (DUD). Compared with the traditional scoring function, namely Chemgauss3, from Fred, we found that our approach showed better performance for 10 out of 13 data sets, and similar performance for the remaining 3 receptor targets. The best data set provides an increase in the enrichment from 18- to 57-fold over random at a false positive rate of 1%. We further challenged our approaches on a data set of 189 compounds with known binding data of the beta adrenergic receptors (βAR) GPCR subfamily. The βARs belong to the extensively explored Class A Rhodopsin like GPCRs, and a plethora of experimental data is available for model evaluation and validation. The three most explored subtypes in the βAR family are

designated β1AR, β2AR, and β3AR. They are widely distributed and play critical biological functions on several tissues. The β1AR subtype is expressed predominantly in mammalian heart, while the β2AR is the major subtype in most vascular and bronchial smooth muscle cells, the β3AR is mostly located in adipose tissues [228,229]. The βAR mediated signaling pathways play a critical role in regulating cardiac function, and βAR mediated relaxation of the smooth muscle also has essential therapeutic values in treating asthma [228-231]. Identification of βAR ligands with the desired selectivity and activity patterns may lead to better drug candidates as the β1-stimmulating-effects of several β2-agonists determines their cardiac side effects in treating asthma. Similarly, the β2-activity of β1-antagonists may exert harmful effects when used in the treatment of heart diseases. With recent publication of the human β2AR and Turkey β1AR crystal structures [232-234], applying structure-based approach to explore ligand's subtype selectivity has attracted further attention for the βAR sub-family. On the other hand, the sequence similarity among these three subtypes is quite high, especially near the ligand binding regions, making them challenging targets to elucidate ligand binding specificities. Herein, we reported the performance of a novel subtype biased scoring protocol in searching for selective ligands. Furthermore, by analyzing the derived computational tools for highly weighted PL/MCT-tess descriptors, we also proposed prospective protein-ligand contacts that may contribute to the ligands' subtype specificity within the βAR family.

4.2. Pilot Study

4.2.1. Data Sets Selection (the DUD Data Set)

The structures were directly downloaded from the Directory of Useful Decoys (DUD) data sets: http://dud.docking.org [235]. The DUD has been recently compiled by the Schoichet

group, specifically for benchmarking the performance of docking methods. The data sets contain a diverse set of 40 interesting protein targets, as well as a list of known ligands and decoys for each target. On average, 36 decoys were chosen for each active compound structure from the free ZINC database of commercially-available compounds. To minimize the physical bias in benchmarking docking methods, the decoys were chosen specifically to match the physical properties but topologically dissimilar to the known active structures. A number of criteria were explicitly considered such as molecular weight, cLogP, number of hydrogen bonding partners, and number of rotatable bonds. Recently, Good et al. carried out additional refinements of the DUD actives by applying a lead-likeness filter (AloP <4.5, MW <450) and clustering analysis. To be comparable with other virtual screening methods, we have also used the 13 data sets that contain at least 15 clusters. (see Table 4.1 for details) In total, there are six members of the kinase protein family (cdk, egfr, p38, pdgfrb, src, and vegfr2), two members of metalloenzymes (ace, pde5), one serine protease (fxa), and several other enzymes. To employ the proposed virtual screening methods, the original DUD protein-ligand crystal structures were used as the reference during pose filter training. For vegfr2 and pdgfrb where no co-crystallized ligand is provided, the top-scored pose from docking a known ligand to the apo protein structure was used. The data set was collected and cleaned by previous lab member, Dr. Jui-hua Hsieh.

4.2.2.   Methods

Docking Methods for Pose Generation

Each protein target were prepared using the Molprobity server to add hydrogen atoms and assign partial charge, as well as correcting potential structure problems including missing atoms, potentially transposed heavy atoms in asparagine, glutamine, and histidine side chains.

The co-crystallized water molecules were removed in order to reduce bias during pose generation. On the other hand, the cofactors or metal atoms were preserved in cases where they are known to be critical for protein function or involved in ligand binding. The DUD actives and decoys were used directly from download, but ionized and converted to three-dimensional conformers with Omega (version 2.2.1)[236] using the default parameters. The Fred docking software (version 2.2.5)[236] from OpenEye Scientific was then employed to generate docking poses for each conformer from the derived library. We employed the default settings for Fred docking except for changing to a larger number of preserved poses. For each molecule, the top 30 poses selected by the Fred's default scoring function, Chemgauss3, were kept for post-docking analysis. For the 6 kinase protein targets, additional restraints were applied to preserve the canonical protein-ligand interactions at the hinge region [237]. The final data of the docking poses and calculated Chemgauss3 scores were provided by previous lab member, Dr. Jui-hua, Hsieh.

Novel Descriptors of the Protein-Ligand Interface Based on Conceptual DFT

Our group recently developed a set of simple yet effective geometrical descriptors, the ENTess descriptors, to describe the physicalchemical properties at the protein-ligand interface. It utilizes the Pauling electronegativity (EN) to annotate atom types and uses Delaunay Tesselation (Tess) to characterize the geometrical property at the interface. Briefly, for each protein-ligand complex, Delaunay Tessellation was first used to partition the space into an ensemble of tetrahedrons. Only interfacial tetrahedrons that are formed by both protein and ligand atoms were kept to define the protein ligand interface. Furthermore, a distance cutoff of 8A (distance for physically meaningful interactions) was used to exclude Delaunay quadruplets with long edges between atom vertices. Theoretically, a total of 554

types of Delaunay quadruplets were defined by their unique four-atom composition, including the atom types and the origins (receptor atom or ligand atom). The descriptor value was calculated by summing up the electronegativity values of the composing vertices/atoms. In our previous studies, we have successfully built quantitative structure-binding affinity relationship models for a data set of 264 x-ray crystallized protein-ligand complexes using the ENTess descriptors and a standard QSAR modeling workflow [238].

More recently, our group improved the original ENTess descriptors by assigning a more physic-chemically rigorous property to be the descriptor values. Instead of using atom's electronegativity values, pair wise atomic potentials for the protein-ligand complexes (PL) based on maximal charge transfer (MCT) were used to derive the novel PL/MCT-tess descriptors. The MCT is based upon the conceptual DFT [239] to determine the maximal electron flow between the donor and acceptor atoms. Assuming that the total energy of the system is perturbed by the charge transfer up to the second order: The MCT is calculated as follows,

$$\Delta E = \mu \Delta N + 1/2 \eta \Delta N^2 \qquad (2)$$

where ΔE and ΔN are energy change and charge transfer, respectively. When the total energy is minimized with respect to the charge transfer, i.e. $\frac{d\Delta E}{d\Delta N} = 0$, we have

$$\Delta N_{max} = -\mu/\eta \equiv MCT \qquad (3)$$

here $\mu = (\partial E/\partial N)_v$, $\eta = (\partial^2 E/\partial^2 N)_v$, are the chemical potential (negative of electronegativity) and the chemical hardness, respectively. v represents the external potential formed by the framework of atomic nuclei.

The values for a specific tetrahedron type m can be calculated from the following equation:

$$PL/MCT - tess_m = \sum_{k=1}^{n} (\sum_{p=1}^{3} \sum_{l=1}^{3} MCT_p \times MCT_l / d_{pl})_k \quad (4)$$

where n is the number of occurrences of this tetrahedron type in a given protein-ligand complex. For a specific tetrahedron k, p iterates through all vertices belong to the protein, while l iterates through all vertices belong to the ligand, and d_{pl} is the distance between the considered pair of protein and ligand atoms. It should be noted that we are only interested in the interfacial Delaunay tetrahedrons; therefore for each tetrahedron it contains at maximum of three protein atoms (p) or ligand atoms (l).

The described descriptors were successfully employed in our previous studies to derive classification models that can differentiate native-like docking poses (showing low RMSD to the co-crystallized pose) versus non-native like decoy poses (showing high RMSD to the co-crystallized pose). After a rigorous validation using the 13 DUD datasets, it was shown to significantly improve the traditional physical force field-based hit scoring functions by combining the pose scoring from the classification models [60]. These studies justified the application of PL/MCT-tess descriptors to explore the protein-ligand interface.

Target Specific Classification Models for Discriminating Irrelevant Ligand Docking Pose

A number of recent studies showed that inaccuracy of the scoring functions have become one of the largest obstacles in applying molecular docking in structure-based virtual screening. It is not uncommon to find that the traditional scoring functions cannot identify the native-like ligand pose from the decoy poses, which have comparable binding energies but depicting irrelevant protein-ligand interactions. Figure 4.1A shows the RMSD to the co-crystallized ligand pose versus the energy-based docking scores for 1000 poses of the cognate ligand generated by the Fred program. In Figure 4.1B, 4.1C, 4.1D, five different

beta2-adrenergic receptor inhibitors (including the cognate ligand, noted in red) were docked to a single protein structure (PDB: 2RH1) [233] using the Glide program [51]. The RMSD were then calculated between the docked poses of each ligand with the corresponding co-crystallized pose (PDB: 3D4S, 3NYA, 3NY8, 3NY9) [240,241]. No relationships could be identified between the RMSD to the native-pose and the docking score. (Figure 4.1a, 4.1c) When evaluating the relationship between the RMSD to the x-ray pose and the similarities to the x-ray pose in terms of Euclidean distance in the PL/MCT-tess descriptor space, we observed stronger correlations (r=0.62 versus 0.38 for Glide scores, Figure 4.1d, 4.1c). By analyzing the average RMSD to X-ray poses against distances in the PL/MCT-tess descriptor space suggested that those docking poses with normalized PL/MCT-tess distance lower than -1 are highly likely to be within 4A RMSD to their corresponding x-ray pose. The 4A threshold is usually considered a reasonable indicator to quantify if a ligand pose is located inside its x-ray binding pocket with a roughly correct global orientation [242]. Therefore, in this study, we define the native-like binding pose as those having similar binding mode to the target protein-ligand complex (i.e. with normalized distance to target ligand in PL/MCT-tess descriptor space lower than -1), and thus are potentially showing low RMSD to their own x-ray binding pose. Furthermore, all the docking poses generated by known binding decoys, no matter how close they are to the co-crystallized ligand in the PL/MCT-tess descriptor space, are considered non-native like poses. By such definition, the problem of separating native-like poses versus other decoy poses can be learned through training a standard classification model based on their PL/MCT-tess descriptors. The detailed workflow for selecting the native-like and decoy poses for model's training set is described in Figure 4.2. Because there

88

is no cognate ligand with the vegfr2 and pdgfrb protein targets, the pose with the lowest

MedusaScore was utilized as the native pose [243].

To develop the classification model that can distinguish native-like poses from other irrelevant poses, we used the RandomForest correlation algorithm implemented in the freely-available R package randomForest [17]. RandomForest was known for its excellent prediction accuracy and efficiency at large data sets. It has an internal unbiased estimate of the generalization error during the growth of the forest, thus no additional model validation is necessary. In addition, it provides reasonable flexibility in dealing with imbalanced class ratio by imposing higher penalties on misclassification of the minority class. For details of the RandomForest algorithm, the parameters used, and the evaluating metrics see methods in Chapter 5.

Virtual Screening Protocols Using Post-docking Pose Filters

As described above, the 13 DUD data sets were docked to their corresponding protein target using the Fred docking program, and the default Chemgauss3 scoring function were used to select 30 poses per ligand for further analysis. We therefore calculated the PL/MCT-tess descriptors for each preserved pose to describe its contacts with the protein residues. Three different approaches were used to assign the fitness score of each compound in the final ranking list: (Figure 4.3)

    i.   (*PL/MCT-dist*) All the poses for the same compound were ranked by its similarity to the cognate ligand's pose in terms of distance in the PL/MCT-tess descriptor space. The distance for the most similar pose was used as the final score of this compound in the ranking list. This approach purely uses the information from the PL/MCT-tess descriptors to select pose and rank the compound, thus providing a

direct evaluation on how this descriptors can distinguish ligands from decoys by its own.

ii. (*PL/MCT-similarityFilter*) In this approach, a pose filter is first developed to exclude poses that are too dissimilar to the cognate ligand's pose in terms of distance in the PL/MCT-tess descriptor space. The cutoff for pose similarity is determined by using all the poses for the binding decoys that were in the randomly selected 10% hold-out data. Assuming the distance between the decoy pose to the native pose follows a normal distribution, we defined the distance cutoff as follows: $D_T = \bar{y} + Z\sigma$ where $\bar{y}$ is the average Euclidean distance between each pose to the native pose, and Z is an arbitrary parameter to control the significance level. Here, we set the default value of Z as -0.5 to obtain reasonable pose coverage. The final score of each compound is then determined by the best Chemgauss3 score for the remaining poses.

iii. (*PL/MCT-RFFilter*) This approach is similar to *PL/MCT-similarityFilter* approach in that both approaches use a pose filter to exclude irrelevant poses of a compound and then rank compounds based on their best Chemgauss3 scored poses. The difference in *PL/MCT-RFFilter* is that it employs both the known active and decoys docking poses from the previously excluded 10% hold-out data to derive the pose filter, as described above. By using the machine learning algorithm random Forest, we are thus able to calibrate the contribution of each PL/MCT-tess descriptor type in the resulting pose filter.

Assessment of Virtual Screening Performance

The conventional receiver operating characteristic (ROC) curve was used in this study to evaluate virtual screening protocol's ability to recover known ligands from the top-ranked screening library. In addition to report the area under the ROC curve (AUC) as an indicator for model performance across the entire screening library (see Chapter 3, methods), we used another metric, the ROC Enrichment (ROCE), to quantify model performance at the early stage. ROCE is defined as the ratio of true positive rate to the false positive rate, at a given percentage of recovered known decoys [244]. This metric is similar to the conventional Enrichment Factor metric but is independent of the ratio of decoys to actives in evaluated screening library. We report ROCE values at 1%, 2%, and 5% as recommended in previous publications. To ensure fair comparison, any compound with missing scores, either from removal of all its poses by pose filter or by docking error, were placed at the bottom of the ranked list. To estimate errors, we did 100 independent runs, and then reported the average and standard deviations to approximate the errors. In each run, 10% of the data was randomly selected and used to train the filter, which will later be applied to rank the remaining 90% data. (Figure 4.3)

Randomization Test

To examine the robustness of the proposed pose filters, we generated two random pose filters and compared their performance with the corresponding real pose filters. The first test resembles the Y-randomization test, where we scrambled the activity labels before training the pose filters. While in the second test we keep the activity labels, but randomly shuffling the PL/MCT-tess descriptor profiles. The active to decoy ratio was preserved in both cases. In short, the *PL/MCT-RFFilter* calculations were repeated with the randomized activities or descriptor profiles of the training set. The estimated prediction error from the

developed randomForest models and their virtual screening performance were compared with models generated from real data, and see if there are significant difference.

4.2.3. Results and Discussion

Before applying the virtual screening protocol for any real analysis, I used the benchmarked DUD data sets to assess the performance of the proposed approach. Thirteen datasets with diverse chemical features were selected in this case. For each data set, 10% of decoys and known actives were excluded randomly at each calculation, and used as the calibration set to determine the pose filter as described (methods).

The average number of poses used in the pose filter construction and the estimated prediction errors are shown in Table 4.2. The sensitivity, specificity and CCR were calculated based on the out of bag (OOB) estimates during model generation [17,163]. The results show that the overall accuracy in terms of CCR exceeds 0.80 for all data sets except ace, ache and fxa data sets. The sensitivities are generally lower than specificities, most probably due to the extreme imbalance of the training set. It would be interesting as a future study to implement different sampling approaches to down-size the decoy set before training the randomForest classifier, and see if that improves models' performance. The successful construction of randomForest classifiers for most data sets in part demonstrated that our definition of native-like poses is practically reasonable.

The ROC curves of the individual 13 DUD datasets using Chemgauss3, *PL/MCT-dist*, *PL/MCT-similarityFilter*, and *PL/MCT-RFFilter* were compared in Figure 4.4. The data in Figure 4.4 took the results from one of the 100 repeated studies as an example. Figure 4.5 and 4.6 illustrates model performances in terms of overall ROC-AUC, ROCE at 1%, 2% and 5%. It is obvious from the plots that the *PL/MCT-RFFilter* performs best compared to other

approaches, showing good performance in both overall ranking and early enrichments. For all data set except cdk2 and hivrt, the *PL/MCT-RFFilter* approach gives AUC higher than ranking by the empirical scoring function Chemgauss3 score. On the contrary, we did not observe consistent improvements in terms of ROC-AUC after applying the *PL/MCT-dist* and *PL/MCT-similarityFilter* approach. It is in fact within our expectation that the *PL/MCT-dist* should show inconsistent performance across different data sets. The assumption behind the *PL/MCT-dist* protocol is simple and most often too optimistic: it uses the similarity of a ligand's docking pose to the x-ray crystallized conformation as the only indicator of this pose's fitness in the protein binding pocket. Any other favorable contacts that could lead to gain of energy are underestimated in this case. In compare with the *PL/MCT-RFFilter*, the *PL/MCT-similarityFilter* ignores the set of pre-excluded known actives when generating pose filter and does not acknowledge the distinct importance of each interaction descriptor type for classifying irrelevant versus native-like poses. Therefore, the *PL/MCT-similarityFilter* achieved better performance than using the Chemgauss3 scoring functions for a number of data sets, such as the ache, cox2, and pde5, with the best case of 32.3 fold enrichment at 1% of FPR. But in other cases, *PL/MCT-similarityFilter* has similar or even worse performance. Therefore, we could conclude that the relative better performance of *PL/MCT-RFFilter* can be attributed to its inclusion of more information during pose filter generation stage, which has been extensively discussed in a number of previous publications [60,226,227].

In virtual screening, early recognition is a very critical criterion for evaluating model performances, because only the top fraction of ranked list will be subjected to further experimental validation in practice. The *PL/MCT-RFFilter* approach also showed clear better performance when early enrichment is considered, with the best example (cox2) giving an

enrichment of 56.9 fold at the false positive rate of 1%. Across all the evaluated data sets, the

PL/MCT-RFFilter approach gives a substantially higher number of data sets with ROCE

higher than 15 (10 out of 13 data sets) compared to using docking scores alone (6 out of 13

data sets, Table 4.3). This further demonstrates the advantages of using *PL/MCT-RFFilter*

approach, and the implementation of this interaction will be useful. Furthermore, the

PL/MCT-dist method performs better than using Chemgauss3 scores in five data sets in terms

of ROCE at 1% of false positive rate, especially for difficult cases like ache, pde5, and

pdgfrb. This demonstrates that the statistical information contained in the protein ligand

interaction descriptors alone can be informative in certain cases. It needs to be noted that the

native-pose for pdgfrb and vegfr2 were not from experimentally solved x-ray complex

structure, and that the receptor structure for pdgfrb is a homology model. So it is not

surprising that the Chemgauss3 scoring function did not give good results for these two

targets. The *PL/MCT-RFFilter* approach, however, gives good enrichment of 24.9 and 32.3

fold, respectively, at the false positive rate of 1%. This result suggests that a reasonable

prediction of the native-like ligand pose and including the knowledge of the critical protein-

ligand contacts can rescue virtual screening performance for likely inaccurate receptor

structural models.

To demonstrate the virtual screening results from *PL/MCT-RFFilter* approach is not

by chance, we applied two types of fake pose filters during virtual screening and compared

their performances with the pose filters constructed from the real data. These two artificial

sets differ in whether the descriptor profiles are randomly shuffled or the activity labels are

randomly shuffled. The pose filter trained from randomized descriptor or activity profile

showed close to random classification accuracy, practically put every instance to the

dominant class, the decoy class. As shown in Figure 4.7, the virtual screening performance derived from scrambling the activity files will result in inseparable class distributions, leading to a pose filter that will classify any pose as decoys and eliminate them from further evaluation. On the other hand, the pose filter generated from randomly shuffled the descriptor files is able to derive a tentative pose filter, but this information is not relevant to the selection of native-like poses. As illustrated in Figure 4.7, this filter excluded indiscriminately of a fraction of the entire collection of poses, resulting in a similar or worse virtual screening performance as using the docking scores alone. In summary, this randomization test showed that the information extracted by the filter is not spurious, and could effectively guide the selection of native-like pose during virtual screening.

4.3. Virtual Screening of Subtype Selective Ligands for the Beta Adrenoceptor Protein Family

The major focus for this study is to evaluate whether the information from PL/MCT descriptors are capable of improving virtual screening of subtypes selective ligands. For the βAR protein family, there have been extensive studies on putative recognition sites in the binding pocket that can differentiate the binding propensities of a small molecule [245-247]. General docking programs that evaluate ligands fitness solely based on energy considerations thus have difficult to discriminate these types of interactions. Imposing a number of intermolecular or pharmacophoric constraints can partially solve the problem, but is often too restrictive and yield few interesting hits. The PL/MCT-tess descriptors provide an option to include ligand information automatically during training the pose filter, for example, by including interfacial descriptors for non-selective ligands in the calibration set. However, it should be employed with great cautious, since it represents a much challenging task than

differentiate binding decoys or geometrical decoys from the active compounds, and requires discerning more delicate preferences of the different protein-ligand contacts. We have carried out a systematic study on a large collection of published human β1AR, β2AR, and β3AR structural models as well as two β2AR crystal structures, i.e., 2RH1 (released by the RCSB Protein Data Bank (PDB) on Oct. 30, 2007) and the latest, 3P0G (released by RCSB PDB on Jan. 19, 2011) (Table 4.4). The *de novo* designed structural models were collected from the same group whose β2AR model has shown a comparable or better virtual screening performance to the X-ray structures (Chapter 3). The five homology models were retrieved from the well-recognized modeling server ModWeb [248] and GPCRDB [249]. Both agonist and antagonist bound models were included when possible to account for any structural features associated with functional activity.

4.3.1. *Small Molecule Data Set*

To perform any meaningful evaluations of the virtual screening performance for subtype selective binders within the beta adrenergic receptors family, we will need a reasonably large and accurate dataset representing different types of compounds targeting the beta-adrenergic receptor family. A dataset of 207 compounds were collected from the ChEMBL database, incorporating the data published from 1996 through 2009. Although the binding data were extracted from multiple sources, the protocols for binding assay were kept similar. Among all the binding data, only measurements using I-CYP as the radio-labeled ligand were considered. For purpose of identify subtype selective ligands, the compounds included in the final data set are mostly tested against two of three or all three subtypes of adrenergic receptors. (Exceptions were made for 11 compounds that are β2AR non-binders. They were included later during the QSAR modeling process to improve models'

classification accuracy). In summary (Figure 4.8), for each subtype of βAR, there are 196 compounds in β1AR data set (Ki from 0.017nM to 1mM) ; 207 compounds in β2AR data set (Ki from 0.14nM to 0.1mM) ; and 166 compounds in β3AR data set (Ki from 1.1 nM to 0.15mM). To simplify the situation, we did not distinguish between agonist, inverse agonist, and neutral antagonist within binders. This is also a legitimate approximation, because as shown in our previous studies, as well as in others' work, the crystal structures represent a putative early activation stage, and is capable of retrieving both agonists and antagonists [58,250-252]. In addition, data for function assays often varies in experimental settings, making it difficult for a direct comparison. However, it would be definitely interesting in the future to explore the applicability of the described approach in predicting the overall complex pharmacological profiles of β adrenoceptor ligands.

To date, NC-IUPHAR (International Union of Basic and Clinical Pharmacology Committee on Receptor Nomenclature and Drug Classification) has not published any recommendations to define a compound as selective ligand for a given receptor subtype within the protein family. However, although the selectivity varies based on the receptor types in consideration, a compound can be considered selective if its ratio of binding affinity is higher than 100 times [253]. Here, due to the limited data sources for highly selective βAR ligands, we define a compound as selective if it has a modest selectivity and binding affinity: difference in pKi ≥1.5, and pKi >=6 for the desired subtype. By this definition, there are 20 β1AR selective ligands, 8 β2AR selective ligands, and 12 β3AR selective ligands. We should note that the available β2 selective ligands are generally in a lower activity range, it represents a difficult case in this task.

4.3.2. Methods

Drug-like Decoys Set:

A standard decoy set was downloaded from Schrodinger's website. It contains 1000 decoys with drug-like properties. Unlike the DUD data sets where the decoys were specially designed to challenge the docking algorithms, the compounds from this decoy set resemble more of a real virtual screening campaign, where non target-specific, drug-like compounds were used.

Protein Set Up and Docking Programs

The protein structures were prepared with Glide protein preparation wizard to add missing hydrogen and assign partial charges (see Chapter 3). We used two different popular docking programs, the Glide (methods in Chapter 3) and Fred (see above in Pilot study). Both of the implemented scoring functions (Chemgauss3 and GlideScore) belong to the class of empirical scoring functions. Because Glide docking pose with GlideScore showed consistent better performance (0~5 fold better in ROCE at 1% of specificity) than Chemgauss3, I only report here the results from Glide docking.

Assessment of Virtual Screening Performance

Same performance evaluation metrics were used as in the pilot study.

4.3.3. Results and Discussion

Virtual Screening Using Pose Filters Generated from Known Inactives

Initially, we tested whether the proposed approaches are able to identify active ligands from known inactives as well as drug-like decoys. To generate the pose filters, the binding pose of a co-crystallized ligands (for $\beta 2Xtal_i$, $\beta 2Xtal_a$), the binding pose derived from a simulation study (for $\beta 1EPI_a$, $\beta 2TA_a$, $\beta 3AJ_a$), or the top Glide-score pose of a selected

potent ligand (for $\beta1GPCRDB_i$, $\beta1ModWeb_i$, $\beta1ModWeb_a$, $\beta3GPCRDB_i$, $\beta3ModWeb_i$) was employed as the reference pose, while the known inactives for each receptor type were utilized as the calibration set to derive the pose filters. Because the Glide docking was observed with good fidelity in pose predictions with our previous studies as well as other's published work, it is reasonably safe to use the top-scored binding pose of the reference ligand from Glide docking to set the reference descriptors. Table 4.5 listed the information of the number of seeds and decoys used for each receptor model. Figure 4.9 shows the model performance using the default empirical scoring function GlideScore, versus the PL/MCT-dist approach, PL/MCT-similarityFilter approach, and the PL/MCT-RFFilter approach as defined in the pilot study. The improvements from applying PL/MCT-similarityFilter or PL/MCT-RFFilter, however, were only marginal for most receptors, with the best case of $\beta1ModWeb_a$ where the PL/MCT-RFFilter increases the ROCE at 1% from 4-fold to 8-fold compared with the Glide scores. There are several possible reasons for this marginal improvement in compare with those observed in the pilot study. First the poses selected via Glide score could be generally more native-like than the poses selected via Chemgauss3 score [242,254], thus the improvements from applying a native-like pose classifier (pose filter) would be less prominent. This is highly likely the major cause, because we observed consistently better virtual screening performance of Glide score versus Chemgauss3 (up to 200% increase in ROCE at 1%, data not shown). This is consistent to our previous observations [60], suggesting that the outcomes of applying the target-specific pose filter is dependent on the parent scoring function used. Second, we used a set of known inactives to generate the pose filter, instead of using the "putative inactive" decoys as in the pilot study. This is good in a sense that those compounds are true inactives, so all the docking poses

generated from this set can be safely put into the class of irrelevant binding poses. However, this set consists of relatively small number of chemical scaffolds, leading to a more biased calibration set and pose filters with less general applicability domain.

Besides evaluating proposed approaches' ability at identifying active compounds, we also evaluated how good they can enrich selective ligands at early stage. As we expected, all structural models except for $\beta2Xtal_a$ showed no preference for subtype selective ligands. The virtual screening results showed a much lower enrichment factor on selective ligands versus all known actives. Similarly, because the pose filters were not designed to differentiate subtype selective ligands, neither of the *PL/MCT-similarityFilter* and *PL/MCT-RFFilter* help to enrich subtype selective ligands. On the other hand, the PL/MCT-dist approach showed surprisingly good virtual screening performance on enriching selective ligands when using $\beta1GPCRDB_i$, $\beta1ModWeb_i$, and $\beta2Xtal_i$ models. This suggests that for certain receptor model and reference ligand combinations, the simple distance calculation between a docking pose and the reference pose could provide reasonable judgment as how likely the pose represents a selective ligand. More detailed discussions on how the selection of reference ligand could influence the virtual screening performance are provided in the next section.

Virtual Screening of Selective Ligands using Subtype-biased Pose Filters

In this study, we evaluated the possibility of selectivity biased pose filter by using the subtype selective ligand as reference and including the non-selective ligands into the calibration set when deduce the pose filters (Figure 4.2). Similar to what have been described in the previous section, the PL/MCT-tess descriptors derived from the top GlideScore pose of the reference ligand was used to define the native-like pose during the pose-filter training stage. Here, the two classes employed to derive randomForest models become putative

native-like poses of known selective ligand versus all poses of known non-selective and inactives. It should be emphasized again that all the pose information used during model training come from the hold-out set, which has been excluded from later model evaluation. As an example, Figure 4.10 showed the virtual screening performance using $\beta1ModWeb_i$, $\beta2Xtal_a$, and $\beta2TA_a$ receptor models. Comparing Figure 4.10F with Figure 4.9B on the performance using $\beta2TA_a$ model, it is clear that the virtual screening protocol of *PL/MCT-similarityFilter and PL/MCT-RFFilter* in combination with the subtype-biased setting (Figure 4.9 B) has the best ability to enrich $\beta2$ selective ligands. The two approach enabled a favorable ROCE reaching approximately 16-fold and 8-fold, respectively, over random at a false positive rate of 1%. Considering the fact that the reference ligand ($\beta2$ selective agonist TA2005) used are same between these two tests, the different results suggest that using the subtype biased pose-filters are indeed helpful in discriminating non-selective ligands from selective ones. The performance on enriching overall $\beta2$ active ligands did not change very much between these two experiments, showing that the employed pose-filters only focus on the desired subset of ligand types. In addition to the implementation of a subtype biased pose-filter, the choice of reference ligand is also essential on the virtual screening results. When $\beta1ModWeb_i$ receptor structure model was used for docking and the non-selective $\beta1$ active compound Cyanopindolol was used as the reference, there is negligible difference between using the non-biased and subtype-biased pose filter (Figure 4.10A versus Figure 4.9A). To further explore the influence of the choice of reference ligands, we used a $\beta2$ selective binder TA-2005 in combination with $\beta2Xtal_a$, and repeated the pose filtering and scoring protocol (Figure 4.10D versus 4.10E). Both *PL/MCT-*

101

similarityFilter and *PL/MCT-RFFilter* showed over 50% increase in early enrichment of β2 selective ligands.

As observed in Chapter 3, the structure models derived directly from homology modeling are usually having a suboptimal set of conformation of its binding pocket residues, making it difficult to establish favorable contact between receptor and known ligands. Although for each homology model (β1GPCRDB$_i$, β1ModWeb$_i$, β1ModWeb$_a$, β3GPCRDB$_i$, β3ModWeb$_i$) we already adapted the receptor side chain conformations to a known potent ligand through the protein preparation wizard in Glide, their virtual screening outcomes are still unsatisfactory. There have been plenty studies trying to optimize structure models after homology modeling in order to promote ligand binding [58,251,252,255-257]. In this study, we did not change the global packing of the trans-membrane regions or their relative orientations, but performed induced fit docking with selected ligands in hope of deriving receptor binding pocket models that are more appropriate for ligand binding. Briefly, the homology models and ligands tested were β1ModWeb$_i$ (with β1 selective ligand ICI89406), β1ModWeb$_a$ (with β1 selective ligand Dobutamine) and β3ModWeb$_a$ (with β3 selective ligand AJ-9677). The derived top-scored 10 receptor-ligand complexes were manually inspected and one protein-ligand complex was selected for further virtual screening analysis. Structure conservation on those experimentally confirmed receptor-ligand interactions such as the Asp113$^{3.32}$ and Asn312$^{7.39}$ were closely examined. As a result, the shape of the selected binding pocket and the orientation of key residues are adapted to a known potent and subtype selective ligand. It should be noted that we only use induced fit docking in preparing the receptors and that the receptor side chains were kept rigid during the process of virtual screening. Take the β1ModWeb$_i$_refined model for example (Figure 4.10C GlideScore versus Figure 4.10B,

GlideScore), the customized binding pocket showed better recognition for β1 active as well as selective ligands, with an increased enrichment (ROCE at 1%) from 1.5 to 12.5 for actives and 0 to 15.1 for selective ligands. Applying the *PL/MCT-dist*, *PL/MCT-similarityFilter* and *PL/MCT-RFFilter* protocols on the docking poses further augmented the enrichment of selective ligands to 25.0-fold , 25.1-fold and 20.7-fold over random, respectively.

The Molecular Basis of Subtype Selectivity from Identified Important PL/MCT-tess Descriptors

Using the RandomForest pose filter allows a systematic analysis of the importance of each protein-ligand interaction descriptors. We then counted the frequency of descriptors out of the 100 runs for their occurrence at the top 20 important descriptors. As an example we showed in Figure 4.11 for three most frequently observed descriptor types when using the combination of β2Xtal$_a$ receptor models with the β2-selective ligand TA2005. The ligand TA-2005 has a selectivity ratio of ~50 and ~150 for β2- versus β1 and β3 receptors, respectively [245,258]. Three of the frequent observed PL/MCT-tess descriptors observed for TA-2005 are ClNlOlOr, ClOlNrOr, and ClClOlNr (Figure 4.11). The ClNlOlOr descriptor depicted the salt bridge between TA-2005 with Asn312$^{7.39}$, and the hydrogen bonds formed with Ser203$^{5.42}$ and Ser204$^{5.43}$. These residues have been shown to be important in binding b2AR agonists. The ClOlNrOr descriptor highlighted two other critical protein ligand interactions with Tyr308$^{7.35}$ and Asn293$^{6.55}$, both of which are believed to be involved in determining ligand selectivity [246]. The ClClOlNr descriptor mapped an interesting contact between the N-substituent of the ligand with Lys305. This residue is an aspartic acid (Asp356) in β1, and a glycine (Gly325) in β3. These three residues are significantly different in size, and charge state, thus could have potential impact on subtype selectivity.

103

4.3.4. Conclusions

In summary, we have proposed a general pose filter using the DFT derived protein-ligand interaction descriptors (PL/MCT-tess) in structure-based virtual screening. The pose filter was validated with the 13 benchmarked DUD datasets with Fred docking program. The pose filter enables consistent improvement over using the default scoring functions by Fred for up to 6 fold in terms of ROCE at 1% of false positive rate. We further challenged the pose filter to identify subtype selective beta-adrenergic ligands. The results showed that it is possible to achieve reasonable results with carefully designed subtype biased pose filters, preferably using a selective ligand as the reference and a calibration set including both known nonselective and inactive molecules. In addition, we found that the homology models showed poor virtual screening performance in general, but can be further improved with an optimized conformation of the binding site residues. For instance, after optimize the binding pocket of the homology model β1ModWeb$_i$ with a selective β1AR ligand ICI89406, we observed an increase of ROCE from 0 to 15 fold over random at a false positive rate of 1%. It should be noted that the success of the proposed approach is heavily dependent on the previous knowledge, especially the selection of proper reference ligands to guide the structure modeling as well as docking pose selection. In future studies, it would be interesting to see how additional knowledge about the receptors, such as the conformational flexibility of the binding pockets can be employed into structure-based docking, and affect the virtual screening performance.

Tables and Figures
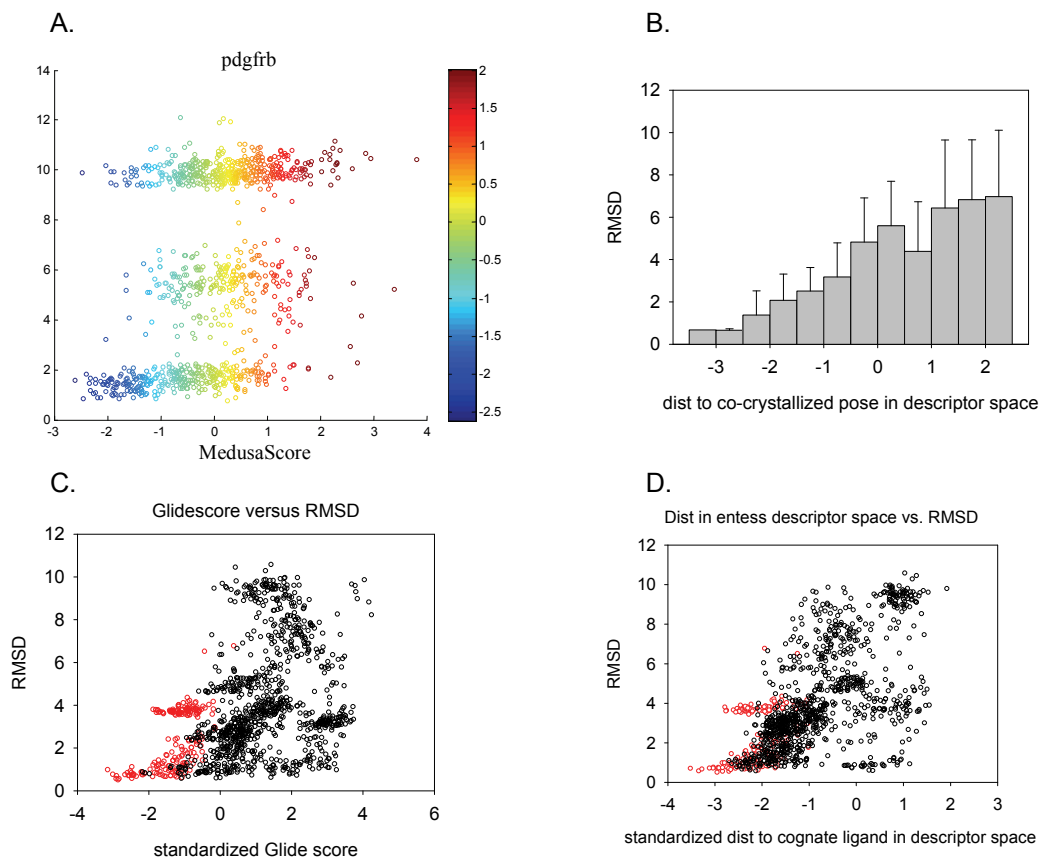
Figure 4.1. The distribution of docking poses' RMS deviation from the X-ray pose versus similarity of PL/MCT-tess descriptors calculated by Euclidean distances. A) for pdgfrb homology model and B-D) for PDB 2RH1. A) The pose with the best MedusaScore is served as the reference to calculate the RMSD value of poses. The plot shows the pose distribution based on RMSD values (y-axis) versus Z-score values of MedusaScore (x-axis, Z-score is calculated based on MedusaScore distribution of the decoy sets). B) Average RMSD and standard deviation values for increasing distance to the reference pose in PL/MCT-descriptor space. A total of 1000 poses from 5 X-ray co-crystallized ligands were used. The distances are converted to Z-scores based on distribution of distances calculated by a drug-like decoy set to the X-ray pose of PDB 2RH1. (see Table 4.5 for detail information of the decoy set). C) The pose distribution based on Z-score values of GlideScore (x-axis) *vs.* RMSD values (y-axis). D) The pose distribution based on Z-score values of distance to the native pose in PL/MCT-tess descriptor space (x-axis) *vs.* RMSD values (y-axis). The data points for the cognate ligand in PDB 2RH1 is indicated with red circles. Plot A. adapted from Hsieh et al.[61].
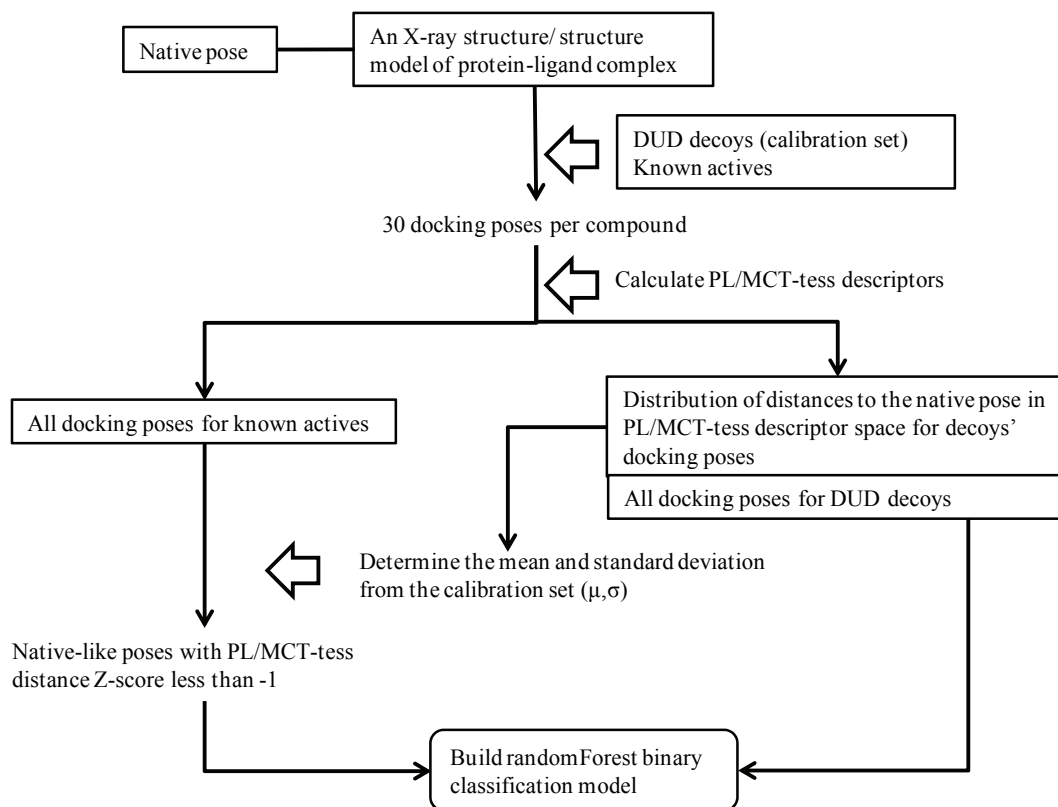
Figure 4.2. Flowchart of pose selection in training the proposed target-specific pose filter, PL/MCT-RFFilter.
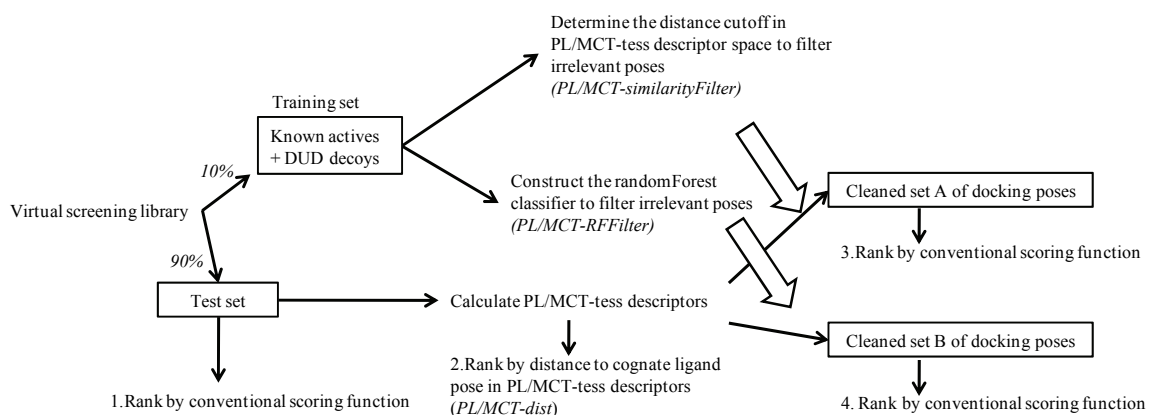
Figure 4.3. Flowchart of using the three described *PL/MCT-dist*, *PL/MCT-similarityFilter*, and *PL/MCT-RFFilter* approaches in structure-based virtual screening.
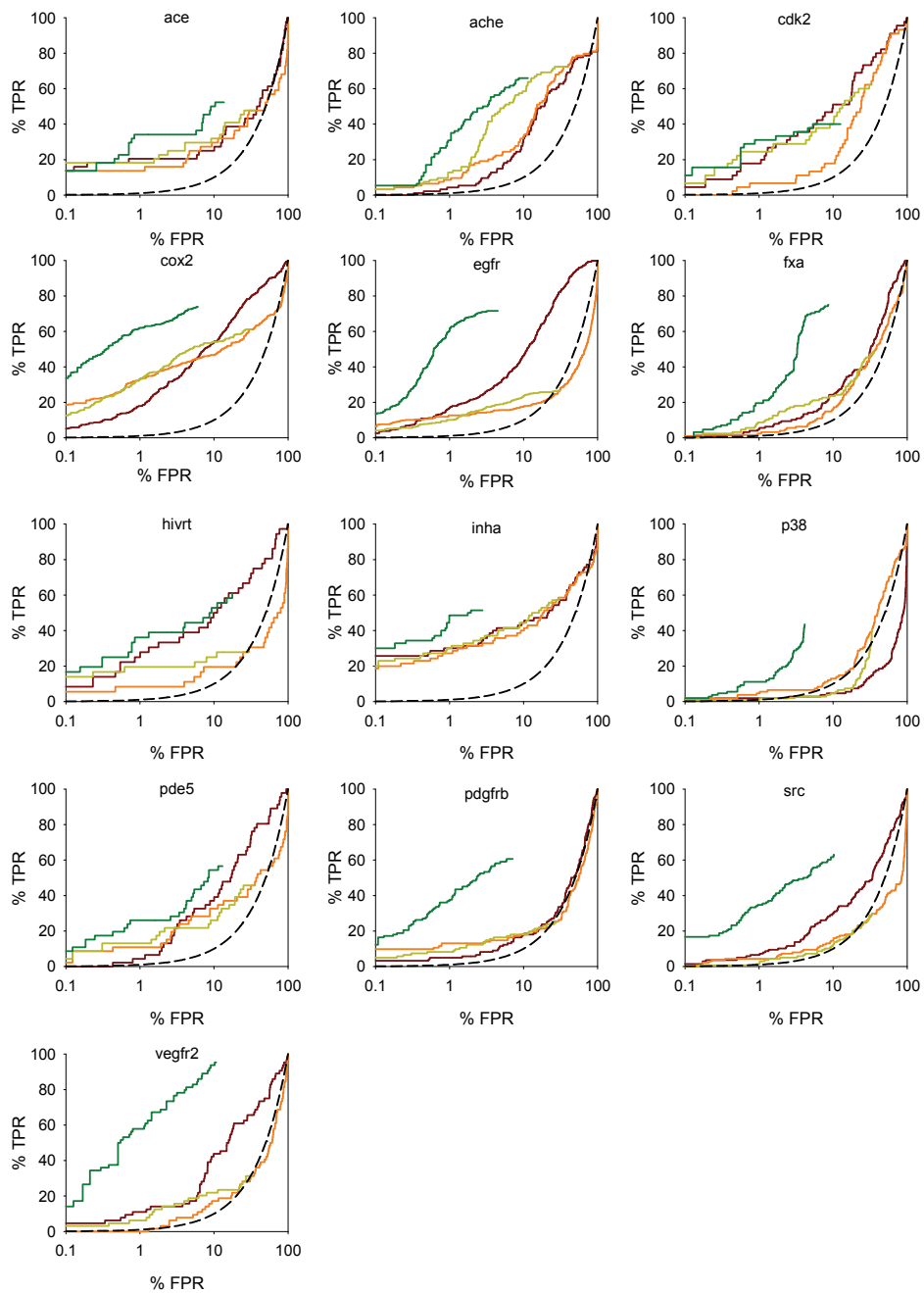
Figure 4.4. ROC enrichment plots for each of the 13 DUD data set. Dark red lines are based on Chemgauss3 scores, orange lines are based on *PL/MCT-dist* score, yellow lines are based on *PL/MCT-similarityFilter* approach, and green lines are based on *PL/MCT-RFFilter* approach.
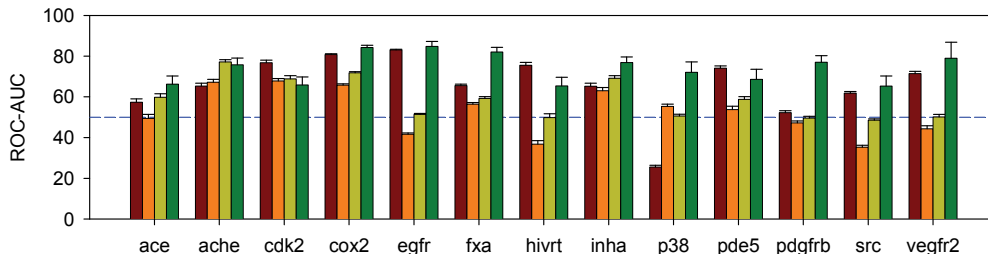
Figure 4.5. ROC-AUC for each of the 13 DUD data set using Chemgauss3 scores (dark red), *PL/MCT-dist* (orange), *PL/MCT-similarityFilter* approach (yellow), and *PL/MCT-RFFilter* approach (green).
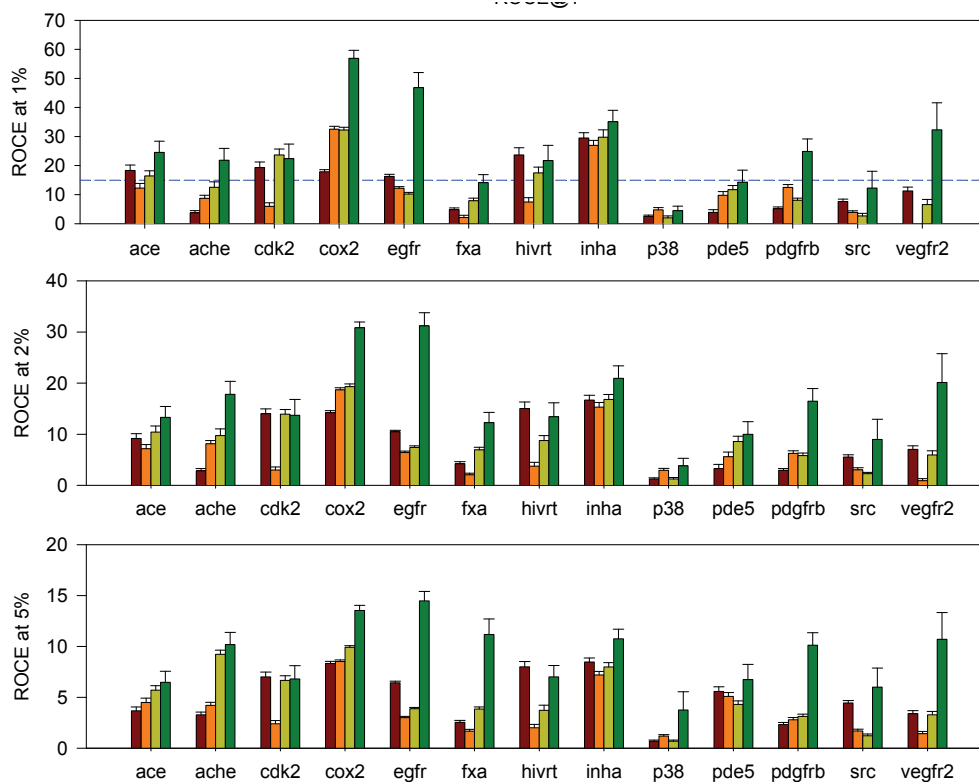
Figure 4.6. ROC enrichment for each of the 13 DUD data set using Chemgauss3 scores (dark red), *PL/MCT-dist* (orange), *PL/MCT-similarityFilter* approach (yellow), and *PL/MCT-RFFilter* approach (green).

E. ROC-AUC

F. ROCE at 1%
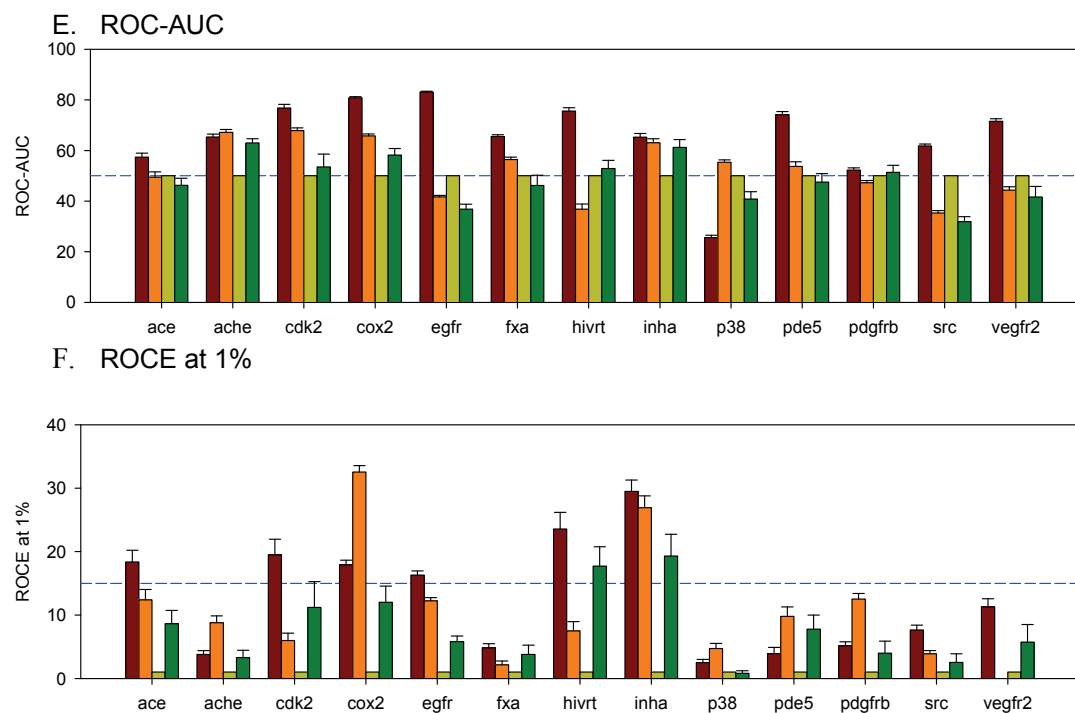
 Figure 4.7. The ROC-AUC metric and ROC enrichment (ROCE) for each of the DUD data set with randomized activity/descriptor profiles. The results using Chemgauss3 scores (dark red) and *PL/MCT-dist* (orange) based on non-randomized profiles were regenerated from Figure 4.5, Figure 4.6 for comparison. Yellow bars represent *PL/MCT-RFFilter* approach with scrambled activities. Green bars represent *PL/MCT-RFFilter* approach with scramble descriptor profiles.
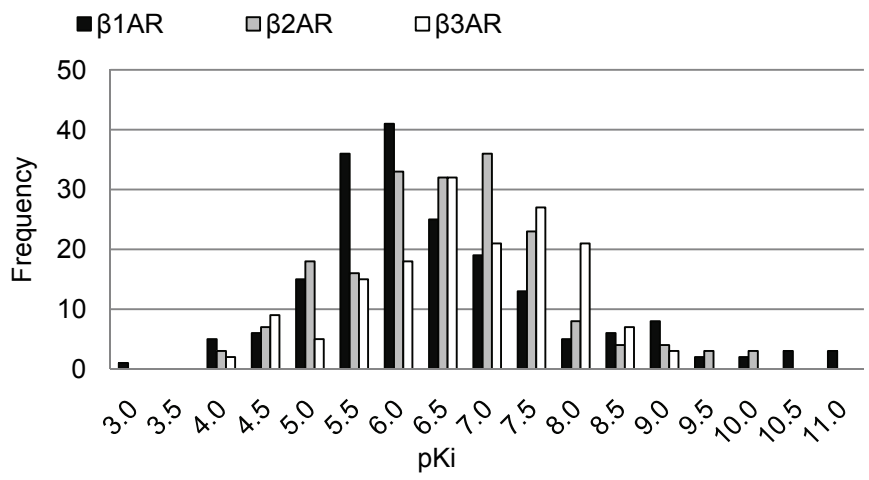
Figure 4.8. Activity distributions of the βAR data set.

A. b1AR-AUC   B. b2AR-AUC   C. b3AR-AUC

GlideScore/active cp.
GlideScore/selective cp.
PL/MCT-dist/active cp.
PL/MCT-dist/selective cp.
PL/MCT simpleFilter/active cp.
PL/MCT simpleFilter/selective cp.
PL/MCT RFFilter/active cp.
PL/MCT RFFilter/selective cp.
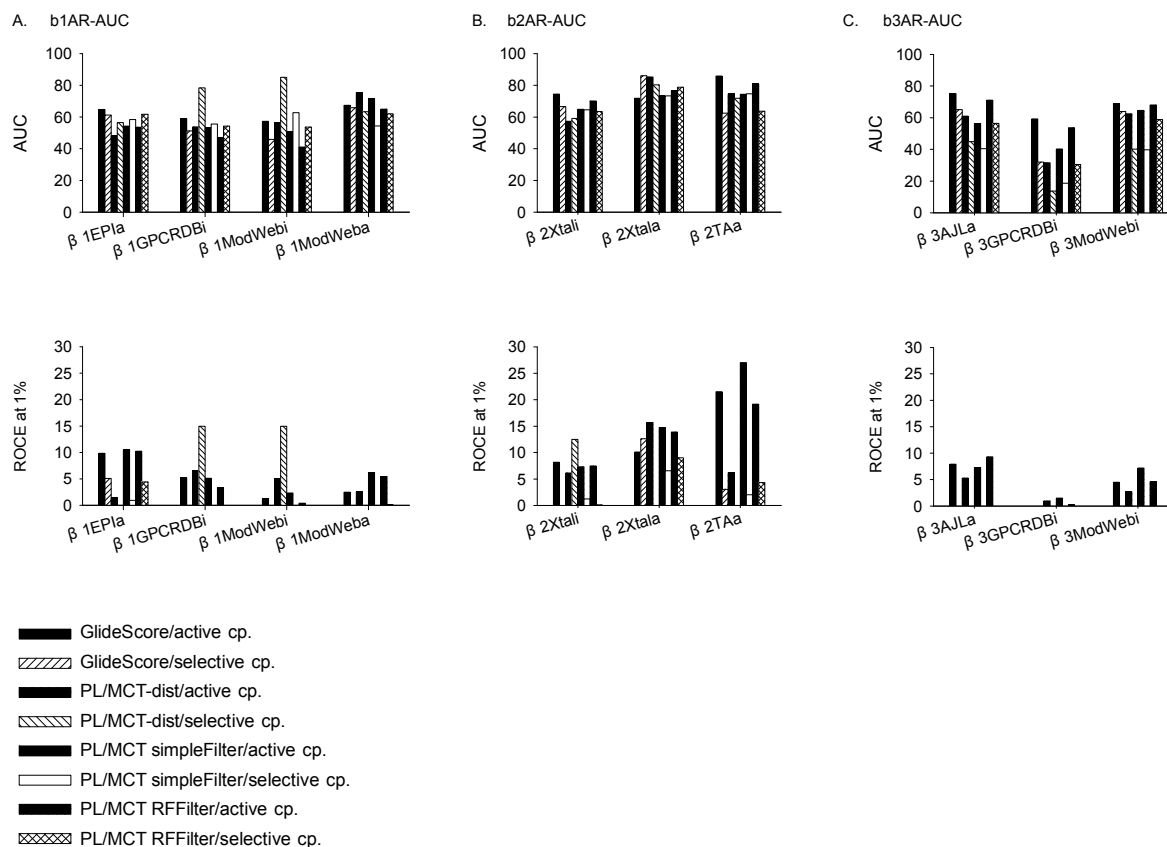
Figure 4.9. The ROC-AUC metric and ROC enrichment (ROCE) for each of the βAR receptor type and structure model.

111
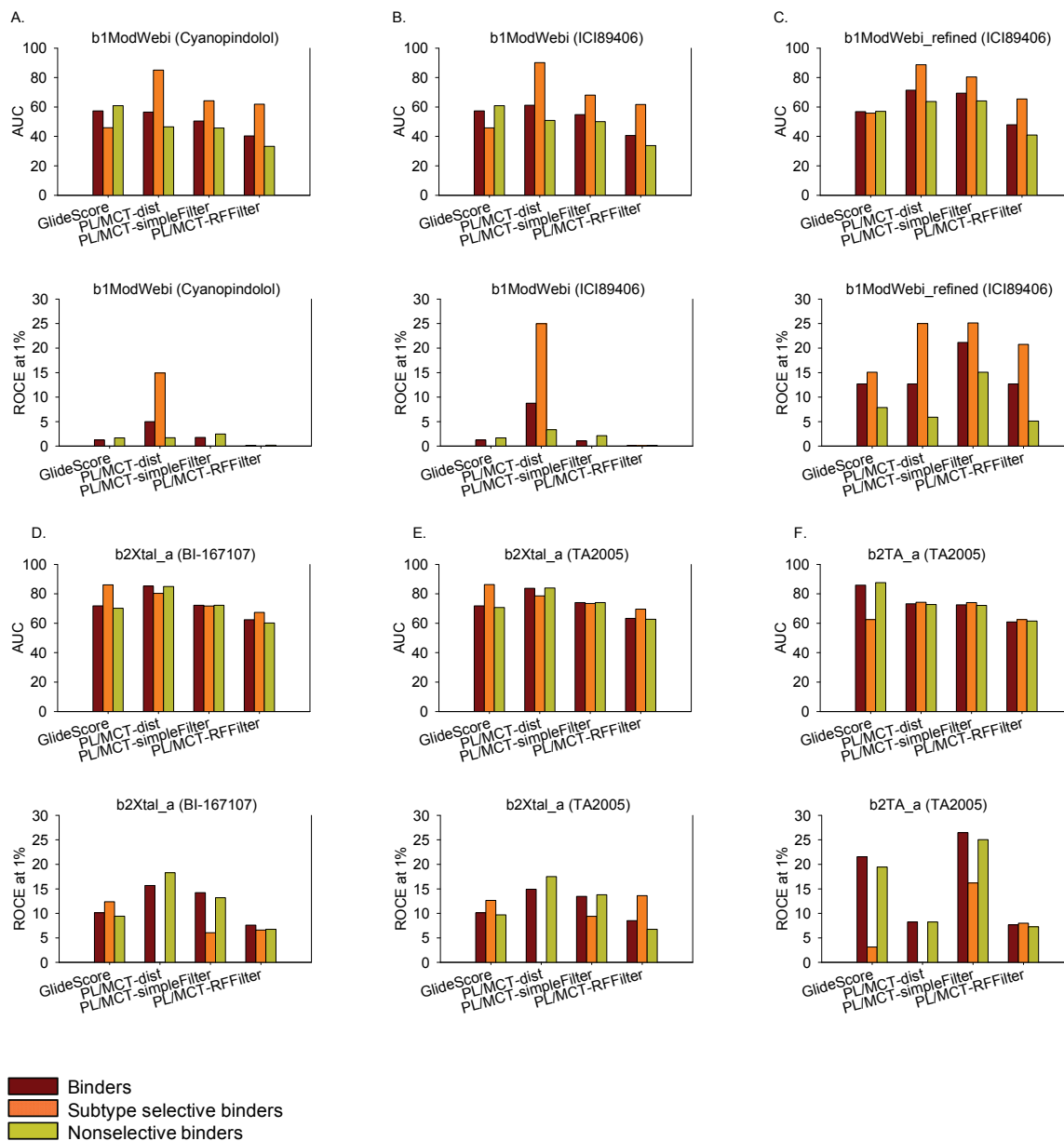
Figure 4.10. The ROC-AUC metric and ROC enrichment (ROCE) for each of the βAR receptor type and structure model with subtype biased pose filters.

A. ClNlOlOr

Asn312$^{7.39}$

Ser204$^{5.43}$

Ser203$^{5.42}$

Asp113$^{3.32}$

B. ClOlNrOr

Tyr308$^{7.35}$

Asn293$^{6.55}$

Asn312$^{7.39}$

Ser204$^{5.43}$

C. ClClOlNr

Lys305$^{7.32}$
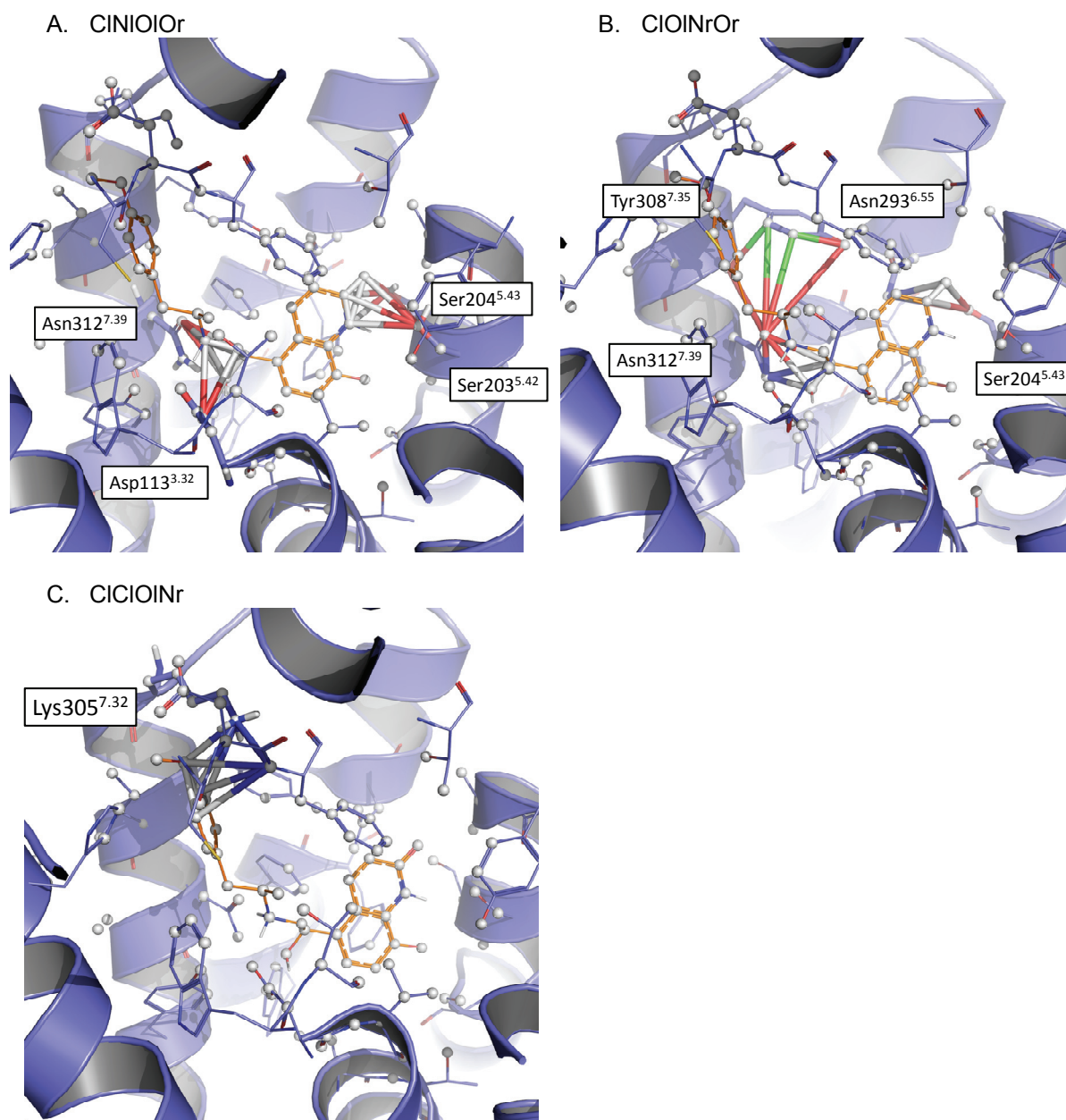
Figure 4.11. Important PL/MCT-tess descriptors summarized from subtype biased *PL/MCT-RFFilter* using β2Xtal$_a$ structure model and TA-2005 as the reference ligand. The β2Xtal$_a$ receptor structure is colored in blue, while the β2 selective ligand is colored in orange. The tetrahedrons corresponding to the specified PL/MCT-tess descriptor are shown in stick.

Table 4.1. Summary of benchmark data sets used for pilot studies. Table adapted from [60]

Target	Function	PDB	# of ligands	# of decoys	# of clusters
ace	metallopeptidase	1o86	46	1726	19
ache	acetylcholine esterase	1eve	99	3631	19
cdk2	serine/threonine kinase	1ckp	47	1776	32
cox2	cyclooxygenase	1cx2	212	11841	44
egfr	tyrosine kinase	1m17	365	14516	40
fxa	serine protease	1f0r	64	1888	19
hivrt	HIV reverse transcriptase	1rti	34	1415	17
inha	enoyl ACP reductase	1p44	57	2501	23
p38	serine/threonine kinase	1kv2	137	6230	20
pde5	phosphodiesterase	1xp0	26	1562	22
pdgfrb	tyrosine kinase	modela	124	5265	22
src	tyrosine kinase	2src	98	5216	21
vegfr2	tyrosine kinase	1vr2b	48	2479	31

[a]: protein structure is homology model, the ligand structure is taken from the DUD website
[b]: apo structure, the ligand structure is taken from DUD website
HIV: Human Immunodeficiency Virus; ACP: Acyl Carrier Protein

Table 4.2. Statistics of target-specific pose filter *PL/MCT-RFFilter* calculated based on out of bag (OOB) estimates. Data is based on the average statistics derived from 100 runs of 10% randomly selected hold-out set.

	TP	FP	TN	FN	SE	SP	CCR
ace	71.71	1.76	5389.24	76.29	0.48	1.00	0.74
ache	77.04	1.38	7782.62	137.96	0.36	1.00	0.68
cdk2	151.74	1.50	6180.30	54.16	0.73	1.00	0.86
cox2	1070.98	7.72	39844.28	208.02	0.84	1.00	0.92
egfr	191.24	1.20	47972.20	79.11	0.70	1.00	0.85
fxa	209.17	7.01	17150.79	229.83	0.48	1.00	0.74
hivrt	95.93	0.55	4556.45	34.07	0.74	1.00	0.87
inha	147.89	1.72	9796.28	90.11	0.62	1.00	0.81
p38	226.89	3.26	7787.14	41.01	0.83	1.00	0.91
pde5	214.79	1.62	5926.38	50.21	0.79	1.00	0.90
pdgfrb	274.52	0.54	13159.06	80.38	0.77	1.00	0.88
src	355.54	1.04	18944.26	95.36	0.79	1.00	0.89
Vegfr2	198.57	1.69	8699.01	57.43	0.77	1.00	0.88

Table 4.3. The ROC enrichment (ROCE) and maximum actives coverage for each of the DUD data set.

		ROC-AUC	1%	2%	5%	maximum % of actives recovered
ace	Fred-chemgauss3	57.4	18.4	9.2	3.7	100
	PL/MCT	49.5	12.3	7.1	4.5	100
	Fred-PL/MCT similarityFilter	59.8	16.4	10.4	5.7	47.7
	Fred-PL/MCT RFFilter	66.2	24.5	13.3	6.5	52.3
cdk2	Fred-chemgauss3	76.8	19.4	14.0	7.0	100
	PL/MCT	67.8	6.0	3.0	2.4	100
	Fred-PL/MCT similarityFilter	68.8	23.6	13.9	6.6	62.2
	Fred-PL/MCT RFFilter	65.8	22.4	13.7	6.8	40.0
egfr	Fred-chemgauss3	83.1	16.3	10.5	6.4	100
	PL/MCT	41.7	12.2	6.4	3.0	100
	Fred-PL/MCT similarityFilter	51.4	10.2	7.4	3.9	26.7
	Fred-PL/MCT RFFilter	84.9	46.8	31.2	14.5	95.6
hivrt	Fred-chemgauss3	75.5	23.7	15.0	8.0	100
	PL/MCT	36.8	7.5	3.8	2.0	100
	Fred-PL/MCT similarityFilter	49.8	17.5	8.8	3.7	27.8
	Fred-PL/MCT RFFilter	65.4	21.8	13.4	7.0	61.1

		ROC-AUC	1%	2%	5%	maximum % of actives recovered
ache	Fred-chemgauss3	65.3	3.8	2.9	3.3	100
	PL/MCT	67.2	8.8	8.1	4.2	100
	Fred-PL/MCT similarityFilter	77.2	12.5	9.7	9.2	72.3
	Fred-PL/MCT RFFilter	75.7	21.8	17.8	10.2	66.0
cox2	Fred-chemgauss3	80.8	17.9	14.2	8.3	100
	PL/MCT	65.8	32.5	18.7	8.5	100
	Fred-PL/MCT similarityFilter	71.8	32.3	19.3	9.9	61.0
	Fred-PL/MCT RFFilter	84.2	56.9	30.8	13.5	73.8
fxa	Fred-chemgauss3	65.5	4.8	4.2	2.5	100
	PL/MCT	56.4	2.2	2.1	1.7	100
	Fred-PL/MCT similarityFilter	59.3	7.9	6.9	3.8	48.4
	Fred-PL/MCT RFFilter	82.1	14.1	12.3	11.2	75.0
inha	Fred-chemgauss3	65.3	29.5	16.7	8.5	100
	PL/MCT	63.0	27.0	15.3	7.2	100
	Fred-PL/MCT similarityFilter	69.1	29.8	16.8	8.0	58.6
	Fred-PL/MCT RFFilter	76.9	35.1	20.9	10.7	51.4

Table 4.3. The ROC enrichment (ROCE) and maximum actives coverage for each of the DUD data set. (cont.)

		ROC-AUC	1%	2%	5%	maximum % of actives recovered
p38	Fred-chemgauss3	25.6	2.5	1.3	0.7	100
	PL/MCT	55.4	4.8	2.9	1.2	100
	Fred-PL/MCT similarityFilter	50.7	2.1	1.3	0.7	40.7
	Fred-PL/MCT RFFilter	72.1	4.5	3.8	3.7	43.5
pdgfrb	Fred-chemgauss3	52.3	5.1	2.9	2.3	100
	PL/MCT	47.2	12.5	6.2	2.8	100
	Fred-PL/MCT similarityFilter	49.6	8.1	5.8	3.1	25.4
	Fred-PL/MCT RFFilter	77.0	24.9	16.4	10.1	60.7
vegfr2	Fred-chemgauss3	71.5	11.3	7.0	3.4	100
	PL/MCT	44.4	0.0	0.9	1.4	100
	Fred-PL/MCT similarityFilter	50.1	6.6	5.9	3.3	31.3
	Fred-PL/MCT RFFilter	79.0	32.3	20.1	10.7	95.3

		ROC-AUC	1%	2%	5%	maximum % of actives recovered
pde5	Fred-chemgauss3	74.1	3.9	3.3	5.6	100
	PL/MCT	53.7	9.8	5.6	5.1	100
	Fred-PL/MCT similarityFilter	58.8	11.7	8.6	4.3	43.5
	Fred-PL/MCT RFFilter	68.7	14.3	10.0	6.7	56.5
src	Fred-chemgauss3	61.8	7.6	5.5	4.4	100
	PL/MCT	35.3	3.9	3.0	1.7	100
	Fred-PL/MCT similarityFilter	48.5	2.7	2.3	1.2	29.7
	Fred-PL/MCT RFFilter	65.3	12.3	9.0	6.0	63.0

Table 4.4. The synonym of ten human βAR structural models employed in this study.

Receptor subtype	Models name	Model type	Details
β1AR	β1EPI$_a$	*de novo* model	In complex with endogenous βAR agonist epinephrine (T. Lybrand group)
	β1GPCRDB$_i$	Homology model	Based on the crystal structure of inactive state turkey β1AR (PDB 2VT4) (retrieved from GPCRDB website, constructed in G. Vriend group)
	β1ModWeb$_i$	Homology model	Based on the crystal structure of inactive state turkey β1AR (PDB 2VT4) (retrieved from ModWeb server, provided by A. Sali group)
	β1ModWeb$_a$	Homology model	Based on the crystal structure of active state turkey β1AR (PDB 2Y00) (retrieved from ModWeb server, provided by A. Sali group)
β2AR	β2TA$_a$	*de novo* model	In complex with selective β2AR agonist TA2005 (T. Lybrand group)
	β2Xtal$_i$	X-ray crystal structure	Crystal structure of human β2AR (PDB 2RH1) in complex with non-selective antagonist carazolol.
	β2Xtal$_a$	X-ray crystal structure	Crystal structure of human β2AR (PDB 3P0G) in complex with non-selective agonist BI-167107.
β3AR	β3AJ$_a$	*de novo* model	In complex with selective β2AR agonist AJ-9677 (T. Lybrand group)
	β3GPCRDB$_i$	Homology model	Based on the crystal structure of inactive state turkey β1AR (PDB 2VT4) (retrieved from GPCRDB website, constructed in G. Vriend group)
	β3ModWeb$_i$	Homology model	Based on the crystal structure of inactive state β2AR (PDB 3kj6) (retrieved from ModWeb server, provided by A. Sali group)

Table 4.5. The actual number of compounds used for deriving pose filters and calculating ROCE and AUC metrics.

Structure/Model	Reference ligand	[a]Binders Num.	[a]Selective binders Num.	[b]Nonbinders Num.
β1EPI$_a$	Epinephrine	42	20	32
β1GPCRDB$_i$	Cyanopindolol	42	20	32
β1ModWeb$_i$	Cyanopindolol	42	20	32
β1ModWeb$_a$	Dobutamine	42	20	32
β2TA$_a$	TA-2005	45	8	43
β2Xtal$_i$	Carazolol	45	8	43
β2Xtal$_a$	BI-167107	45	8	43
β3AJ$_a$	AJ-9677	58	12	21
β3GPCRDB$_i$	Cyanopindolol	58	12	21
β3ModWeb$_i$	Cyanopindolol	58	12	21

[a]The bound ligand for each structure model were excluded from the calculation.

[b]A number of decoys were filtered by Glide before the refinement stage. They thus did not have a Glide score and were treated to be at the end of the ranked database.

Chapter 5. Development of Predictive *in vivo* toxicity Models from Combined Knowledge of Chemical Structure Information and the ToxCastTM *in vitro* Data

5.1. Introduction

Chemical toxicity is associated with many hazardous biological effects in rodents and humans, such as gene damage, carcinogenicity, or induction of lethal diseases. It is essential to evaluate potential toxicities of all commercial chemicals before releasing them into the market, among which High Production Volume (HPV) compounds and drugs are of highest priority. However, traditional research in toxicology mostly relies on animal toxicity tests which are both labor and resource intensive, so only a very limited number of the chemicals in commerce have been evaluated [145-148]. With the recent ban on using animal for toxicity testing of cosmetics in the European Union, it becomes more urgent for industry to develop novel solutions for toxicity assessment [23,149]. In this context, the use of fast computational toxicology and high-throughput *in vitro* toxicity assays is gaining widespread interest because of their promise for supporting reduction, refinement, and/or replacement (the 3Rs) of the reliance on animal toxicity experiments[147,150].

However, accurate prediction of *in vivo* toxicity using computational tools is always challenging. Although cheminformatics approaches such as Quantitative Structure-Activity Relationship (QSAR) modeling have been widely used in computational toxicology [151,152], most existing tools can not generate models that afford reasonable prediction accuracy when applied to external compounds, resulting in inadequate model applicability for regulatory purposes [153,154]. Possible causes for such unsatisfactory performance are first, small

molecules and their metabolites can act through multiple mechanisms and hit various physiologically important targets to cause distinct adverse effects [20,21], herein, little variance in molecular structure could lead to distinct biological outcomes, making it hard to generalize rational rules from QSAR modeling. Also, the small molecules in the screening database were designed for different reasons with various scaffolds (e.g. pesticides, food additives, cosmetic ingredients). Most previous QSAR models were constructed from cogeneric compounds and had limited extrapolation power to deal with these structurally diverse compounds. On the other hand, the development of various *in vitro* toxicity testing methods, such as cell-based and cell-free HTS techniques, as well as toxicogenomic technologies, offered potential biological basis for estimating the adverse effects of chemicals[22-24,147,155]. It is intriguing to incorporate the knowledge from *in vitro* testing data to improve traditional QSAR modeling.

In 2007, the U.S. Environmental Protection Agency (U.S. EPA) initiated a chemical prioritization research program, the ToxCast[TM] project, to outline future toxicity evaluation *in vivo* [22]. The overall goal of this program is to explore a diverse array of *in vitro* toxicity assays, such as cell-based and cell-free HTS techniques, as well as toxicogenomic technologies, to estimate the adverse effects of environmental chemicals and prioritize candidates for animal testing in the future. To support ToxCast[TM]'s endeavor of predictive toxicology, U.S. EPA compiled and curated an array of high-quality historical animal toxicity data on several hundred chemicals in the Toxicity Reference Database (ToxRefDB) [156]. Phase I of ToxCast[TM] is primarily consisted of chemicals from these available animal toxicity data, and produced data from ~600 high-throughput and high content *in vitro* toxicity assays. Hence, the ToxCast[TM] program provides a valuable data set that could be used to

study the relationships between chemical structures, short term *in vitro* data and long term *in vivo* toxicity experiments.

In this study, we endeavored to improve the risk estimates for chemical toxicity through a series of novel computational approaches. To have an unbiased assessment of the derived models, these models were all developed and validated with the extensive collection of data enabled by the ToxCast[TM] project. The QSAR modeling workflow developed in-house for many years has been successfully applied to develop computational models for mutagenicity [157], carcinogenicity [25], aquatic toxicity [130] and acute toxicity[158,159]. In the first part, we employed a similar workflow to evaluate the possibility to build statistically robust, validated, and externally predictive QSAR models based on chemical descriptors alone. We employed multiple machine-learning algorithms, namely Random Forrest (RF), Support Vector Machines with linear kernel (SVM-linear), and Support Vector Machine with RBF kernel (SVM-RBF). The modeling results showed that four (two chronic and two reproductive rat toxicity endpoints) out of the eighteen evaluated *in vivo* toxicity endpoints yielded reasonable Correct Classification Rate (CCR) for external sets: consistently above 60% using all three types of modeling algorithms. In addition, another group member in our lab demonstrated that by using specific chemical scaffolds to pre-cluster the original data set, we could further improve the external predictivity of the resulting model. Thus, the results indicate that the development of externally validated toxicity predictors, while through careful study design, is feasible for at least some of the ToxRefDB endpoints.

In addition to advancing the traditional QSAR modeling approaches that depend on the explicit information about chemical structure alone, we have continued to develop novel Quantitative Structure *in vitro-in vivo* Relationship (QSIIR) approaches to

enable robust and predictive cheminformatic predictors of animal toxicity. Our recent studies have shown that it is possible to utilize the correlation between rodent acute toxicity data (*in vivo* data) and cytotoxicity data (*in vitro* data) to enhance the performance of traditional chemical-descriptors-based QSAR models [27,28]. In this project, we further explored our methodologies by applying a similar modeling workflow to three rat reproductive toxicity endpoints for the ToxCast[TM] Phase I compounds. The prediction accuracy for the best models was in the range of (70-71%) for all three *in vivo* endpoints, while that achieved by conventional QSAR models was only (57-64%) for the same external set. Furthermore, the external predictivity of individual models was employed to infer important ToxCast[TM] *in vitro* assays. The validated hierarchical models could be helpful for future toxicity testing by prioritizing high-risk compounds for animal tests, identifying informative *in vitro* assays, as well as providing hypothesis for specific signaling or biochemical pathways that are relevant to potential disease development and thus have the possibility of going beyond hazard identification. [160]

5.2. Data Set Overview

The ToxCast[TM] phase I study consisted of 320 molecules with diverse chemical scaffolds. Five duplicates and three triplicates of eight randomly selected compounds were deliberately included to quantify the reproducibility of the bioassay protocols.[161] In this study, the eleven replicated compounds were excluded. Furthermore, we excluded all molecules that cannot be appropriately handled by conventional cheminformatics techniques, e.g. inorganic and organometallic compounds or mixtures. The final ToxCast[TM] data set used in this study contained 291 unique organic compounds.

The relevant animal toxicity data were obtained from ToxRefDB, developed by the National Center for Computational Toxicology (NCCT) in partnership with U.S. EPA's Office of Pesticide Programs (OPP). The *in vivo* toxicity endpoints included in ToxRefDB were based on rat, mouse and rabbit 2-year chronic/cancer, 2-generation reproductive, and developmental toxicity study. Each of the 291 ToxCast[TM] compounds has experimental results for up to 78 different *in vivo* toxicity endpoints in ToxRefDB. Most of the *in vivo* endpoints had few numbers of active responses among all ToxCast[TM] compounds and were not suitable for QSAR modeling. For this reason, we selected 18 (Table 5.1) out of the 78 published *in vivo* endpoints that had the most experimentally active results. In addition, in the original ToxRefDB record, toxic compounds were stored with their associated Lowest Effect Level (LEL) values with units of mg/kg/day. We converted it to binary form, where compounds' activities are defined according to NCCT as either active (toxic) or inactive (non-toxic)[148]. The numbers of ToxCast[TM] phase I compounds in each toxicity endpoint subset range from 224 to 235, and the active compound ratio ranges from 17.4% to 44.6% (Table 5.1).

The ToxCast[TM] Phase I compounds have been tested against 615 various bioassays as listed in Table 5.2. These *in vitro* assays aimed to characterize a wide range of chemical biology interactions through a number of protein function assays, transcriptional reporter assays, multi-cell analysis, and developmental studies using zebra fish embryos [162]. These assays were developed by different biological companies and have been used for toxicity screening tests over years. All endpoints were presented as Inhibition Concentration by 50% (IC_{50}), Lowest Effect Level (LEL) or Lethal Dose Concentration by 50% (LD_{50}). Similar to how we handled the *in vivo* data above, we converted the experimental values into binary

form for active/inactive responses. It is noticeable that some bioassays are not informative due to the low ratio of signals on the 291 compounds. (Table 5.2) To avoid possible complications introduced by these assays during QSAR modeling process, they were excluded from the scope of work. Finally, 284 out of 615 assays were selected for which there are at least 10 active responses for the curated chemical data set.

5.3. Methods:

5.3.1. Generation of Chemical Descriptors

A set of 2,489 theoretical molecular descriptors was initially generated from the canonical SMILES notation using the software *Dragon v.5.5* (version 5.5; Talete SRL, Milano, Italy). Only 0D, 1D and 2D descriptors were considered in this study. We then removed descriptors that are constant or near constant (all, or all but one value is constant). If two descriptors are highly correlated (pair wise correlation over 95%), one of them, chosen randomly, was also deleted. The final set include 1,128 descriptors, and were range scaled to 0~1. A detailed description for descriptor generation and preparation can be found in [10].

5.3.2. In vitro – In vivo Correlation for Data Classification

The relationship between the results obtained from an *in vitro* test and a specific *in vivo* toxicity endpoint could be summarized as 4 categories. (Figure 5.1) The Class B and C compounds have consistent results *in vivo* and *in vitro*. In this study, we merge them together into a new class, which is called Group 1; and Class A and D compounds, which have conflicting results *in vivo* and *in vitro*, were combined to be Group 2. Through this way, we could generate new compound classes based on the *in vitro – in vivo* correlation between each *in vivo* endpoint results in ToxRefDB and every individual ToxCast[TM] bioassay testing result.

124

5.3.3. *QSAR Modeling Workflow and Model Validations*

A rigorous modeling workflow was carried out in order to ensure the selected models are statistically significant and externally predictive. (Figure 5.2) Key steps of the workflow are described below. Binary classification QSAR models were generated based on the chemical information only (traditional QSAR workflow) [10], or employing the chemical information and *in vitro* vs. *in vivo* correlation data (novel hierarchical workflow).

Balancing the Dataset

For each toxicity endpoint in this study, there were many more inactives than actives. This imbalanced class ratio is notorious in skewing the modeling procedure and result in biased statistics, e.g., much lower sensitivity than specificity of predictions. To address this bias, the following methodology excluded a considerable fraction of inactive compounds from the dataset to balance the active/inactive ratio. We used the active compounds from each endpoint results to create a probe subset and calculated the similarity between each inactive compound and the probe subset based on the Euclidean distance of all chemical descriptors between this inactive compound and the most similar active compound. We selected the inactive compound most similar to the active compounds into the modeling set and repeated this procedure for each active compound until the ratio of inactives in each modeling set was between 50% and 55%. This effort resulted in final datasets for all 18 toxicity endpoints for the classification QSAR modeling consisting of 45-50% actives and 50-55% inactive compounds. (Table 5.4)

Modeling Algorithms

Random forest and Support Vector Machines (SVM) algorithms were used in this study. In machine learning, a Random Forest (RF) predictor [163,164] consists of many decision

trees and calculates a prediction based upon the combined output from individual trees. For the RF modeling procedure, *n* samples are randomly drawn from the original data. These samples are then used to construct *n* training sets and to build *n* trees. For each node of the tree, *m* variables are randomly chosen from the descriptor set. The best data split for each training set was calculated using these *m* variables. In this study, only the defined parameters ($n = 500$ and $m = 13$) were used for the model development [17].

The Support vector machine (SVM), developed by Vapnik [165], serves as a general data modeling methodology where both the training set error and the model complexity are incorporated into a special loss function that is minimized during model development. SVM has since become a popular method in statistical learning because of its consistently outstanding performance in many studies and lower risk of over fitting [166,167]. In brief, an SVM model finds a separating hyperplane with a maximal margin in the feature space by minimizing a special-loss function that incorporates both the training set error and the model complexity. To cope better with different classification tasks, e.g. linear vs. nonlinear correlations, a handful of kernel functions were developed to map the original descriptor space to a higher dimensional feature space for modeling purpose.

In this study, we used the SVM implementation in the R package kernlab to build models with both linear kernels and Radial Basis Function (RBF) kernels. The RBF kernel was chosen due to its superior performance in a previous research [18]. In developing our models, a hard margin of cost=10 was used, and the RBF kernel parameter γ was automatically estimated with the sigest function in the kernlab package [125].

Model Applicability Domain

A properly defined applicability domain (AD) of a model was considered critical to determine if a given query compound can be predicted by this model within a reasonable error. For each model used in this study, an AD was determined by a threshold distance D_T between a query compound and its nearest neighbors in the training set, calculated as follows: $D_T = \bar{y} + Z\sigma$ Where $\bar{y}$ is the average Euclidean distance between each compound in the training set and its k nearest neighbors (here, k=1), σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. Here, we set the default value of Z as 0.5 to obtain reasonable prediction coverage. Thus, if the distance of a query compound from any of its k nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable and will not be included.

Model Validation

As emphasized in our earlier reports[14], modeling of only training sets is insufficient to achieve predictive power. For this reason, model validation in this study was carried out in three levels: 1. 5-fold internal cross validation prediction accuracy for the test set CCR_{test}; 2. 5-fold external cross validation prediction accuracy for the external validation set CCR_{ext}; 3. Y-randomization test that consists of rebuilding models using randomly shuffled activity labels, and subsequent evaluation of their predictive power on the external validation set. Finally, in the novel hierarchical modeling workflow (Figure 5.2), individual models each associated with a specific *in vitro* vs. *in vivo* correlation will be employed for consensus prediction of the external set. The new compounds will first be predicted as Group 1 or 2 for their *in vitro* – *in vivo* correlations. The *in vitro* testing results for the same compounds will then be used to convert the predicted correlation groups to the predicted *in vivo* toxicity.

Finally, the consensus prediction value of any compound from the external set was calculated as follows: $y_{pred} = (\sum_{j=1}^{s} y_{pred}^{j} I_{AD}^{j})/(\sum_{j=1}^{s} I_{AD}^{j})$; $I_{AD}^{j} = \begin{cases} 1 & \text{in AD} \\ 0 & \text{out of AD} \end{cases}$ For a query compound, we calculated the mean of the prediction values (y_{pred}^{j}) for this compound from all of the selected models (j=1~s). Each model has a model specific applicability domain (I_{AD}^{j}) as defined above. The compound will be labeled active or inactive based on the average prediction value.

To develop models from imbalanced data, the overall classification accuracy is less objective to evaluate the performance of models. To obtain an unbiased metric for classification ability, sensitivity and specificity are used to separately monitor the classification accuracy on two classes, and the CCR is calculated by averaging the prediction accuracies.

$$Sensitivity = \frac{TP}{TP + FN}; \ Specificity = \frac{TN}{TN + FP}; \ CCR = \frac{(Sensitivity + Specificity)}{2}$$

where TP and FP represent the numbers of predicted true and false positive compounds, and TN and FN represent the number of predicted true and false negative compounds, respectively. A permutation test (10,000 permutations) was developed to identify hierarchical models with significant better CCR_{ext} in comparison with the corresponding traditional QSAR models utilizing chemical descriptors only. If the differences between the real pair of CCR_{ext} fell within the upper 95th percentile ($p < 0.05$) of the permuted data, then the corresponding hierarchical model was considered statistically superior than traditional QSAR models. The permuted CCR_{ext} percentile values for all hierarchical models were then employed to rank the importance of the corresponding *in vitro* assays for a specific *in vivo* toxicity effect.

5.4. Results

5.4.1. Traditional QSAR Modeling with chemical descritors only

For each individual *in vivo* endpoint, we first evaluated the performance of predictors developed by conventional QSAR modeling approaches. All 18 *in vivo* toxicity subsets were compiled based on the 291 ToxCast™ compounds and their toxicity data in ToxRefDB. The QSAR toxicity models were developed for each *in vivo* endpoint subset using the Dragon descriptors of the relevant compounds. We employed RF, SVM-linear, and SVM-RBF algorithms with default parameters and five-fold internal/external cross validation for model training and testing. Therefore, we developed fifteen individual models for each toxicity endpoint, five models for each modeling algorithms. External cross-validation predictions were characterized by Correct Classification Rate (CCR$_{ext}$), sensitivity, and specificity, and are summarized in Figure 5.3A, 5.3B and 5.3C, respectively. The bean plots in Figure 5.3 simultaneously show the distribution of the 15 cross-validation predictions, and the mean of the performance evaluation metrics for each *in vivo* endpoint, grouped by the choice of applicability domain. Models for two (ChrRatCholinesteraseInhib and MgrRatLiver) of these eighteen toxicity endpoints demonstrated CCR above 60% for all three types of QSAR models. After implementing the applicability domain for each individual model, four toxicity endpoints, including two chronic (ChrRatCholinesteraseInhib and ChrMouseTumorigen) and two reproductive rat toxicity endpoints (MgrRatLiver and MgrRatKidney), were shown to have CCR over 0.60 (Figure 5.3, Table 5.3). However, in general, there is no consistent enhancement in models' external predictivity after the application of model AD. In fact, the CCRs of four toxicity endpoints, namely ChrRatLiverhypertrophy, DevRabbitPregEmbryoFetalLoss, DevRabbitPregMaternalPregLoss, and MgrRatLiver were

reduced after AD application. Therefore, we showed that we can achieve moderate improvement for some but not all toxicity models by using AD.

In addition, the models typically had poor prediction performance on the minority class (less than 0.4 in sensitivity for 14 toxicity endpoints), which is common when dealing with imbalanced data sets. Although we only selected the endpoints which were relatively rich in toxic signal, the curated data still had considerably more non-toxic than toxic chemicals for each toxicity endpoint subset. To address this bias, we balanced each modeling set by excluding non-toxic chemicals that are dissimilar to the toxic set. In summary, up to 70 % of non-toxic compounds were excluded to achieve a relatively balanced data sets (with class ratio in the range of 1~1.2). Classification QSAR models were then re-trained only on compounds remaining in the balanced modeling set, and their predictive power was estimated using the unmodified external sets. With this procedure, we made it more challenging to attain predictive QSAR model as it has to discriminate toxic chemicals from most chemically similar non-toxic chemicals.

As expected, balanced datasets afforded improved prediction accuracy for the toxic class (Figure 5.4, Table 5.4). The sensitivity increased by 0.15 in most cases. Models for six out of eighteen endpoints showed reasonable accuracy for the toxic class with sensitivities higher than 0.6. However, the overall external prediction accuracy in CCR showed only moderate improvements due to a decrease in specificity after such data modification. The best external predictivity was achieved for ChrRatCholinesteraseInhib endpoint using SVM_Rbf and SVM_Linear models with CCR 0.88; ChrRatLiverhypertrophy endpoints using all three models with CCR 0.61; MgrRatKidney endpoint using RF models with CCR 0.67; and MgrRatLiver endpoint using RF and SVM_rbf models with CCR 0.62. It should be

noted that differ from the rest endpoints, the ChrRatCholinesteraseInhib endpoint was derived from single-target focused assays, which measure compounds' inhibition rates on Cholinesterase. So it represents an "easy" case for QSAR modeling, and was used as a positive control in this study. These results suggest that employing traditional QSAR modeling with chemical descriptors alone; it is possible to develop externally validated toxicity predictors for several of the ToxRefDB toxicity endpoints.

5.4.2. *Evaluation of the Activity Landscapes of Toxicity Data Sets*

The failure of conventional QSAR models to achieve statistically significant models for most *in vivo* toxicity endpoints may be due to lack of sufficient chemical space coverage and inability of chemical descriptors to account for the toxicity mechanisms involved with the relevant endpoints. To visualize the level of activity landscape discontinuity in the dataset, we plotted pair wise similarities between compounds of the same (i.e. both toxic, or both non-toxic) and different (i.e. one toxic and another non-toxic) *in vivo* toxicity labels. Figure 5.5 illustrated such visualization for the rat toxicity endpoint measured on kidney microscopic and gross pathologies (MgrRatKidney). We found a large proportion of pairs with divergent toxicity effects which are chemically and/or biologically similar when measured with Dragon chemical descriptors, and ToxCast[TM] assays, respectively. The comparison of Metolachlor and its nearest neighbor in chemical space, Acetochlor, may serve as a good example for such discontinuity (so called "activity cliff"[168]). Since Acetochlor and Metolachlor are chemically similar (with the Tanimoto similarity coefficient of 0.82 calculated using MACCS keys), they are expected to have similar biological activities. However, their toxicity profiles in three rodent reproductive toxicity endpoints are remarkably different; they only have similar effects for MGR_Rar_Viability endpoint (Figure

5.5B). Although there are examples where small changes to key molecular features give rise to significant changes in activity, existence of a large number of such conflicting pairs in the database lowers its generalization power during modeling and, as a result, make it difficult to achieve externally predictive models. On the other hand, Acetochlor and Metolachlor can be viewed as quite dissimilar when, instead of structural features, they are compared by their results in *in vitro* assays (Figure 5.5B). Therefore, we tried to rebuild QSAR modeling using the *in vitro* assay results as the biological descriptors or using the hybrid descriptors by combining chemical descriptors and biological descriptors. Depite previous successful application of these protocols[26,159], we could not achieve significant improvents using either one for the ToxCast data. Possible causes could be that the available *in vitro* assays have low informative signals for the concerned *in vivo* endpoints, especially when compared with the influence from the chemical descriptors. These observations suggest that it is, indeed, challenging to develop robust QSAR models with either chemical descriptors or biological descriptors alone. However, as shown in the Figure 5.5A, most chemical-similar pairs of compounds with conflicting toxicity lables can be differentiated in the biological space, and vice versa. It is reasonable to expect that the additional knowledge embedded in the *in vitro* biological profiles could be useful to differentiate pairs of chemically similar compounds that have different toxicity profiles.

5.4.3. *Hierarchical QSAR Modeling Using in vivo versus in vitro Correlations*

It is well known that *in vitro* testing results, especially those obtained from HTS bioassays, have unsatisfactory correlations with systemic toxicity endpoints when any relatively large set of compounds is considered. The ToxCast[TM] bioassay data is no exception. There is no direct correlation between most ToxCast[TM] bioassay and any ToxRefDB *in vivo*

endpoint results (e.g. 1.5% of assay-*in vivo* toxicity endpoints pairs showed p-value of less than 0.01, calculated with fisher's exact test). As expected, no significant improvements in the model accuracies were observed whether using the *in vitro* assay alone as descriptors, or using the hybrid descriptors that directly combine the chemical and *in vitro* biological descriptors. How to extract relevant information from the ToxCast[TM] bioassay data and to apply it for QSAR modeling became a critical question. In this study, we extended the concept of "*in vitro vs. in vivo* correlation" that has been successfully employed to utilize information from cytotoxicity bioassays to predict rodent acute toxicity [158]. Similar approaches were reviewed elsewhere [169]. Using this criterion, for each of the reproductive toxicity data set, we can first select one *in vitro* assay as the basis and then partition the ToxCast[TM] compounds into two groups: Group 1, in which compound's *in vivo* toxicity agrees with the *in vitro* results of the ToxCast[TM] bioassay; and Group 2, if otherwise (Figure 5.2). Picking a different ToxCast assay as a basis, partitioning can be redefined. As a result, multiple assay-specific QSAR models were developed to classify compounds into assay-specific partitions, instead of directly predicting compounds' *in vivo* toxicities. The derived QSAR models were then used to assign compounds in an external dataset to one of the *in vitro* vs. *in vivo* correlation groups (i.e. group 1 or group 2). The group membership was converted into the associated *in vivo* toxicity based on the known *in vitro* response of the assay used as a basis.

To demonstrate the effectiveness of this hierarchical workflow, we applied it to three rodent reproductive endpoints: 1) Multigeneration Reproductive Rat Kidney toxicity (MgrRatKidney); 2) Multigeneration Reproductive Rat Liver toxicity (MgrRatLiver); and 3) Multigeneration Reproductive Rat Endpoint for Viability Index (MgrRatViabilityPND4).

After processed the original data with the above described procedure, the size ratio of the created partitions was again balanced with the same down-sampling approach, until it reached a balanced ratio in the range of 1~1.2. We applied the hierarchical modeling workflow for each of the 284 ToxCast[TM] *in vitro* assays and, therefore, 284 individual QSAR models of one type (RF, SVM_linear or SVM_rbf) were generated for each reproductive toxicity endpoint. Next, the consensus prediction results were derived by averaging all predictions from the 284 models.

Compared to their conventional QSAR model counterparts, only marginal improvement seems to have been achieved by using the consensus hierarchical models. Considering the fact that there are many irrelevant bioassays included for consensus prediction, it is possible that the models based on such bioassays would only contribute noise and thus lower the accuracy of the final consensus estimation. Indeed, we observed similar prediction accuracies by using hierarchical models only from a subset of the bioassays (Table 5.5). The subset of assays was selected based on their active ratio for the compounds in the modeling set. For each modeling set, a different list of assays with the highest active ratio was collected, and the predictions from their associated hierarchical models were employed for consensus prediction of the independent external set. While using a very small number of assays was shown to weaken the predictive power by consensus prediction (data not shown), we observed comparable model performance by including a number of 80 top-ranked bioassays. This result suggests that current consensus hierarchical models could be improved with a rational selection of the relevant bioassays, such as by incorporating the knowledge of the toxicity pathways to identify essential ToxCast[TM] assays. On the other hand, considering the fact that the bioassays are not entirely independent, generally, with multiple assays

targeting the same signaling pathways, or even the same macromolecular targets, consensus predictions can be biased towards pathways that have been tested more often than the others. To reduce such artificial noise, instead of using individual assays, it might be helpful to employ "composite" assays, for instance, merge all signals from nuclear receptor assays to one. Then build hierarchical QSAR models and calculate consens predictions based on those composite assays.

Here, we employed another solution to reduce the influence of irrelevant bioassays is to exclude a fraction of compounds whose consensus scores did not reach certain level. Ideally, the distribution of predicted values for the toxic and non-toxic compounds should yield clear separation (Figure 5.6B). Irrelevent predictions tend to blur the boundaries of the consensus values, but the prediction confidence increases as the consensus score reaches a higher level. For instance, we can arbitrarily defined a compound as "toxic" if more than 70% of individual models predicted it to be toxic and as "non-toxic" if more than 90% of individual models predicted it as non-toxic. Predictions for compounds that did not satisfy these two confidence levels were excluded as "inconclusive". Using this definition, the external predictivity of all the models for the three rodent endpoints was improved (up to 0.08 increase in average CCRext) at the cost of reduced prediction coverage (27% to 45% decrease; Table 5.5). Figures 5.7A, 5.7C and 5.7E show the relationship between external CCR and different toxic/non-toxic confidence breakpoints for MGR_Rat_Liver models. It should be noted that all three types of models for this specific endpoint have the highest external predictivity when the confidence breakpoints for toxic compounds are between 60-90% and the confidence breakpoints for non-toxic compounds are between 80-95%. The performance of RF and SVM-RBF models were relatively insensitive to different breakpoint

values, whereas the external accuracy of the SVM-linear model strongly depended on the breakpoint values being used. In addition to external cross-validation, we also tried to assess model robustness by Y randomization test, where QSAR models were rebuilt using modeling sets with shuffled toxicity data. The prediction accuracy of shuffled models was close to random (Figures 5.7B, 5.7D and 5.7F).

5.4.4. The Relationship between Individual ToxCastTM Bioassays and the Reproductive Toxicity Endpoints

We observed significant variation in prediction performance of hierarchical models based various bioassays. By analyzing hierarchical models with high external predictivity (CCR$_{ext}$) values, we expected to identify critical bioassays that are informative of *in vivo* toxicity effects. Therefore, we performed a permutation test (10,000 times) to evaluate each model's performance in compariaon to the conventional chemical-descriptors-only QSAR model. Table 5.6 lists the top 20 best-performing ToxCast[TM] bioassays for each multi-generation reproductive toxicity endpoint. Many of these assays were targeting genes in the family of cytochrome P450 enzymes, which are involved with the xenobiotic metabolism pathways. Several conjugating–enzymes that are active in placental xenobiotic metabolism are also identified: glutathione S-transferase (GSTA2), uridine 5'-diphosphate-glucuronosyltransferase (UGT1A), and sulfotransferase (SULT2A1). During pregnancy, the mothers are exposed to a wide variety of environmental chemicals. Placental xenobiotic-metabolizing enzymes will respond to those foreign chemicals, and can either reactivate or detoxify those compounds [170,171]. Those metabolites were observed, at least partially, to cross the placenta into the fetal circulation[172]. Therefore, understanding how placenta xenobiotic metabolizing enzymes react to the environmental hazards could provide the basis for predicting and analyzing reproductive and developmental toxicity. Estrogen receptor alpha

(ATG_ERa_TRANS) and estrogen-related receptor gamma (ATG_ERRg_TRANS) were also shown to be informative for estimating reproductive toxicity, especially on early offspring survival. Environmental chemicals are believed to interfere with the endocrine system by acting through estrogen receptors, thus generating adverse reproductive effects, including decreased fertility and implantation loss [173]. In addition, our results also showed high relevance of multigenerational reproductive toxicity to assays targeting serotonin receptors (HTR4, HTR6, and HTR7) and nuclear receptors (human pregnane X receptor (NR1I2), NR1I3). Thus, based on current understanding of mechanisms of toxicity, these data suggest that our methods are capable of selecting physiologically relevant assays, and thus can also be applied to guide potential mode-of-action analysis for future toxicity testing.

5.5. Discussion

A great number of computational technologies such as QSAR are increasingly involved in all aspects of risk assessment of environmental chemicals. Many of such approaches, however, were initially developed for use in drug discovery, which differs from toxicity evaluation in a number of important ways [174]. For example, chemical diversity of environmental toxicants is usually higher than in case of drug candidates, making it a challenging task to derive statistically robust models. In addition, unlike drug candidates that are developed with specific targets in mind, and have optimized physicochemical properties for proper absorption, distribution, metabolism, and excretion, environmental chemicals do not have uniform and well understood mechanisms behind their toxicity effects. The requirements to the QSAR models for regulatory purpose are also different from that for drug design. Accurately predicting "toxic" class (i.e. high recovery rate) and not so accurate prediction of "non-toxic" class is preferred for regulatory purpose, because predicted as non-

toxic would label compounds for low-priority in experimental tests, so toxic compounds predicted as "non-toxic" falsely would not be tested. The opposite may be desired in drug design, where the emphasis is to identify drug candidates with minimus experimental effort (high hit rate). Large number of false positives, inactive compounds predicted as active, will thus not favored. In sum, it is crucial to take into account of these differences during model development.

The compounds included in Phase I data of ToxCastTM project are also quite structurally diverse, as indicated by the distribution of Tanimoto coefficient (MACCS key) calculated for each compound to its most chemical similar counterpart. Around 30% of the ToxCastTM compound pairs have Tanimoto coefficients lower than 0.7, the cutoff commonly used to define structurally similar compounds. In comparison, the 644 compounds in the aquatic toxicity dataset[130] are much more structurally similar. Only 10% of compound pairs in the aquatic toxicity data set have Tanimoto less than 0.7, and over 44% of compound pairs have Tanimoto over 0.9. To evaluate the correlation of structural similarity/diversity with the performance of our models, we calculated the average Euclidean distances between each toxic compound and its most similar toxic compound for the 18 data sets. The dragon descriptors were used here for distance calculation because they were employed during model development. Our results showed that the toxic compounds are most similar to each other in the best chemical-descriptors-based QSAR model, the ChrRatCholinesteraseinhibition model. Thus, it indicates that the structural elements responsible for *in vivo* toxicity may not be sufficiently present in the ToxCast[TM] compounds to afford their statistical identification/generalization. Additional experiments for a diverse set of chemicals should be beneficial for the development of robust predictors.

It is understandably challenging to establish, for large and diverse data sets, direct correlation between chemical structures and their observed effects in biological systems. Chemical compounds and their metabolites can elicit similar toxicity effect through different mode-of-actions and perturbing different biological pathways. Indeed, we could only derive reasonably predictive QSAR models for four out of the eighteen *in vivo* toxicity endpoints. One of the endpoints ChrRatCholineInhibitor showed exceptionally high accuracy at CCR over 0.80, primarily because this toxicity response was measured on the interaction with a specific target, unlike the rest toxicity endpoints. As described above, a marginal structural variation from Metolachlor to Acetochlor leads to very different *in vivo* toxicity profiles for these two compounds. On the other hand, test results from a wide series of *in vitro* assays in ToxCast^{TM} could provide preliminary information on how chemicals interact with various toxicity pathways, and thus are expected to be useful for predictive models. To unite these two sources of information most effectively, we propose a novel hierarchical workflow to incorporate data from *in vitro* assays and chemical structure information into the prediction of systemic toxicity effects observed in animal tests. The rationale of the hierarchical modeling described in this chapter is that by grouping compounds exhibiting the same agreement between selected *in vitro* and *in vivo* responses, we are in a better position to identify and generalize those factors affecting *in vitro* to *in vivo* extrapolation (i.e. similar patterns of metabolism or modes of action). In our studies, the models derived from hierarchical modeling workflow has best prediction accuracy in the range of (0.70~0.71) in comparison to traditional QSAR models (0.57~0.64). However, predictions from general consensus of the hierarchical models are not always optimal; we can achieve comparable modeling performance with 80 *in vitro* assays with the highest signal contesnts. Due to the

fact that many irrelevant *in vitro* assays may contribute noise to the consensus predictions, we expect to see an increase in model performance if those assays could be identified and removed from consideration. Indeed, another group member, Liying Zhang, showed that by pre-clustering the *in vitro* assays according to known biological pathways, and using only the assays that fall into several biologically relevant pathways could enhance the model performance in certain cases. To better test capability of this modeling workflow, it would be interesting to apply it to other ToxCast^{TM} data sets.
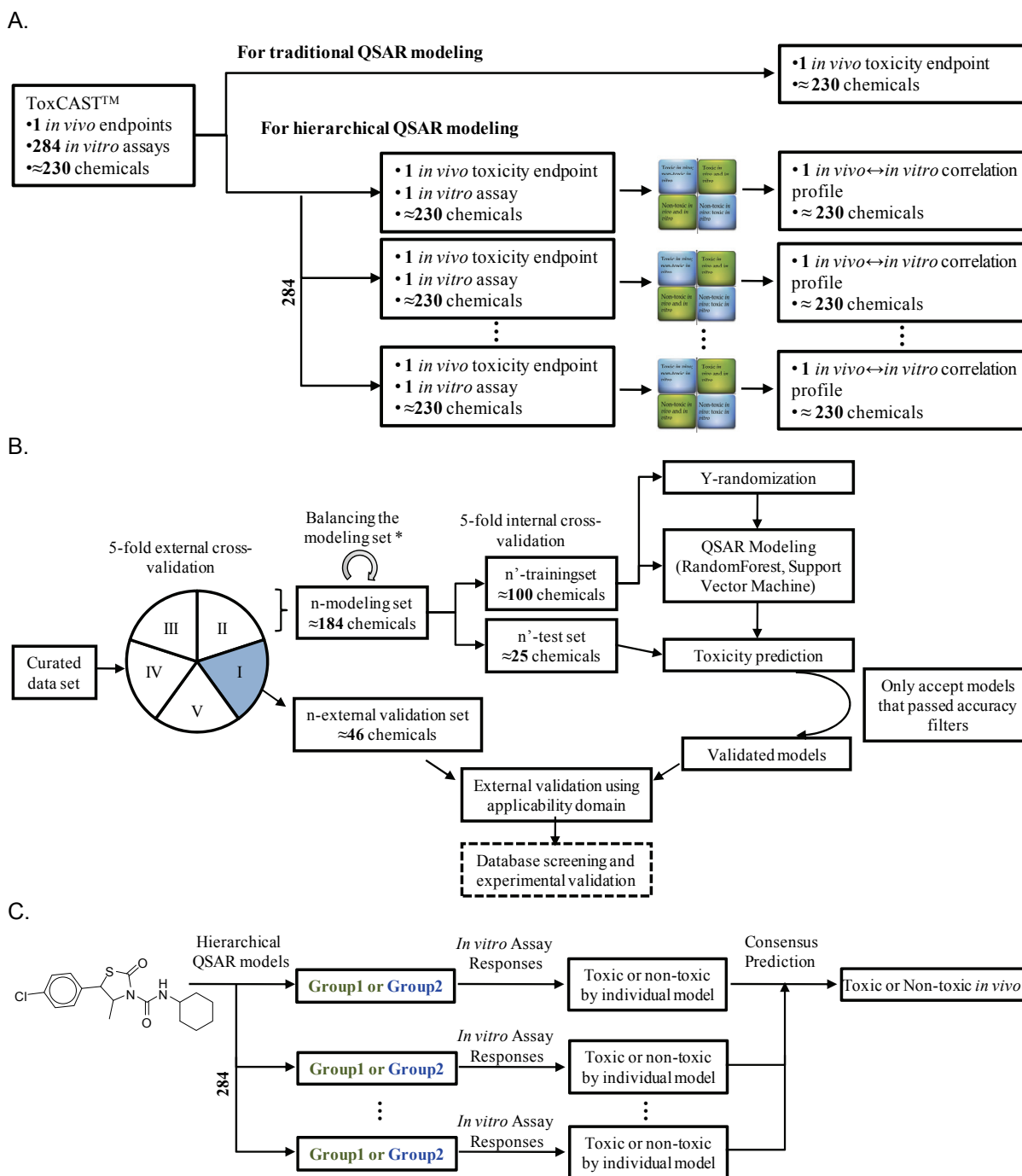
5.6. Conclusions

In conclusion, we have examined the utility of the chemical structure descriptors and the ToxCast^{TM} bioassay data for predicting *in vivo* toxicity of environmental chemicals using QSAR modeling approaches. Our results indicate that the conventional QSAR models using chemical descriptors alone and the ToxCast^{TM} bioassay results individually have limited predictive power. For some endpoints, e.g. MGRRatLiver, it is possible to use conventional QSAR models to achieve good predictivity for a subset of compounds with restricted scaffolds. Furthermore, although the ToxCast^{TM} bioassay data have low correlations with *in vivo* toxicity data, they can still be useful for improving the predictive power of QSAR models when implemented within the novel two step hierarchical QSAR modeling workflow. This hierarchical QSAR modeling workflow, although dependent on availability of *in vitro* data, can help to identify relevant *in vitro* toxicity assays for particular *in vivo* endpoints and thus could be a helpful tool for elucidating mechanisms of toxicity. Overall, the derived models could be used to guide future toxicity studies by choosing *in vitro* assays and by prioritizing compounds for *in vivo* toxicity evaluation.

Tables and Figures

Toxic *in vivo;* non-toxic *in vitro*	Toxic *in vivo* and in *vitro*
Non-toxic *in vivo* and *in vitro*	Non-toxic *in vivo;* toxic *in vitro*

Group1: *in vivo/ in vitro* **agree**
Group2: *in vivo/ in vitro* **disagree**

Figure 5.1. Data set partitioning based on the *in vivo* and *in vitro* toxicity responses. Class 1 consists of compounds whose *in vitro* and *in vivo* responses agree: toxic *in vivo* and *in vitro*, or non-toxic *in vivo* and *in vitro*; Class 2 consists of compounds whose *in vitro* and *in vivo* responses disagree: toxic *in vivo* but non-toxic *in vitro*, or non-toxic *in vivo* but toxic *in vitro*.

A.

For traditional QSAR modeling

ToxCAST™
• **1** *in vivo* endpoints
• **284** *in vitro* assays
• ≈**230** chemicals

• **1** *in vivo* toxicity endpoint
• ≈ **230** chemicals

For hierarchical QSAR modeling

284

• **1** *in vivo* toxicity endpoint
• **1** *in vitro* assay
• ≈**230** chemicals

• **1** *in vivo*↔*in vitro* correlation profile
• ≈ **230** chemicals

• **1** *in vivo* toxicity endpoint
• **1** *in vitro* assay
• ≈**230** chemicals

• **1** *in vivo*↔*in vitro* correlation profile
• ≈ **230** chemicals

• **1** *in vivo* toxicity endpoint
• **1** *in vitro* assay
• ≈**230** chemicals

• **1** *in vivo*↔*in vitro* correlation profile
• ≈ **230** chemicals

B.

Balancing the modeling set *

5-fold external cross-validation

5-fold internal cross-validation

Y-randomization

n-modeling set ≈**184** chemicals

n'-trainingset ≈**100** chemicals

QSAR Modeling (RandomForest, Support Vector Machine)

Curated data set

III II
IV I
V

n'-test set ≈**25** chemicals

Toxicity prediction

Only accept models that passed accuracy filters

n-external validation set ≈**46** chemicals

Validated models

External validation using applicability domain

Database screening and experimental validation

C.

Hierarchical QSAR models

In vitro Assay Responses

Consensus Prediction

Group1 or Group2

Toxic or non-toxic by individual model

Toxic or Non-toxic *in vivo*

284

In vitro Assay Responses

Group1 or Group2

Toxic or non-toxic by individual model

In vitro Assay Responses

Group1 or Group2

Toxic or non-toxic by individual model

* In the later section, we compared the results for after applying or not the strategies to re-balance class ratio of the modeling set

Figure 5.2. Modeling workflow. (A) Preparation of the target data set. (B) Modeling procedure for the data set.(C) external prediction procedure
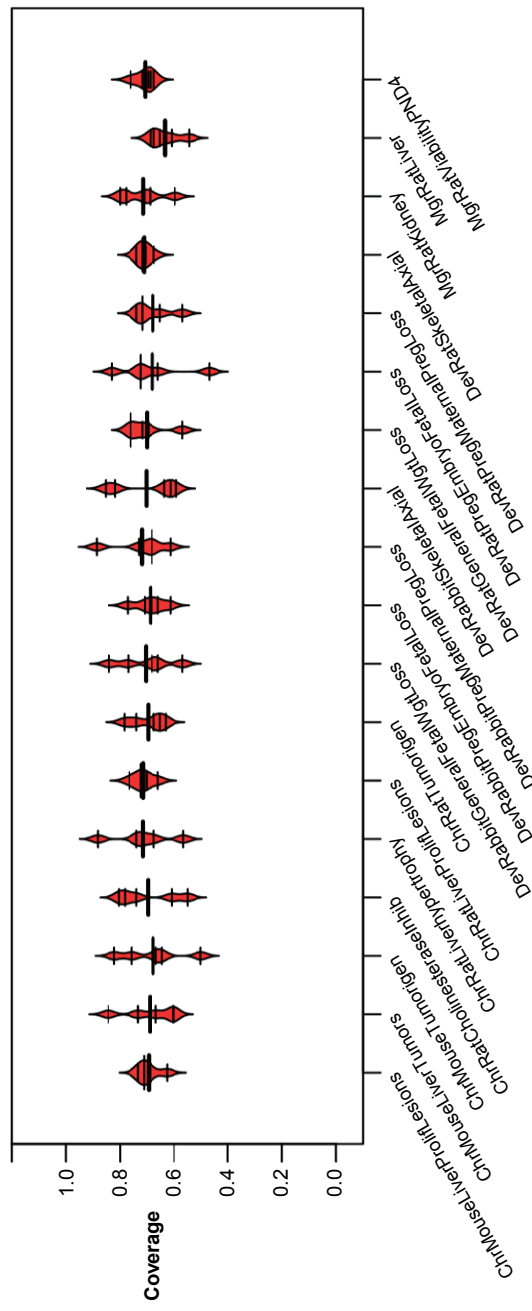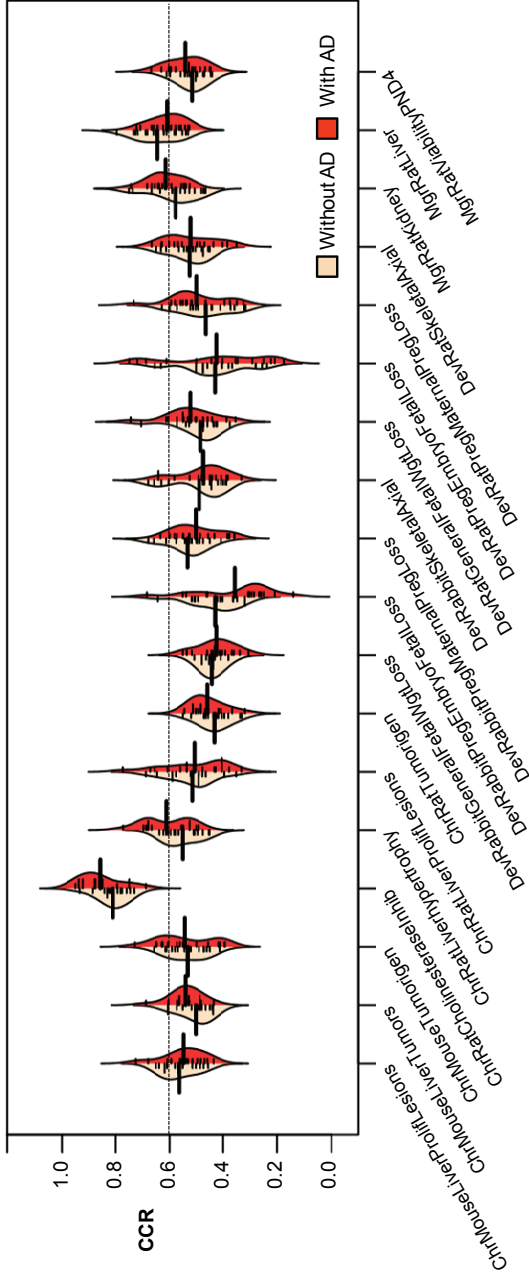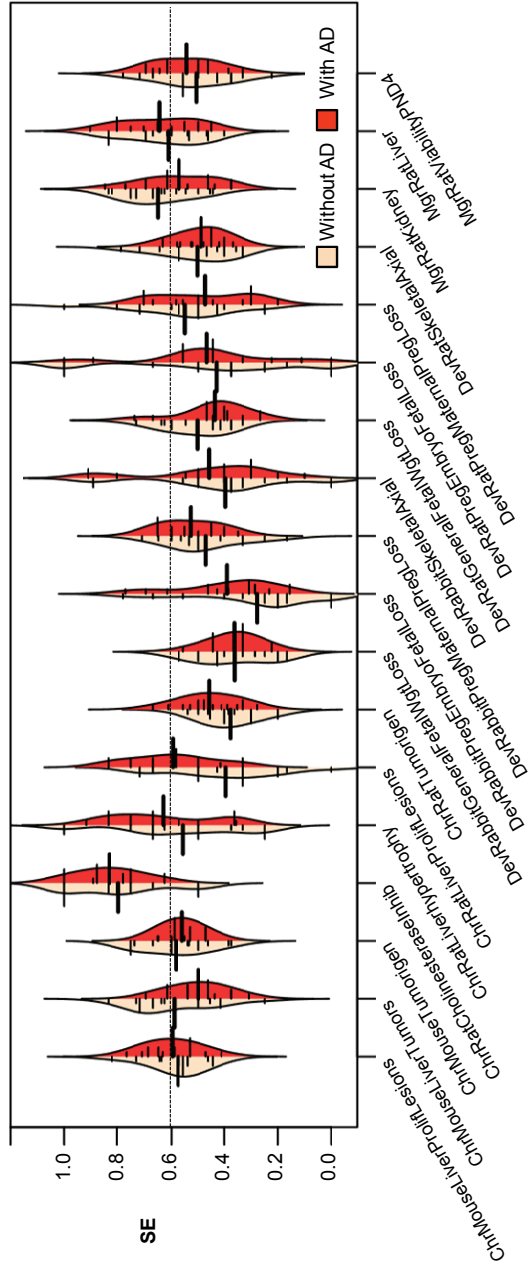
A. CCR

B. SE

143

C. SP
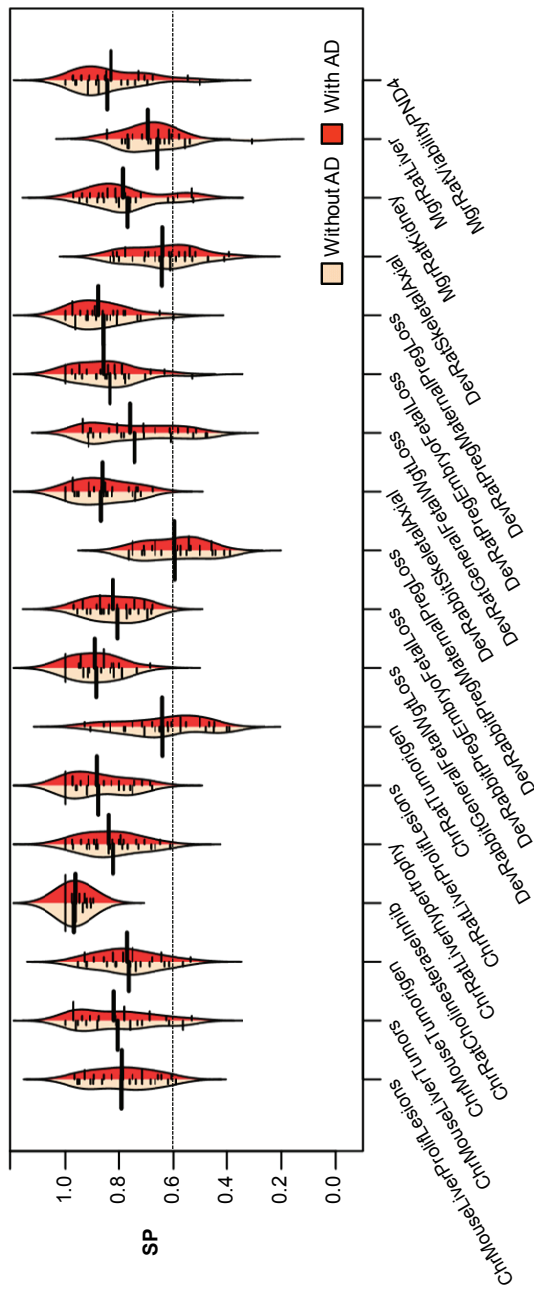
D. Coverage

Figure 5.3. Five-fold cross validation results of RF, SVM-linear, and SVM-RBF QSAR toxicity models for 18 *in vivo* endpoints: (A) CCR; (B) Sensitivity; (C) Specificity; and (D) Coverage
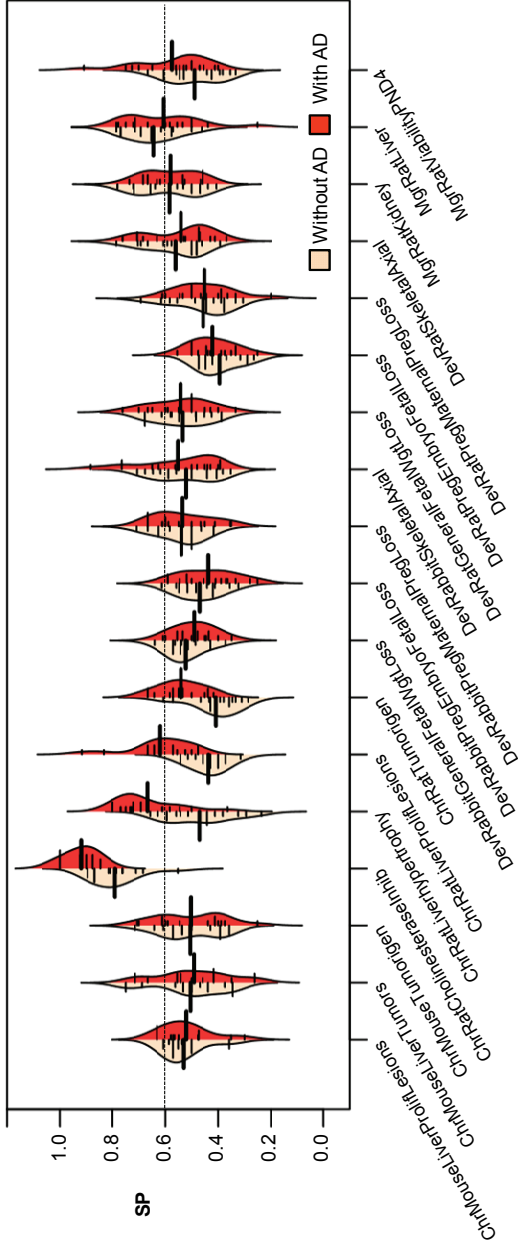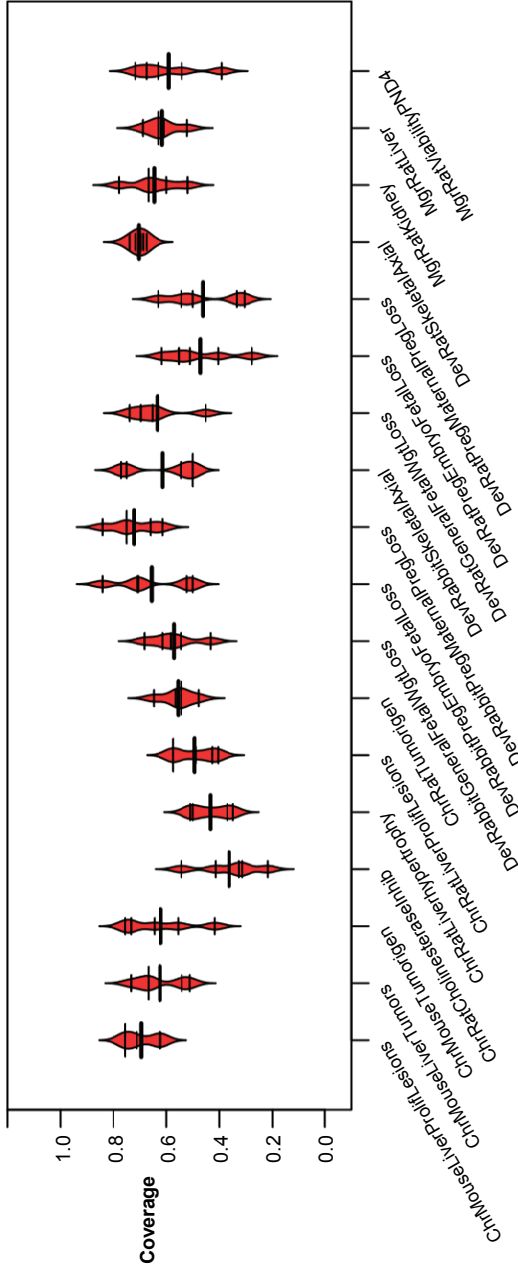
A. CCR

B. SE

C. SP

D. Coverage

Figure 5.4. Five-fold cross validation results of RF, SVM-linear, and SVM-RBF QSAR toxicity models for the balanced 18 *in vivo* endpoints: (A) CCR; (B) Sensitivity; (C) Specificity; and (D) Coverage

A.

B.

	Compounds	Metolachlor	NN1-Acetochlor
In vivo	MGR_Rat_Kidney	0*	1
	MGR_Rat_Liver	0	1
	MGR_Rar_Viability	0	0
In vitro	ACEA_IC50	0	1
	ATG_AHR_CIS	0	0
	CELLLOSS_24HR	0	1
	P53ACT_24HR	0	1
	NVS_ADME_RCYP3A2	1	0
	SOLIDUS_P450	0	1

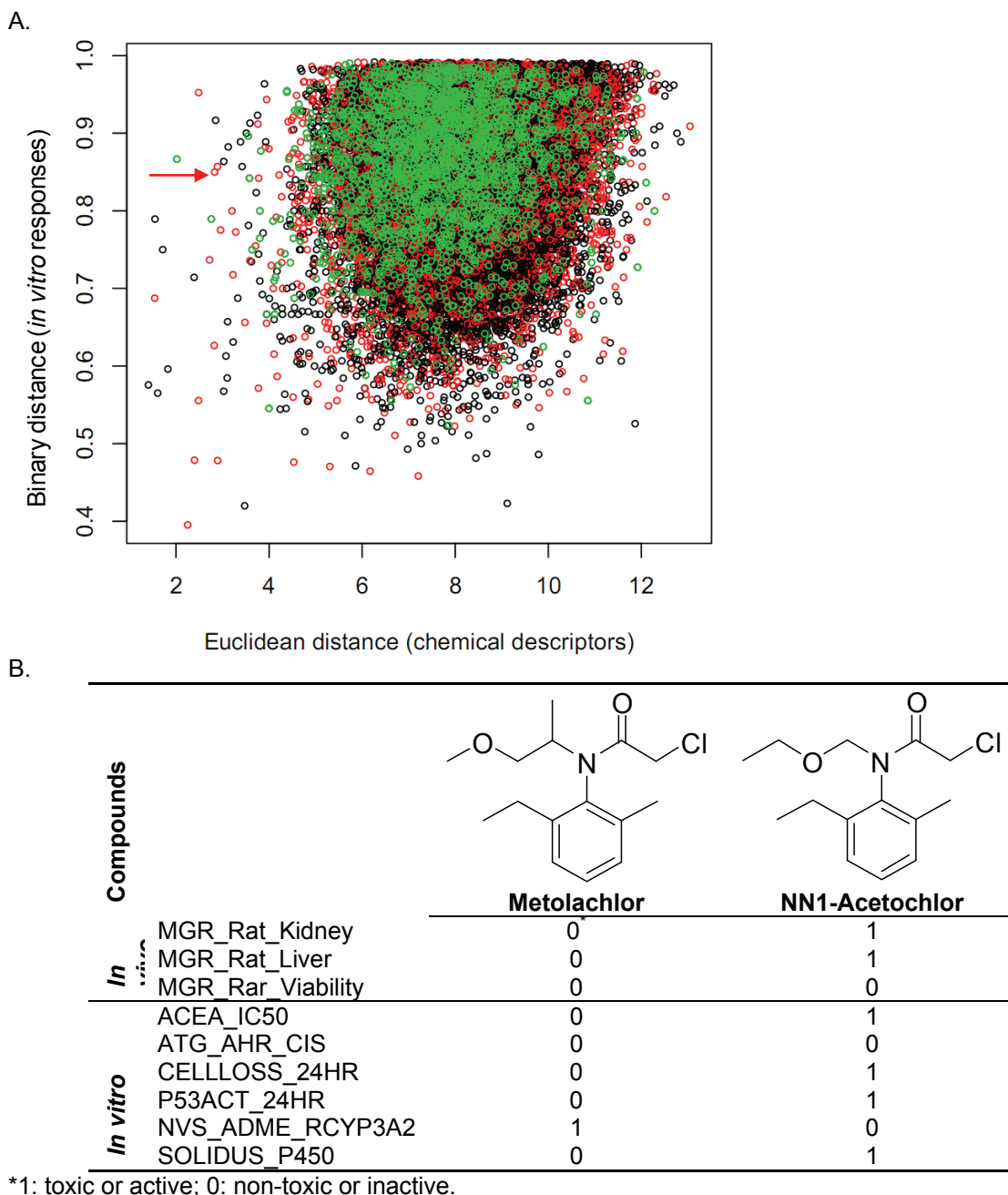*1: toxic or active; 0: non-toxic or inactive.

Figure 5.5. Illustration of activity discontinuities in the chemical and biological space. A. Pair wise distance map with data pairs color-coded by *in vivo* toxicity class measured for MGR_Rat_Kidney (black, pairs of nontoxic compounds; green, pairs of toxic compound; red, pairs of toxic and nontoxic compound.) The arrow indicates the pair of compounds: Metolachlor and Acetochlor. B. *In vivo* and *in vitro* toxicity profiles of Metolachlor, and its nearest neighbor in chemical descriptor space, NN1-Acetochlor.
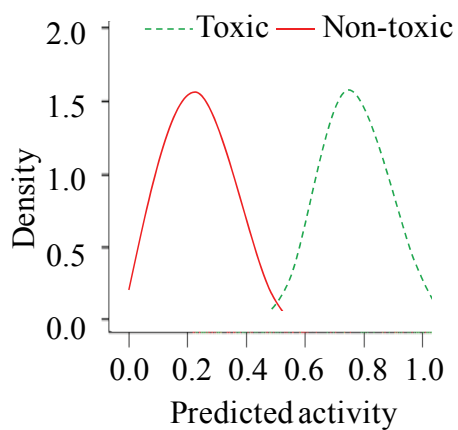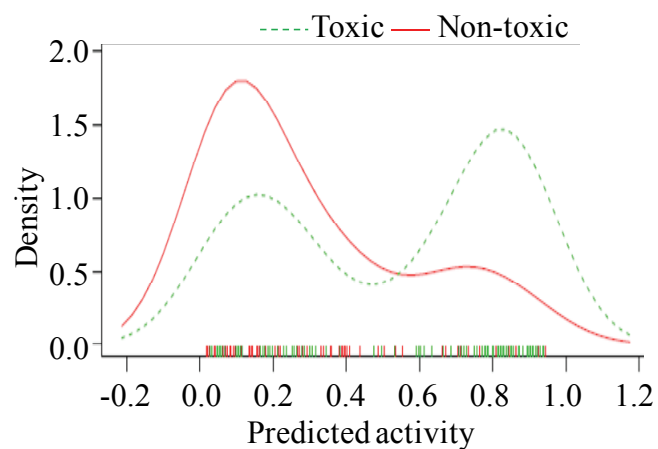
Figure 5.6. Distribution of predictions of the ideal case (A) and the hierarchical QSAR models (RandomForest approach) for MgrRatLiver endpoint (B). As can be seen better separation of two experimental categories is achieved by using two breakpoints at 0.10 and 0.70 (such classification would correspond to CCR=0.70 and Coverage~37%).

A. RF

B. RF with randomized toxicity data

C. SVM-linear

D. SVM-linear with randomized toxicity data

E. SVM-RBF

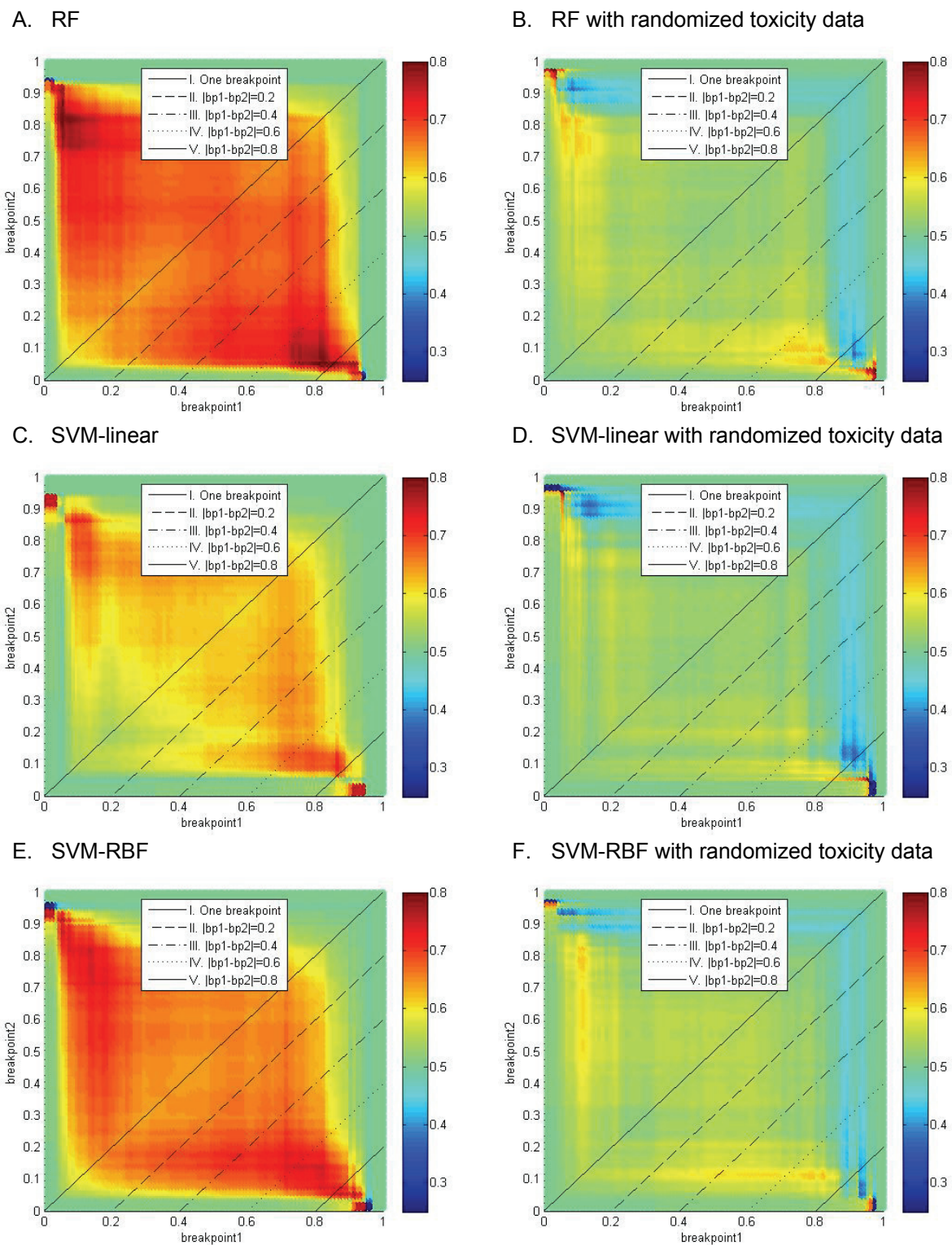### F. SVM-RBF with randomized toxicity data

Figure 5.7. Heatmap illustration of CCR_{ext} for MGR_Rat_Liver endpoint when different breakpoints values are used to define toxic or non-toxic compounds.

Table 5.1. Overview of respective chemical subsets for the 18 *in vivo* toxicity endpoints used in this study (see U.S. EPA [161] for further information of these assays)

Endpoints	Synonym	Category	Test species	Number of Compounds	Active Ratio (%)
CHR_Mouse_LiverProliferativeLesions	ChrMouseLiverProlifLesions	Chronic toxicity	Mouse	228	38.2
CHR_Mouse_LiverTumors	ChrMouseLiverTumors	Chronic toxicity	Mouse	228	29.8
CHR_Mouse_Tumorigen	ChrMouseTumorigen	Chronic toxicity	Mouse	228	38.6
CHR_Rat_Cholinesteraseinhibition	ChrRatCholinesteraseInhib	Chronic toxicity	Rat	235	17.4
CHR_Rat_LiverHypertrophy	ChrRatLiverhypertrophy	Chronic toxicity	Rat	235	26.0
CHR_Rat_LiverProliferativeLesions	ChrRatLiverProlifLesions	Chronic toxicity	Rat	235	25.5
CHR_Rat_Tumorigen	ChrRatTumorigen	Chronic toxicity	Rat	235	39.1
DEV_Rabbit_General_FetalWeightReduction	DevRabbitGeneralFetalWgtLoss	Developmental toxicity	Rabbit	224	20.0
DEV_Rabbit_PregncyRelated_EmbryoFetalLoss	DevRabbitPregEmbryoFetalLoss	Developmental toxicity	Rabbit	224	29.5
DEV_Rabbit_PregncyRelated_MaternalPregLoss	DevRabbitPregMaternalPregLoss	Developmental toxicity	Rabbit	224	44.6
DEV_Rabbit_Skeletal_Axial	DevRabbitSkeletalAxial	Developmental toxicity	Rabbit	224	22.8
DEV_Rat_General_FetalWeightReduction	DevRatGeneralFetalWgtLoss	Developmental toxicity	Rat	235	33.2
DEV_Rat_PregncyRelated_EmbryoFetalLoss	DevRatPregEmbryoFetalLoss	Developmental toxicity	Rat	235	21.7
DEV_Rat_PregncyRelated_MaternalPregLoss	DevRatPregMaternalPregLoss	Developmental toxicity	Rat	235	19.1
DEV_Rat_Skeletal_Axial	DevRatSkeletalAxial	Developmental toxicity	Rat	235	43.0
MGR_Rat_Kidney	MgrRatKidney	Reproductive toxicity	Rat	232	29.3
MGR_Rat_Liver	MgrRatLiver	Reproductive toxicity	Rat	232	43.1
MGR_Rat_ViabilityPND4	MgrRatViabilityPND4	Reproductive toxicity	Rat	232	28.0

Table 5.2. Overview of the selected 284 ToxCast bioassays. (see U.S. EPA [161,175] for further information of these assays)

Company Source	Description	Number of assays selected (provided assays)	Species	Endpoint	Maximum tested concentration (µM)	Active data ratio (%)
ACEA	Real-time Cell Electronic Sensing	6 (7)	Human	IC50	100	17.4
Attagene	Transcription factor assays	31 (73)	Human	LEL	100	22.0
BioSeek	Cell-based protein level assays	101 (174)	Human	LEL	40	22.7
Cellumen	Cell imaging assays	46 (57)	Human	IC50	200	12.7
CellzDirect	Transcription assays	31 (42)	Human	LEL	40	11.6
Gentronix	HTS genotoxicity	1 (1)	Human	LEL	200	10.3
NCGC	Nuclear receptor assays	7 (18)	Human; Rat;	IC50	200	2.0
NovaScreen	Receptor binding and enzyme inhibition assays	57 (239)	Human; Rat; Mouse; Rabbit; Pig; Guinea Pig; Sheep; Cow	IC50	20 and 50	2.8
Solidus	P450 vs. cytotoxicity assays	4 (4)	Human	LC50	960	22.1

151

Table 5.3. Prediction accuracy of 5-fold external validation for RandomForest (RF) and two types of SVM (SVM_linear, SVM_rbf) models. Here, only shows toxicity modeling results for endpoints that have models with reasonable prediction accuracy (CCR >=0.6).

| Endpoint | Model_type | Active ratio | Modeling set | External set | | | | | | | |
| | | | % accuracy | AD-[a] | | | AD+[b] | | | |
				CCR	SE	SP	CCR	SE	SP	Coverage
ChrMouseLiverProlifLesions	RF	38.2	0.52~0.57	0.56	0.22	0.90	0.57	0.21	0.93	0.69
	SVM_linear		0.50~0.59	0.57	0.47	0.67	0.55	0.43	0.67	0.69
	SVM_rbf		0.53~0.58	0.60	0.41	0.79	0.61	0.45	0.77	0.69
ChrMouseTumorigen	RF	38.6	0.57~0.64	0.58	0.28	0.89	0.61	0.35	0.87	0.68
	SVM_linear		0.47~0.58	0.55	0.43	0.67	0.55	0.42	0.68	0.68
	SVM_rbf		0.49~0.64	0.57	0.38	0.77	0.59	0.43	0.76	0.68
ChrRatCholinesteraseInhib	RF	17.4	0.73~0.80	0.75	0.51	0.98	0.76	0.54	0.98	0.69
	SVM_linear		0.80~0.86	0.83	0.71	0.95	0.85	0.75	0.96	0.69
	SVM_rbf		0.79~0.82	0.81	0.66	0.97	0.84	0.71	0.97	0.69
MgrRatKidney	RF	29.3	0.58~0.61	0.59	0.25	0.93	0.62	0.32	0.91	0.72
	SVM_linear		0.52~0.62	0.51	0.37	0.66	0.54	0.43	0.65	0.72
	SVM_rbf		0.60~0.64	0.59	0.37	0.82	0.62	0.45	0.78	0.72
MgrRatLiver	RF	43.1	0.59~0.68	0.68	0.53	0.83	0.63	0.46	0.79	0.64
	SVM_linear		0.48~0.60	0.60	0.55	0.66	0.57	0.52	0.62	0.64
	SVM_rbf		0.54~0.61	0.63	0.58	0.68	0.61	0.58	0.64	0.64

Statistics range for 18 endpoints: 0.44~0.85 0.00~0.75 0.53~0.98 0.64~0.72

a. Model performance without using applicability domain. Model coverage is 100%
b. Model performance after using applicability domain

Table 5.4. Prediction accuracy of 5-fold external validation on balanced modeling sets using RandomForest (RF) and two types of SVM (SVM_linear, SVM_rbf) models. Here, only shows toxicity modeling results for endpoints that have models with reasonable prediction accuracy (CCR >=0.6).

Endpoint	Model_type	Active ratio	Modeling set	External set						
			AD-	AD-			AD+			
			CCR_CV*	CCR	SE	SP	CCR	SE	SP	Coverage
ChrMouseLiverProlifLesions	RF	49.1	0.53~0.64	0.61	0.64	0.58	0.62	0.66	0.58	0.69
	SVM_linear		0.49~0.67	0.55	0.59	0.52	0.52	0.55	0.48	0.69
	SVM_rbf		0.50~0.67	0.57	0.62	0.51	0.57	0.61	0.53	0.69
ChrRatCholinesteraseInhib	RF	45.5	0.75~0.88	0.82	0.78	0.85	0.79	0.65	0.94	0.36
	SVM_linear		0.71~0.86	0.80	0.83	0.77	0.88	0.85	0.91	0.36
	SVM_rbf		0.72~0.86	0.82	0.88	0.76	0.88	0.85	0.91	0.36
ChrRatLiverhypertrophy	RF	47.3	0.54~0.62	0.59	0.66	0.53	0.61	0.52	0.70	0.43
	SVM_linear		0.45~0.60	0.54	0.59	0.49	0.61	0.55	0.67	0.43
	SVM_rbf		0.47~0.66	0.53	0.66	0.41	0.61	0.55	0.67	0.43
DevRatGeneralFetalWgtLoss	RF	47.3	0.45~0.59	0.52	0.49	0.55	0.59	0.60	0.58	0.63
	SVM_linear		0.40~0.51	0.46	0.41	0.52	0.49	0.46	0.52	0.63
	SVM_rbf		0.42~0.56	0.47	0.44	0.50	0.49	0.50	0.49	0.63
MgrRatKidney	RF	47.7	0.55~0.70	0.60	0.56	0.64	0.64	0.67	0.60	0.63
	SVM_linear		0.47~0.58	0.50	0.49	0.52	0.53	0.51	0.55	0.63
	SVM_rbf		0.48~0.58	0.55	0.50	0.59	0.59	0.59	0.60	0.63
MgrRatLiver	RF	48.8	0.58~0.63	0.69	0.67	0.70	0.63	0.59	0.67	0.62
	SVM_linear		0.51~0.54	0.58	0.60	0.56	0.56	0.59	0.53	0.62
	SVM_rbf		0.53~0.57	0.64	0.63	0.65	0.59	0.59	0.59	0.62
Statistics range for 18 endpoints:							0.35~0.88	0.26~0.85	0.33~0.94	0.36~0.72

Table 5.5. Prediction accuracy of external validation sets for three rat *in vivo* reproductive toxicity endpoints using conventional QSAR modeling workflow utilizing chemical descriptors only versus hierarchical QSAR modeling workflow. Bold font shows prediction accuracy after using the confidence threshold (>0.7 for toxic and <=0.1 for non-toxic). Abbreviations: Conv. QSAR, Conventional QSAR; Hier. QSAR-284, Hierarchical QSAR using the entire biological profile from the 284 *in vitro* assays for consensus prediction; Hier. QSAR-80, Hierarchical QSAR using 80 assays with the highest active ratio for consensus prediction (see text for details).

Endpoints	Model type	RF					SVM-Linear					SVM-RBF				
		CCR	SE	SP	Coverage	ROC-AUC	CCR	SE	SP	Coverage	ROC-AUC	CCR	SE	SP	Coverage	ROC-AUC
MgrRatKidney	Conv. QSAR	0.64	0.67	0.60	0.63		0.53	0.51	0.55	0.63	0.62	0.59	0.59	0.60	0.63	0.67
	Hier. QSAR-284	0.65	0.51	0.79	0.72	0.70	0.56	0.51	0.61	0.72	0.62	0.61	0.55	0.68	0.72	
		0.71	**0.75**	**0.67**	**0.35**	**0.69****	**0.62**	**0.83**	**0.41**	**0.30**		**0.69**	**0.85**	**0.52**	**0.33**	0.67*
	Hier. QSAR-80	0.63	0.45	0.81	0.72	0.70	0.58	0.53	0.63	0.72	0.63	0.58	0.51	0.65	0.72	
		0.75	**0.84**	**0.67**	**0.17**	0.68*	**0.56**	**0.82**	**0.31**	**0.10**		**0.62**	**0.81**	**0.42**	**0.15**	0.66
MgrRatLiver	Conv. QSAR	0.63	0.59	0.67	0.62		0.56	0.59	0.53	0.62	0.56	0.59	0.59	0.59	0.62	0.59
	Hier. QSAR-284	0.64	0.52	0.75	0.64	0.69**	0.58	0.58	0.58	0.64	0.60	0.63	0.60	0.67	0.64	
		0.70	**0.73**	**0.66**	**0.37**		**0.59**	**0.82**	**0.36**	**0.31**		**0.68**	**0.82**	**0.55**	**0.36**	0.67*
	Hier. QSAR-80	0.61	0.51	0.72	0.64	0.68*	0.61	0.60	0.62	0.64	0.60	0.64	0.60	0.69	0.64	
		0.65	**0.68**	**0.63**	**0.20**		**0.40**	**0.70**	**0.10**	**0.09**		**0.67**	**0.86**	**0.48**	**0.15**	0.65
MgrRatViabilityPND4	Conv. QSAR	0.57	0.55	0.58	0.56		0.47	0.40	0.55	0.56		0.49	0.43	0.56	0.56	
	Hier. QSAR-284	0.62	0.45	0.79	0.71	0.69**	0.54	0.52	0.57	0.71	0.55	0.54	0.48	0.60	0.71	
		0.70	**0.73**	**0.67**	**0.35**		**0.56**	**0.84**	**0.29**	**0.26**		**0.53**	**0.70**	**0.37**	**0.34**	0.56**
	Hier. QSAR-80	0.61	0.48	0.75	0.71	0.66*	0.55	0.50	0.61	0.71	0.58*	0.53	0.40	0.66	0.71	
		0.68	**0.71**	**0.64**	**0.16**		**0.55**	**0.80**	**0.31**	**0.10**		**0.62**	**0.80**	**0.44**	**0.12**	0.43

* p <0.05, **p<0.01, by two-tailed Student's t test on ROC-AUC between models derived from Conv.QSAR and those from Hier.QSAR.

154

Table 5.6. The top 20 ToxCast[TM] bioassays for each reproductive toxicity endpoint based on the prediction accuracy of their associated hierarchical models.

A. MGR_Rat_Kidney

Assay	Gene	Gene name/ Description
ATG_VDRE_CIS[a]	CYP27B1	cytochrome P450, family 27, subfamily B, polypeptide 1
BSK_3C_Vis_down[a]		HUVEC (IL-1b+TNFa+IFN-g): Visual
BSK_4H_Eotaxin3_down[a]	CCL26	chemokine (C-C motif) ligand 26
BSK_4H_Pselectin_down[a]	SELP	selectin P (granule membrane protein 140kDa, antigen CD62)
BSK_BE3C_hLADR_up[a]	HLA-DRA	major histocompatibility complex, class II, DR alpha
BSK_hDFCGF_EGFR_up[a]	EGFR	epidermal growth factor receptor
BSK_KF3CT_IL1a_down[a]	IL1A	interleukin 1, alpha
BSK_KF3CT_IP10_down[b]	CXCL10	chemokine (C-X-C motif) ligand 10
BSK_LPS_VCAM1_down[a]	VCAM1	vascular cell adhesion molecule 1
BSK_SAg_CD40_up[a]	CD40	CD40 molecule, TNF receptor superfamily member 5
BSK_SAg_PBMCCytotoxicity_down[a]		HUVEC/PBMC (TCR): Cytotoxicity
CLM_MicrotubuleCSK_Destabilizer_24hr[a]		HCS Microtubule Destabilization
CLM_p53Act_1hr[a]	TP53	tumor protein p53
CLM_p53Act_72hr[a]	TP53	tumor protein p53
CLZD_ABCB11_48[a]	ABCB11	ATP-binding cassette, sub-family B (MDR/TAP), member 11
CLZD_GSTA2_24[a]	GSTA2	glutathione S-transferase alpha 2
CLZD_SULT2A1_24[a]	SULT2A1	sulfotransferase family, cytosolic, 2A, dehydroepiandrosterone (DHEA)-preferring, member 1
NCGC_PXR_Agonist_human[a]	NR1I2	nuclear receptor subfamily 1, group I, member 2
NVS_GPCR_g5HT4[a]	HTR4	5 hydroxytryptamine (serotonin) receptor 4
NVS_GPCR_h5HT6[a]	HTR6	5-hydroxytryptamine (serotonin) receptor 6

B. MGR_Rat_Liver

Assay	Gene	Gene name/ Description
ATG_ERRg_TRANS[b]	ESRRG	estrogen-related receptor gamma
ATG_HIF1a_CIS	HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
ATG_Pax6_CIS	PAX6	paired box 6
BSK_4H_VCAM1_down	VCAM1	vascular cell adhesion molecule 1
BSK_hDFCGF_CollagenIII_up[a]	COL3A1	collagen, type III, alpha 1
BSK_KF3CT_IL1a_down[a]	IL1A	interleukin 1, alpha
BSK_LPS_MPC1_down	CCL2	chemokine (C-C motif) ligand 2
CLM_Hepat_CellLoss_1hr[a]		Cellumen_Hepat_CellLoss
CLM_Hepat_NuclearSize_48hr[a]	Ratus Norvegicus	Cellumen_Hepat_NuclearSize
CLM_MicrotubuleCSK_Destabilizer_24hr[a]		HCS Microtubule Destabilization
NCGC_ERalpha_Agonist[a]	ESR1	estrogen receptor 1
NVS_ADME_hCYP1A2	CYP1A2	cytochrome P450, family 1, subfamily A,

		polypeptide 2
NVS_ADME_hCYP2C8[a]	CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8
NVS_ADME_rCYP2A2	CYP2A2	cytochrome P450, subfamily 2A, polypeptide 1
NVS_ADME_rCYP2C6[a]	CYP2C6	cytochrome P450, subfamily IIC6
NVS_GPCR_g5HT4[a]	HTR4	5 hydroxytryptamine (serotonin) receptor 4
NVS_GPCR_hOpiate_mu[a]	OPRM1	opioid receptor, mu 1
NVS_NR_bPR[a]	PGR	
NVS_NR_hCAR[173]	NR1I3	nuclear receptor subfamily 1, group I, member 3
NVS_NR_hPPARg	PPARG	peroxisome proliferator-activated receptor gamma

C. MGR_Rat_Viability		
Assay	**Gene**	**Gene name/ Description**
ATG_ERa_TRANS[a]	ESR1	estrogen receptor 1
ATG_ERRg_TRANS[a]	ESRRG	estrogen-related receptor gamma
ATG_THRa1_TRANS[a]	THRA	thyroid hormone receptor, alpha (erythroblastic leukemia viral (v-erb-a) oncogene homolog, avian)
BSK_3C_uPAR_down[b]	PLAUR	
BSK_3C_Vis_down[a]		HUVEC (IL-1b+TNFa+IFN-g): Visual
BSK_BE3C_hLADR_up[b]	HLA-DRA	major histocompatibility complex, class II, DR alpha
BSK_hDFCGF_MIG_down[a]	CXCL9	chemokine (C-X-C motif) ligand 9
BSK_KF3CT_ICAM1_down	ICAM1	intercellular adhesion molecule 1
BSK_KF3CT_MMP9_down[b]	MMP9	
BSK_LPS_PGE2_down[a]	PTGER2	prostaglandin E receptor 2 (subtype EP2), 53kDa
BSK_LPS_PGE2_up[b]	PTGER2	prostaglandin E receptor 2 (subtype EP2), 53kDa
BSK_SM3C_Proliferation_down[a]		HUVEC/Primary Human Vascular Smooth Muscle Cells (IL-1b+TNF-a+IFN-g): Proliferation
CLZD_GSTA2_48	GSTA2	glutathione S-transferase alpha 2
CLZD_UGT1A1_6[a]	UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1
NVS_ADME_hCYP2C9[b]	CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9
NVS_ADME_hCYP3A5[b]	CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5
NVS_ADME_rCYP2C6[b]	Cyp2c6	cytochrome P450, subfamily IIC6
NVS_ENZ_rAChE[b]	Ache	acetylcholinesterase
NVS_GPCR_h5HT7[a]	HTR7	5-hydroxytryptamine (serotonin) receptor 7 (adenylate cyclase-coupled)
NVS_GPCR_hDRD1	DRD1	dopamine receptor D1

a.   Assays considered to be related with *in vivo* test results, based on permutation test ($p < 0.05$, 10,000 times) to show statistically better external predictivity CCR_{ext} than conventional QSAR model.
b.   Assays that showed high correlation with *in vivo* test results based on association analysis using Fisher's exact test ($p < 0.05$) , in addition to better external predictivity than conventional QSAR model in permutation test.

Chapter 6. Conclusions and Future Studies

6.1. Combi-QSAR Modeling Approach and Virtual Screening for Novel Inhibitors of Human Histone Deacetylase (HDAC)

Histone deacetylases inhibitors (HDACIs) have emerged as a new class of drugs for the treatment of human cancers and other diseases due to their effects on cell growth, differentiation, and apoptosis. In chapter 2, we have employed a combinatorial QSAR approach to generate models for 59 chemically diverse compounds tested for their inhibitory activity against HDAC1. The SVM and kNN QSAR methods were used in combination with MolConnZ and MOE descriptors independently to identify the best approach with the highest external predictive power. Rigously validated QSAR models were used to screen our in-house database collection of a total of over 9.5 million compounds. This study resulted in 45 consensus hits that were predicted to be potent HDAC1 inhibitors. Two hit compounds that were not present in the original dataset were nevertheless reported recently as HDAC1 inhibitors [131,132]. Four hit compounds with interesting chemical features were purchased and experimentally validated. Three of them were confirmed to have inhibitory activities to HDAC1 (Class I HDACs) and the best activity obtained was IC_{50} of 1.00 μM. The fourth compound was later identified to be a selective inhibitor to HDAC6, a Class II HDACs. Moreover, two of the confirmed hits are marketed drugs which could potentially expedite their development as anticancer drugs acting via HDAC1 inhibition. We believe that the technologies described in this study could be used for data analysis and hit identification for other drug discovery studies.

6.2. Development, Validation and Application of Target-specific Pose Filters in Structure-based Virtual Screening of Subtype Selective Ligands

In Chapter 3and 4, we have proposed a general pose filter using the DFT derived protein-ligand interaction descriptors (PL/MCT-tess) in structure-based virtual screening. The pose filter was validated with the 13 benchmarked DUD datasets and FRED2.2.5 docking program. The pose filter enables a constant improvement over using the default Fred scoring functions for active compounds. We further challenged the pose filter to identify subtype selective beta-adrenergic ligands. The results showed that it is possible to achieve reasonable early enrichment with carefully designed subtype biased pose filters, preferably using a selective ligand as the reference and a calibration set including both known nonselective and inactive molecules. It should be noted that the success of the proposed approach is heavily dependent on the previous knowledge, especially the selection of proper reference ligands to guide the structure modeling as well as docking pose selection.

For future studies, it would be interesting to see how additional knowledge about the receptors, such as the conformational flexibility of the binding pockets can be employed into structure-based docking, and affect the virtual screening performance. It is believed that the binding sites of GPCRs are very flexible in the absence of a bound ligand. Therefore, virtual screening with one static binding pocket from individual structure models or X-ray crystal structure may not be able to capture the real binding mode of some ligands. One possible solution is to include a set of experimentally solved structures, or in case of βAR, a set of representative structures extracted from molecular dynamic simulation study, to perform virtual screening in parallel, and combine the derived ranking lists for a consensus rank. In addition, the current pose filter does not explicitly apply any kind of model applicability domain. It is expected that small molecules with larger molecular volume could form more

non-specific contacts with receptor side-chains, and contribute to the false positive class. It should be beneficial to filter out or penalize those molecules from the ranking list. However, preliminary studies on applying a QSAR-like applicability domain for the pose filter did not show significant benefits for virtual screening performance. Further studies are needed to explore this issue.
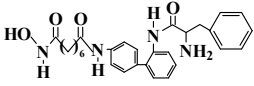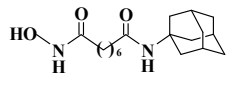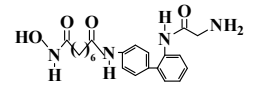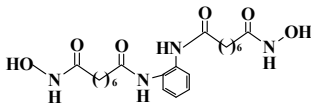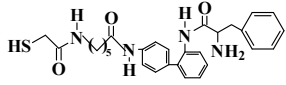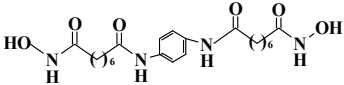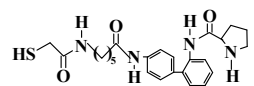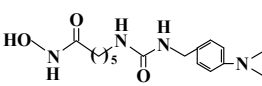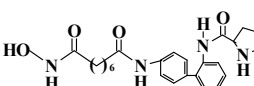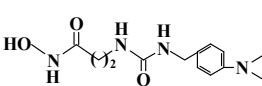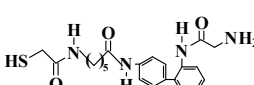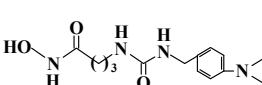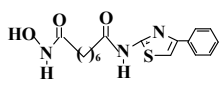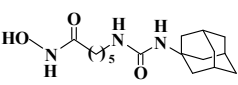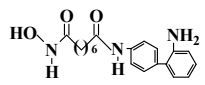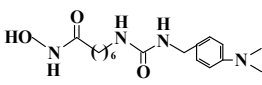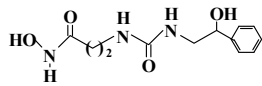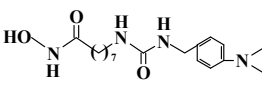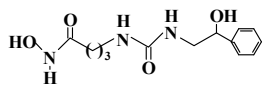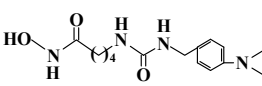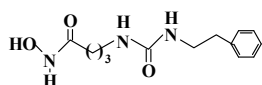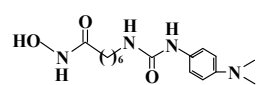
6.3. Development of Predictive *in vivo* Toxicity Models from Combined Knowledge of Chemical Structure Information and the ToxCast^TM *in vitro* Data

In Chapter 5, we have examined the utility of the chemical structure descriptors and the ToxCast^TM bioassay data for predicting *in vivo* toxicity of environmental chemicals. Our results indicate that the conventional QSAR models using chemical descriptors alone and the ToxCast^TM bioassay results individually have limited predictive power. For some endpoints, e.g. MGRRatLiver, it is possible to use conventional QSAR models to achieve good predictivity for a subset of compounds with restricted scaffolds. On the other hand, although the ToxCast^TM bioassay data have low correlations with *in vivo* toxicity data, they can still be useful for improving the predictive power of QSAR models when implemented within the novel two step hierarchical QSAR modeling workflow. This novel hierarchical QSAR modeling workflow can also help to estimate possible mechanisms of the toxicity pathways by evaluating the *in vitro* toxicity assays associated with top-ranked models. The derived models thus could be used to guide the future toxicity studies by choosing *in vitro* assays, prioritizing compounds for *in vivo* toxicity evaluation, as well as directing potential mode-of-action analysis by evaluation of the molecular targets and pathways of the most predictive *in vitro* assays.
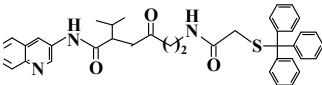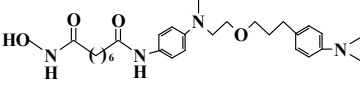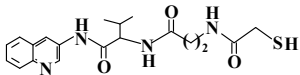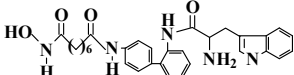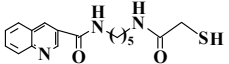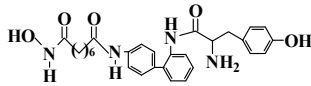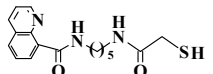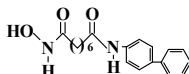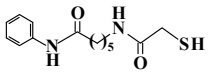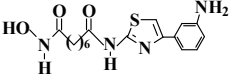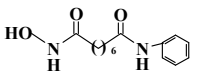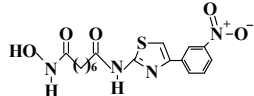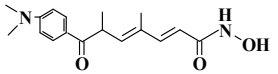
As shown from our studies, the proposed QSAR modeling workflow can improve toxicity predictions with a reduced applicability domain. The coming phase II ToxCast[TM] data that was tested against an additional 1000 compounds would provide an objective test set to evaluate models' applicability as well as prediction accuracy. In addition, it is well known that the toxicity effects are dose-dependent. For current studies, the toxicity observations as well as the bioassay responses have been simplified into active (toxic) or inactive (non-toxic) classes. It would be interesting if we can include the dose-response data in the QSAR modeling, and test whether we can achieve better performance. Indeed, a recent publication by Sedykh et al.[26] showed that it is possible to use the concentration-response data in the prediction of rodent acute toxicity. Furthermore, it would be beneficial to combine the prediction results with the results from bioinformatics analysis of the inherent connections between protein targets and bioassays.

APPENDICES

Appendix I: Structure and *pIC₅₀* of fifty-nine HDAC1inhibitors used for model building and validation.

Mol.	Comp. ID	pIC_{50}	Mol.	Comp. ID	pIC_{50}
	AG-biph-08	7.10		BC-2-83	6.00
	AG-biph-15	6.70		BC-2-84	6.00
	AG-biph-35	5.15		BC-2-87	6.00
	AG-biph-36	5.82		BC-3-10	6.32
	AG-biph-38	7.10		BC-3-14	5.00
	AG-biph-40	5.52		BC-3-18	4.30
	Ag-THIA-01	7.82		BC-3-22	6.38
	Ag-b-57	7.10		BC-3-42	6.30
	BC-1-30-2	4.00		Bc-3-46	6.70
	BC-2-45	4.00		Bc-3-5	6.72
	BC-2-48	4.00		Bc-3-52	6.87

Mol.	Comp. ID	pIC₅₀	Mol.	Comp. ID	pIC₅₀
	Bc-3-63	5.00		BC-4-56	5.96
	BC-3-70	5.00		BC-4-81	6.10
	BC-3-94	5.00		BC-4-84	6.10
	BC-4-14	6.12		BC-4-86	5.33
	BC-4-2	5.00		BC-4-87	4.89
	BC-4-3	6.20		BC-4-93	5.77
	BC-4-31	6.00		BC-4-96	6.22
	BC-4-4	5.30		Bc-5-44	6.05
	BC-4-44	6.05		Bc-6-12	6.70
	BC-4-46	6.35		Bc-6-25	7.36
	BC-4-54	6.26		Bc-6-26	7.30
	BC-4-55	6.70		BC-6-30	5.52

Mol.	Comp. ID	pIC_{50}	Mol.	Comp. ID	pIC_{50}
	BC-6-33	4.30		YC-03065	6.52
	BC-6-34	4.60		Ychdac044	7.30
	BC-6-38	6.40		Ychdac045	7.52
	BC-6-40	6.60		Yc-Ⅱ-84	7.26
	BC-6-83	6.52		Yc-Ⅱ-88	8.10
	SAHA	7.10		Yc-Ⅱ-90	7.66
	TSA	8.46			

REFERENCES

1. Schreiber, S.L. Small molecules: the missing link in the central dogma. *Nat Chem Biol* **1**, 64-6 (2005).

2. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nature Reviews Drug Discovery* **1**, 727-730 (2002).

3. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat Rev Genet* **5**, 262-75 (2004).

4. Spring, D.R. Chemical genetics to chemical genomics: small molecules offer big insights. *Chem Soc Rev* **34**, 472-82 (2005).

5. Stockwell, B.R. Exploring biology with small organic molecules. *Nature* **432**, 846-54 (2004).

6. Tan, D.S. Diversity-oriented synthesis: exploring the intersections between chemistry and biology. *Nat Chem Biol* **1**, 74-84 (2005).

7. Bajorath, J. Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery. (2008).

8. Bajorath, J. Computational analysis of ligand relationships within target families. *Current opinion in chemical biology* **12**, 352-358 (2008).

9. Bajorath, J. & SpringerLink. *Chemoinformatics and Computational Chemical Biology*, (Humana Press, 2011).

10. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **29**, 476-488 (2010).

11. Stouch, T.R. et al. In silico ADME/Tox: why models fail. *Journal of computer-aided molecular design* **17**, 83-92 (2003).

12. Tropsha, A., Gramatica, P. & Gombar, V.K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Combinatorial Science* **22**, 69-77 (2003).

13. Kubinyi, H. Validation and predictivity of QSAR models. 30?3 (2006).

14. Golbraikh, A. & Tropsha, A. Beware of q2! *Journal of Molecular Graphics and Modelling* **20**, 269-276 (2002).

15. Bender, A. Bayesian methods in virtual screening and chemical biology. *Methods in molecular biology (Clifton, NJ)* **672**, 175 (2011).

16. Zheng, W. & Tropsha, A. Novel variable selection quantitative structure--property relationship approach based on the k-nearest-neighbor principle. *J Chem Inf Comput Sci* **40**, 185-94 (2000).

17. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R news* **2**, 18-22 (2002).

18. Burbidge, R., Trotter, M., Buxton, B. & Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem* **26**, 5-14 (2001).

19. Warmuth, M.K. et al. Active learning with support vector machines in the drug discovery process. *J Chem Inf Comput Sci* **43**, 667-73 (2003).

20. Martin, Y.C., Kofron, J.L. & Traphagen, L.M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry* **45**, 4350-4358 (2002).

21. Guengerich, F.P. & MacDonald, J.S. Applying mechanisms of chemical toxicity to predict drug safety. *Chemical research in toxicology* **20**, 344-369 (2007).

22. Dix, D.J. et al. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol Sci* **95**, 5-12 (2007).

23. Krewski, D. et al. Toxicity testing in the 21st century: a vision and a strategy. *J Toxicol Environ Health B Crit Rev* **13**, 51-138 (2010).

24. Martin, M.T. et al. Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem Res Toxicol* **23**, 578-90 (2010).

25. Zhu, H., Rusyn, I., Richard, A. & Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environmental Health Perspectives* **116**, 506 (2008).

26. Sedykh, A. et al. Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ Health Perspect* **119**, 7 (2011).

27. Zhu, H., Rusyn, I., Richard, A. & Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ Health Perspect* **116**, 506-13 (2008).

28. Zhu, H. et al. A novel two-step hierarchical quantitative structure-activity relationship modeling work flow for predicting acute toxicity of chemicals in rodents. *Environ Health Perspect* **117**, 1257-64 (2009).

29. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J. & Corbeil, C.R. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British journal of pharmacology* **153**, S7-S26 (2008).

30. Greer, J., Erickson, J.W., Baldwin, J.J. & Varney, M.D. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *J Med Chem* **37**, 1035-54 (1994).

31. Gubernator, K. & Bohm, H.J. *Structure-based ligand design*, (Wiley-Vch: Weinheim, Germany, 1998).

32. Wlodawer, A. & Vondrasek, J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct* **27**, 249-84 (1998).

33. Rognan, D. Structure-Based Approaches to Target Fishing and Ligand Profiling. *Molecular Informatics* **29**, 176-187 (2010).

34. Rockey, W.M. & Elcock, A.H. Rapid computational identification of the targets of protein kinase inhibitors. *Curr Opin Drug Discov Devel* **9**, 326-31 (2006).

35. Fanelli, F. & De Benedetti, P.G. Computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem Rev* **105**, 3297-351 (2005).

36. Klabunde, T. & Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* **3**, 928-44 (2002).

37. Ellison, K.E. & Gandhi, G. Optimising the use of beta-adrenoceptor antagonists in coronary artery disease. *Drugs* **65**, 787-97 (2005).

38. Novoa, E.M., de Pouplana, L.R., Barril, X. & Orozco, M. Ensemble Docking from Homology Models. *Journal of Chemical Theory and Computation* **6**, 2547-2557 (2010).

39. Fan, H. et al. Molecular Docking Screens Using Comparative Models of Proteins. *J.Chem Inf.Model.* (2009).

40. Bissantz, C., Bernard, P., Hibert, M. & Rognan, D. Protein-based virtual screening of chemical databases. II. Are homology models of G-Protein Coupled Receptors suitable targets? *Proteins* **50**, 5-25 (2003).

41. Kairys, V., Fernandes, M.X. & Gilson, M.K. Screening drug-like compounds by docking to homology models: a systematic study. *J.Chem.Inf.Model.* **46**, 365-379 (2006).

42. McGovern, S.L. & Shoichet, B.K. Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J.Med.Chem.* **46**, 2895-2907 (2003).

43.   Bordogna, A., Pandini, A. & Bonati, L. Predicting the accuracy of protein 托 igand docking on homology models. *J Comput Chem* (2010).

44.   Ivanov, A.A., Barak, D. & Jacobson, K.A. Evaluation of homology modeling of G-protein-coupled receptors in light of the A(2A) adenosine receptor crystallographic structure. *J Med Chem* **52**, 3284-92 (2009).

45.   Cherezov, V. et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318**, 1258-1265 (2007).

46.   Rosenbaum, D.M. et al. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318**, 1266-1273 (2007).

47.   Rasmussen, S.G. et al. Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383-387 (2007).

48.   Kontoyianni, M., McClellan, L.M. & Sokol, G.S. Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* **47**, 558-65 (2004).

49.   Lyne, P.D. Structure-based virtual screening: an overview. *Drug Discov Today* **7**, 1047-55 (2002).

50.   Vangrevelinghe, E. et al. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**, 2656-62 (2003).

51.   Friesner, R.A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Vol. 47 1739-1749 (2004).

52.   Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W. & Taylor, R.D. Improved protein-ligand docking using GOLD. *Proteins* **52**, 609-23 (2003).

53.   Wu, S.Y. et al. Discovery of a novel family of CDK inhibitors with the program LIDAEUS: structural basis for ligand-induced disordering of the activation loop. *Structure* **11**, 399-410 (2003).

54.   Hindle, S.A., Rarey, M., Buning, C. & Lengaue, T. Flexible docking under pharmacophore type constraints. *J Comput Aided Mol Des* **16**, 129-49 (2002).

55.   Chuaqui, C., Deng, Z. & Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J Med Chem* **48**, 121-33 (2005).

56.   Deng, Z., Chuaqui, C. & Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem* **47**, 337-44 (2004).

57.   de Graaf, C. & Rognan, D. Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. *J Med Chem* **51**, 4978-85 (2008).

58. de Graaf, C. & Rognan, D. Customizing G Protein-coupled receptor models for structure-based virtual screening. *Curr Pharm Des* **15**, 4026-48 (2009).

59. Weill, N. & Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J Chem Inf Model* **49**, 1049-62 (2009).

60. Hsieh, J.H. et al. Cheminformatics Meets Molecular Mechanics: A Combined Application of Knowledge-based Pose Scoring and Physical Force Field-based Hit Scoring Functions Improves the Accuracy of Structure-Based Virtual Screening. *submitted* (2011).

61. Hsieh, J.H. Cheminformatics Approaches to Structure Based Virtual Screening: Methodology Development and Applications. *Doctoral dissertation* (2011).

62. Oprea, T.I., May, E.E., Leitao, A. & Tropsha, A. Computational systems chemical biology. *Methods Mol Biol* **672**, 459-88 (2011).

63. Hassig, C.A. & Schreiber, S.L. Nuclear histone acetylases and deacetylases and transcriptional regulation: HATs off to HDACs. *Curr Opin Chem Biol* **1**, 300-8 (1997).

64. Wolffe, A.P. Histone deacetylase: a regulator of transcription. *Science* **272**, 371-2 (1996).

65. Yamagoe, S. et al. Interaction of histone acetylases and deacetylases in vivo. *Mol Cell Biol* **23**, 1025-33 (2003).

66. Mork, C.N., Faller, D.V. & Spanjaard, R.A. A mechanistic approach to anticancer therapy: targeting the cell cycle with histone deacetylase inhibitors. *Curr Pharm Des* **11**, 1091-104 (2005).

67. Gui, C.Y., Ngo, L., Xu, W.S., Richon, V.M. & Marks, P.A. Histone deacetylase (HDAC) inhibitor activation of p21WAF1 involves changes in promoter-associated proteins, including HDAC1. *Proc Natl Acad Sci U S A* **101**, 1241-6 (2004).

68. Johnstone, R.W. Histone-deacetylase inhibitors: novel drugs for the treatment of cancer. *Nat Rev Drug Discov* **1**, 287-99 (2002).

69. Johnstone, R.W. & Licht, J.D. Histone deacetylase inhibitors in cancer therapy: is transcription the primary target? *Cancer Cell* **4**, 13-8 (2003).

70. Wolffe, A.P. Chromatin remodeling: why it is important in cancer. *Oncogene* **20**, 2988 (2001).

71. Bolden, J.E., Peart, M.J. & Johnstone, R.W. Anticancer activities of histone deacetylase inhibitors. *Nat Rev Drug Discov* **5**, 769-84 (2006).

72.   Rodriquez, M. et al. Chemistry and Biology of Chromatin Remodeling Agents: State of Art and Future Perspectives of HDAC Inhibitors. *Current Medicinal Chemistry* **13**, 1119-1139 (2006).

73.   Yoshida, M., Kijima, M., Akita, M. & Beppu, T. Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. *Journal of Biological Chemistry* **265**, 17174 (1990).

74.   Curtin, M. & Glaser, K. Histone deacetylase inhibitors: the Abbott experience. *Curr Med Chem* **10**, 2373-92 (2003).

75.   Juvale, D.C. et al. 3D-QSAR of histone deacetylase inhibitors: hydroxamate analogues. *Org Biomol Chem* **4**, 2858-68 (2006).

76.   Wang, D.F., Helquist, P., Wiech, N.L. & Wiest, O. Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases. *J Med Chem* **48**, 6936-47 (2005).

77.   Woo, S.H. et al. Structurally simple trichostatin A-like straight chain hydroxamates as potent histone deacetylase inhibitors. *J Med Chem* **45**, 2877-85 (2002).

78.   Richon, V.M. et al. Second generation hybrid polar compounds are potent inducers of transformed cell differentiation. *Proc Natl Acad Sci U S A* **93**, 5705-8 (1996).

79.   Grant, S., Easley, C. & Kirkpatrick, P. Vorinostat. *Nat Rev Drug Discov* **6**, 21-2 (2007).

80.   Landry, J. et al. The silencing protein SIR2 and its homologs are NAD-dependent protein deacetylases. *Proc Natl Acad Sci U S A* **97**, 5807-11 (2000).

81.   Gregoretti, I.V., Lee, Y.M. & Goodson, H.V. Molecular evolution of the histone deacetylase family: functional implications of phylogenetic analysis. *J Mol Biol* **338**, 17-31 (2004).

82.   Finnin, M.S. et al. Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **401**, 188-93 (1999).

83.   Somoza, J.R. et al. Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure* **12**, 1325-34 (2004).

84.   Vannini, A. et al. Crystal structure of a eukaryotic zinc-dependent histone deacetylase, human HDAC8, complexed with a hydroxamic acid inhibitor. *Proc Natl Acad Sci U S A* **101**, 15064-9 (2004).

85.   Ragno, R., Simeoni, S., Valente, S., Massa, S. & Mai, A. 3-D QSAR studies on histone deacetylase inhibitors. A GOLPE/GRID approach on different series of compounds. *J Chem Inf Model* **46**, 1420-30 (2006).

86. Tropsha, A. & Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr Pharm Des* **13**, 3494-504 (2007).

87. Chen, B. et al. Chemistry and biology of mercaptoacetamides as novel histone deacetylase inhibitors. *Bioorg Med Chem Lett* **15**, 1389-92 (2005).

88. Kozikowski, A.P. et al. Functional differences in epigenetic modulators-superiority of mercaptoacetamide-based histone deacetylase inhibitors relative to hydroxamates in cortical neuron neuroprotection studies. *J Med Chem* **50**, 3054-61 (2007).

89. Kozikowski, A.P., Dritschilo, A., Jung, M., Petukhov, P.A. & Chen, B. Histone deacetylase inhibitors for treatment of neurological diseases and cancer. in *PCT Int. Appl.* Vol. 2005007091 (2005).

90. Kozikowski, A.P., Dritschilo, A., Jung, M., Petukhov, P.A. & Chen, B. Preparation of w-ureido alkanohydroxamic acid and related urea drivatives as histone deacetylase inhibitors. in *U.S. Pat. Appl. Publ.* Vol. 2005014839 (2005).

91. Kozikowski, A.P. & Chen, B. Preparation of hydroxyamides and mercaptoacetamides as histone deacetylase inhibitors for treatment of neurological dieseases and cancer. in *U.S. Pat. Appl. Publ.* Vol. 2005032831 (2005).

92. Kozikowski, A.P., Jung, M. & Dritschilo, A. Isoform-selective HDAC inhibitors including biphenyl hydroxamic acid- and mercaptoacetamide-containing amino acid amides, their preparation and use for treating cancer, neurological diseases and malaria. in *PCT Int. Appl.* Vol. 2008019025 (2008).

93. Ohtani, M., Matsuura, T., Shirahase, K. & Sugita, K. (2E)-5-[3-[(Phenylsulfonyl)amino]phenyl]- pent-2-en-4-ynohydroxamic Acid and Its Derivatives as Novel and Potent Inhibitors of ras Transformation. *Journal of Medicinal Chemistry* **39**, 2871-2873 (1996).

94. Jung, M. et al. Amide Analogues of Trichostatin A as Inhibitors of Histone Deacetylase and Inducers of Terminal Cell Differentiation. *Journal of Medicinal Chemistry* **42**, 4669-4679 (1999).

95. Lu, Q. et al. Zn2+-chelating motif-tethered short-chain fatty acids as a novel class of histone deacetylase inhibitors. *J Med Chem* **47**, 467-74 (2004).

96. Suzuki, T. et al. Synthesis and histone deacetylase inhibitory activity of new benzamide derivatives. *J Med Chem* **42**, 3001-3 (1999).

97. Jimenez, C. & Crews, P. Novel marine sponge derived amino acids 13. Additional psammaplin derivatives from Psammaplysilla purpurea. *Tetrahedron* **47**, 2097-2102 (1991).

98. Irwin, J.J. & Shoichet, B.K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**, 177-82 (2005).

99.   World Drug Index (WDI).

100.   Synergy Libraries.

101.   InterBioScreen Libraries.

102.   Progenitor Databases.

103.   MolconnZ. version 4.05 edn (eduSoft, LC: Ashland, VA, 2006).

104.   Wiener, H. Structural determination of paraffin boiling points. *Journal of the American Chemical Society* **69**, 17-20 (1947).

105.   Randic, M. Characterization of molecular branching. *Journal of the American Chemical Society* **97**, 6609-6615 (1975).

106.   Kier, L.B. & Hall, L.H. An electrotopological-state index for atoms in molecules. *Pharm Res* **7**, 801-7 (1990).

107.   Kier, L.B., Hall, L.H. & Frazer, J.W. An index of electrotopological state for atoms in molecules. *Journal of Mathematical Chemistry* **7**, 229-241 (1991).

108.   Kier, L.B. & Hall, L.H. *Molecular Connectivity in Structure-Activity analysis*, (Wiley: New York, 1986).

109.   Kier, L.B. & Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research*, (Academic Press: New York, 1976).

110.   Petitjean, M. 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**, 331-337 (1992).

111.   Kier, L.B. & Hall, L.H. *Molecular Structure Description: The Electrotoplogical State*, (Academic Press: New York, 1999).

112.   Molecular Operation Environment version 2006.08 edn (Chemical Computing Group Inc. :Montreal, Quebec, Canada, Montreal, Quebec, Canada, 2006).

113.   Wiener, H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *Journal of the American Chemical Society* **69**, 2636-2638 (1947).

114.   Balaban, A.T. Highly discriminating distance-based topological index. *Chemical Physics Letters* **89**, 399-404 (1982).

115.   Balaban, A.T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **53**, 355-375 (1979).

116.	Gasteiger, J. & Marsili, M. Iterative partial equalization of orbital electronegativity--a rapid access to atomic charges. *Tetrahedron* **36**, 3219-3228 (1980).

117.	Stanton, D.T. & Jurs, P.C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical chemistry* **62**, 2323-2329 (1990).

118.	Golbraikh, A. & Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers* **5**, 231-43 (2002).

119.	Golbraikh, A. et al. Rational selection of training and test sets for the development of validated QSAR models. *J Comput Aided Mol Des* **17**, 241-53 (2003).

120.	Sharaf, M.A., Illman, D.L. & Kowalski, B.R. *Chemometrics*, (Wiley-Interscience, 1986).

121.	Tropsha, A. & Zheng, W. Identification of the descriptor pharmacophores using variable selection QSAR: applications to database mining. *Curr Pharm Des* **7**, 599-612 (2001).

122.	Vapnik, V. The Nature of Statistical Learning Theory. Springer-Verlag. *New York* (1995).

123.	Oloff, S., Mailman, R.B. & Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J Med Chem* **48**, 7322-32 (2005).

124.	Shen, M. et al. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J Med Chem* **47**, 2356-64 (2004).

125.	Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab an S4 package for kernel methods in R. *Journal of Statistical Software* **11**, 1-20 (2004).

126.	Team, R.D.C. R: A language and environment for statistical computing. . in *R Foundation for Statistical Computing* (Vienna, Austria, 2005).

127.	Golbraikh, A. & Tropsha, A. Beware of q2! *J Mol Graph Model* **20**, 269-76 (2002).

128.	Votano, J.R. et al. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J Med Chem* **49**, 7169-81 (2006).

129.	Zhang, S., Golbraikh, A., Oloff, S., Kohn, H. & Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J Chem Inf Model* **46**, 1984-95 (2006).

130. Zhu, H. et al. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J Chem Inf Model* **48**, 766-84 (2008).

131. Wang, D.F., Wiest, O., Helquist, P., Lan-Hargest, H.Y. & Wiech, N.L. QSAR studies of PC-3 cell line inhibition activity of TSA and SAHA-like hydroxamic acids. *Bioorg Med Chem Lett* **14**, 707-11 (2004).

132. Remiszewski, S.W. et al. Inhibitors of human histone deacetylase: synthesis and enzyme and cellular activity of straight chain hydroxamates. *J Med Chem* **45**, 753-7 (2002).

133. Shen, M. et al. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J Med Chem* **45**, 2811-23 (2002).

134. Hsieh, J.H., Wang, X.S., Teotico, D., Golbraikh, A. & Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J Comput Aided Mol Des* **22**, 593-609 (2008).

135. Medina-Franco, J.L., Golbraikh, A., Oloff, S., Castillo, R. & Tropsha, A. Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J Comput Aided Mol Des* **19**, 229-42 (2005).

136. Wang, X.S., Tang, H., Golbraikh, A. & Tropsha, A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J Chem Inf Model* **48**, 997-1013 (2008).

137. Zhang, S. et al. Antitumor agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J Comput Aided Mol Des* **21**, 97-112 (2007).

138. O'Connor, K.A. & Roth, B.L. Finding new tricks for old drugs: an efficient route for public-sector drug discovery. *Nat Rev Drug Discov* **4**, 1005-14 (2005).

139. Chen, Y.D., Jiang, Y.J., Zhou, J.W., Yu, Q.S. & You, Q.D. Identification of ligand features essential for HDACs inhibitors by pharmacophore modeling. *J Mol Graph Model* **26**, 1160-8 (2008).

140. Kozikowski, A.P. et al. Chemistry, biology, and QSAR studies of substituted biaryl hydroxamates and mercaptoacetamides as HDAC inhibitors-nanomolar-potency inhibitors of pancreatic cancer cell growth. *ChemMedChem* **3**, 487-501 (2008).

141. Ragno, R. et al. Class II-selective histone deacetylase inhibitors. Part 2: alignment-independent GRIND 3-D QSAR, homology and docking studies. *Eur J Med Chem* **43**, 621-32 (2008).

142.   Guo, Y. et al. Exploration of a binding mode of indole amide analogues as potent histone deacetylase inhibitors and 3D-QSAR analyses. *Bioorg Med Chem* **13**, 5424-34 (2005).

143.   Wagh, N.K., Deokar, H.S., Juvale, D.C., Kadam, S.S. & Kulkarni, V.M. 3D-QSAR of histone deacetylase inhibitors as anticancer agents by genetic function approximation. *Indian J Biochem Biophys* **43**, 360-71 (2006).

144.   Xie, A. et al. Quantitative structure-activity relationship study of histone deacetylase inhibitors. *Curr Med Chem Anticancer Agents* **4**, 273-99 (2004).

145.   Applegate, J.S. & Baer, K. Strategies for closing the chemical data gap. *Center for Progressive Reform (White Paper)* **602**, 1-19 (2006).

146.   Guth, J.H. BACKGROUND PAPER FOR REFORM NO. 5 OF THE LOUISVILLE CHARTER FOR SAFER CHEMICALS.

147.   Gibb, S. Toxicity testing in the 21st century: a vision and a strategy. *Reprod Toxicol* **25**, 136-8 (2008).

148.   Judson, R. et al. The toxicity data landscape for environmental chemicals. *Environ Health Perspect* **117**, 685-95 (2009).

149.   Collins, F.S., Gray, G.M. & Bucher, J.R. Transforming environmental health protection. *Science (New York, NY)* **319**, 906 (2008).

150.   Russell, W.M. The development of the three Rs concept. *Altern Lab Anim* **23**, 298-304 (1995).

151.   Johnson, D.E., Smith, D.A. & Park, B.K. Linking toxicity and chemistry: think globally, but act locally? *Curr Opin Drug Discov Devel* **7**, 33-5 (2004).

152.   Dearden, J.C. In silico prediction of drug toxicity. *J Comput Aided Mol Des* **17**, 119-27 (2003).

153.   Stouch, T.R. et al. In silico ADME/Tox: why models fail. *J Comput Aided Mol Des* **17**, 83-92 (2003).

154.   Benigni, R. et al. The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **25**, 53-97 (2007).

155.   Rabinowitz, J.R., Goldsmith, M.R., Little, S.B. & Pasquinelli, M.A. Computational molecular modeling for evaluating the toxicity of environmental chemicals: prioritizing bioassay requirements. *Environ Health Perspect* **116**, 573-7 (2008).

156. Knudsen, T.B. et al. Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. *Reprod Toxicol* **28**, 209-19 (2009).

157. Votano, J.R. et al. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **19**, 365-77 (2004).

158. Zhu, H. et al. A Novel Two-Step Hierarchical Quantitative Structure-Activity Relationship Modeling Work Flow for Predicting Acute Toxicity of Chemicals in Rodents. *Environmental Health Perspectives* **117**, 1257-1264 (2009).

159. Zhu, H. et al. Quantitative Structure-Activity Relationship Modeling of Rat Acute Toxicity by Oral Exposure. *Chemical Research in Toxicology* **22**, 1913-1921 (2009).

160. Claxton, L.D., Umbuzeiro, G.D. & Demarini, D.M. The Salmonella Mutagenicity Assay: The Stethoscope of Genetic Toxicology for the 21st Century. *Environ Health Perspect* **118**(2010).

161. U.S.EPA. EPA ToxCast. (2009).

162. Cohen Hubal, E.A. et al. Exposure science and the U.S. EPA National Center for Computational Toxicology. *J Expo Sci Environ Epidemiol* **20**, 231-6 (2010).

163. Breiman, L. Random forests. *Machine learning* **45**, 5-32 (2001).

164. Dalgaard, P. *Introductory statistics with R*, (Springer Verlag, 2008).

165. Vapnik, V.N. *The nature of statistical learning theory*, (Springer Verlag, 2000).

166. Pochet, N., De Smet, F., Suykens, J.A. & De Moor, B.L. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **20**, 3185-95 (2004).

167. Pirooznia, M., Yang, J.Y., Yang, M.Q. & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* **9 Suppl 1**, S13 (2008).

168. Peltason, L., Iyer, P. & Bajorath, J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J Chem Inf Model* **50**, 1021-33 (2010).

169. Mayer, P. & Reichenberg, F. Can highly hydrophobic organic substances cause aquatic baseline toxicity and can they contribute to mixture toxicity? *Environmental Toxicology and Chemistry* **25**, 2639-2644 (2006).

170. Collier, A.C., Tingle, M.D., Paxton, J.W., Mitchell, M.D. & Keelan, J.A. Metabolizing enzyme localization and activities in the first trimester human placenta:

the effect of maternal and gestational age, smoking and alcohol consumption. *Hum Reprod* **17**, 2564-72 (2002).

171. Myllynen, P., Pasanen, M. & Vahakangas, K. The fate and effects of xenobiotics in human placenta. *Expert Opin Drug Metab Toxicol* **3**, 331-46 (2007).

172. Karl, P.I., Gordon, B.H., Lieber, C.S. & Fisher, S.E. Acetaldehyde production and transfer by the perfused human placental cotyledon. *Science* **242**, 273 (1988).

173. Witorsch, R.J. & Thomas, J.A. Personal care products and endocrine disruption: A critical review of the literature. *Critical Reviews in Toxicology* **40**, 1-30 (2010).

174. Kavlock, R.J. et al. Computational toxicology--a state of the science mini review. *Toxicol Sci* **103**, 14-27 (2008).

175. Judson, R.S. et al. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect* **118**, 485-92 (2010).

176. Becker, O.M., Shacham, S., Marantz, Y. & Noiman, S. Modeling the 3D structure of GPCRs: advances and application to drug discovery. *Curr.Opin.Drug Discov.Devel.* **6**, 353-361 (2003).

177. Freddolino, P.L. et al. Predicted 3D structure for the human beta 2 adrenergic receptor and its binding site for agonists and antagonists. *Proc.Natl.Acad.Sci.U.S.A* **101**, 2736-2741 (2004).

178. Furse, K.E. & Lybrand, T.P. Three-dimensional models for beta-adrenergic receptor complexes with agonists and antagonists. *J.Med.Chem.* **46**, 4450-4462 (2003).

179. Kontoyianni, M., DeWeese, C., Penzotti, J.E. & Lybrand, T.P. Three-dimensional models for agonist and antagonist complexes with beta 2 adrenergic receptor. *J Med Chem* **39**, 4406-20 (1996).

180. Zhang, Y., Devries, M.E. & Skolnick, J. Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS.Comput.Biol.* **2**, e13 (2006).

181. Becker, O.M. et al. G protein-coupled receptors: in silico drug discovery in 3D. *Proc.Natl.Acad.Sci.U.S.A* **101**, 11304-11309 (2004).

182. Chen, J.Z., Wang, J. & Xie, X.Q. GPCR structure-based virtual screening approach for CB2 antagonist search. *J.Chem.Inf.Model.* **47**, 1626-1637 (2007).

183. Evers, A. & Klebe, G. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J.Med.Chem.* **47**, 5381-5392 (2004).

184. Evers, A. & Klabunde, T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. *J.Med.Chem.* **48**, 1088-1097 (2005).

185. Zhang, S., Golbraikh, A. & Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J.Med.Chem.* **49**, 2713-2724 (2006).

186. Oloff, S., Zhang, S., Sukumar, N., Breneman, C. & Tropsha, A. Chemometric analysis of ligand receptor complementarity: identifying Complementary Ligands Based on Receptor Information (CoLiBRI). *J Chem.Inf.Model.* **46**, 844-851 (2006).

187. Costanzi, S. On the applicability of GPCR homology models to computer-aided drug discovery: a comparison between in silico and crystal structures of the beta2-adrenergic receptor. *J Med Chem* **51**, 2907-14 (2008).

188. Wu, B. et al. Structures of the CXCR4 Chemokine GPCR with Small-Molecule and Cyclic Peptide Antagonists. *Science* (2010).

189. Jaakola, V.P. et al. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **322**, 1211-1217 (2008).

190. Palczewski, K. et al. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**, 739-745 (2000).

191. Warne, T. et al. Structure of a beta1-adrenergic G-protein-coupled receptor. *Nature* **454**, 486 (2008).

192. Park, J.H., Scheerer, P., Hofmann, K.P., Choe, H.W. & Ernst, O.P. Crystal structure of the ligand-free G-protein-coupled receptor opsin. *Nature* **454**, 183-7 (2008).

193. Visiers, I., Ballesteros, J.A. & Weinstein, H. Three-dimensional representations of G protein-coupled receptor structures and mechanisms. *Methods Enzymol.* **343**, 329-371 (2002).

194. Wishart, D.S. et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668-D672 (2006).

195. Okuno, Y., Yang, J., Taneishi, K., Yabuuchi, H. & Tsujimoto, G. GLIDA: GPCR-ligand database for chemical genomic drug discovery. *Nucleic Acids Res.* **34**, D673-D677 (2006).

196. Roth, B.L., Lopez, E., Patel, E.S. & Kroeze, W.K. The Multiplicity of Serotonin Receptors: Uselessly diverse molecules or an embarrasment of riches? *Neuroscientist* **6**, 252-262 (2006).

197. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).

198. Ghose, A.K., Viswanadhan, V.N. & Wendoloski, J.J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J.Comb.Chem.* **1**, 55-68 (1999).

199. Friesner, R.A. et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J.Med.Chem.* **47**, 1739-1749 (2004).

200. Huey, R., Morris, G.M., Olson, A.J. & Goodsell, D.S. A semiempirical free energy force field with charge-based desolvation. *J.Comput.Chem.* **28**, 1145-1152 (2007).

201. Morris, G.M. et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19**, 1639-1662 (1998).

202. Zsoldos, Z., Reid, D., Simon, A., Sadjad, S.B. & Johnson, A.P. eHiTS: a new fast, exhaustive flexible ligand docking system. *J.Mol.Graph.Model.* **26**, 198-212 (2007).

203. Tripos Inc. 1699 South Hanley Rd, S.L., Missouri 63144, USA. 8.0; Tripos Inc., 1699 South Hanley Rd, St. Louis, Missouri 63144, USA.

204. OpenEye Scientific Software Inc. Santa Fe, N., USA, www.eyesopen.com, 2010. OpenEye Scientific Software Inc. *Santa Fe, New Mexico* (2006).

205. Clark, A.M. & Labute, P. 2D depiction of protein-ligand complexes. *J Chem.Inf.Model.* **47**, 1933-1944 (2007).

206. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, V., Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

207. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).

208. Molecular Operating Environment. Chemical Computing Group Inc., M., Canada. CCG. Molecular Operation Eenvironment. 2007.09.

209. Spijker, P., Vaidehi, N., Freddolino, P.L., Hilbers, P.A. & Goddard, W.A., III. Dynamic behavior of fully solvated beta2-adrenergic receptor, embedded in the membrane with bound agonist or antagonist. *Proc.Natl.Acad.Sci.U.S.A* **103**, 4882-4887 (2006).

210. Avlani, V.A. et al. Critical role for the second extracellular loop in the binding of both orthosteric and allosteric G protein-coupled receptor ligands. *J.Biol.Chem.* **282**, 25677-25686 (2007).

211. Stern, P.S. et al. Molecular modeling of the binding of pheromone biosynthesis activating neuropeptide to its receptor. *J.Insect Physiol* **53**, 803-818 (2007).

212. Katritch, V. et al. Analysis of full and partial agonists binding to beta2-adrenergic receptor suggests a role of transmembrane helix V in agonist-specific conformational changes. *J Mol Recognit* **22**, 307-18 (2009).

213. Kota, P., Reeves, P.J., Rajbhandary, U.L. & Khorana, H.G. Opsin is present as dimers in COS1 cells: identification of amino acids at the dimeric interface. *Proc.Natl.Acad.Sci.U.S.A* **103**, 3054-3059 (2006).

214. Liang, Y. et al. Organization of the G protein-coupled receptors rhodopsin and opsin in native membranes. *J.Biol.Chem* **278**, 21655-21662 (2003).

215. Neve, K.A. et al. Structural Basis of Dopamine Receptor Activation. in *The Dopamine Receptors* (ed. Neve, K.A.) 47-73 (Humana Press, 2010).

216. Truchon, J.F. & Bayly, C.I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J.Chem.Inf.Model.* **47**, 488-508 (2007).

217. Clark, R.D. & Webster-Clark, D.J. Managing bias in ROC curves. *J Comput Aided Mol Des* **22**, 141-6 (2008).

218. Lefkowitz, R.J., Sun, J.P. & Shukla, A.K. A crystal clear view of the beta2-adrenergic receptor. *Nat.Biotechnol.* **26**, 189-191 (2008).

219. Vriend, G. WHAT IF: a molecular modeling and drug design program. *J.Mol.Graph.* **8**, 52-6, 29 (1990).

220. Pieper, U. et al. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **34**, D291-D295 (2006).

221. Rognan, D. Chemogenomic approaches to rational drug design. *British journal of pharmacology* **152**, 38-52 (2007).

222. Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. How many drug targets are there? *Nat Rev Drug Discov* **5**, 993-996 (2006).

223. Olsson, T. & Oprea, T.I. Cheminformatics: a tool for decision-makers in drug discovery. *Curr Opin Drug Discov Devel* **4**, 308-13 (2001).

224. Stumpfe, D., Geppert, H. & Bajorath, J. Methods for computer-aided chemical biology. Part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem Biol Drug Des* **71**, 518-28 (2008).

225. Vogt, I., Stumpfe, D., Ahmed, H.E. & Bajorath, J. Methods for computer-aided chemical biology. Part 2: Evaluation of compound selectivity using 2D molecular fingerprints. *Chem Biol Drug Des* **70**, 195-205 (2007).

226. Cross, S., Baroni, M., Carosati, E., Benedetti, P. & Clementi, S. FLAP: GRID molecular interaction fields in virtual screening. validation using the DUD data set. *J Chem Inf Model* **50**, 1442-50 (2010).

227. Radestock, S., Weil, T. & Renner, S. Homology model-based virtual screening for GPCR ligands using docking and target-biased scoring. *J Chem Inf Model* **48**, 1104-17 (2008).

228. Hoffmann, C., Leitz, M.R., Oberdorf-Maass, S., Lohse, M.J. & Klotz, K.N. Comparative pharmacology of human beta-adrenergic receptor subtypes--characterization of stably transfected receptors in CHO cells. *Naunyn Schmiedebergs Arch Pharmacol* **369**, 151-9 (2004).

229. Taira, C.A. et al. Therapeutic implications of beta-adrenergic receptor pharmacodynamic properties. *Curr Clin Pharmacol* **3**, 174-84 (2008).

230. Berkin, K.E. & Ball, S.G. Essential hypertension: the heart and hypertension. *Heart* **86**, 467-75 (2001).

231. Waldeck, B. Beta-adrenoceptor agonists and asthma--100 years of development. *Eur J Pharmacol* **445**, 1-12 (2002).

232. Rosenbaum, D.M. et al. GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science* **318**, 1266-73 (2007).

233. Cherezov, V. et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318**, 1258-65 (2007).

234. Rasmussen, S.G. et al. Structure of a nanobody-stabilized active state of the beta(2) adrenoceptor. *Nature* **469**, 175-80 (2011).

235. Huang, N., Shoichet, B.K. & Irwin, J.J. Benchmarking sets for molecular docking. *J Med Chem* **49**, 6789-801 (2006).

236. OpenEye Scientific Software. (Santa Fe, NM, USA, 2008).

237. Liu, Y. & Gray, N.S. Rational design of inhibitors that bind to inactive kinase conformations. *Nature Chemical Biology* **2**, 358-364 (2006).

238. Zhang, S., Golbraikh, A. & Tropsha, A. Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces. *J Med Chem* **49**, 2713-24 (2006).

239. Liu, S.B. Conceptual Density Functional Theory and Some Recent Developments. **25**, 590-600 (2009).

240. Hanson, M.A. et al. A Specific Cholesterol Binding Site Is Established by the 2.8 ?Structure of the Human [beta] 2-Adrenergic Receptor. *Structure* **16**, 897-905 (2008).

241. Wacker, D. et al. Conserved Binding Mode of Human beta2 Adrenergic Receptor Inverse Agonists and Antagonist Revealed by X-ray Crystallography. *Journal of the American Chemical Society* **132**, 11443-11445 (2010).

242. Warren, G.L. et al. A critical assessment of docking programs and scoring functions. *J Med Chem* **49**, 5912-31 (2006).

243. Yin, S., Biedermannova, L., Vondrasek, J. & Dokholyan, N.V. MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J Chem Inf Model* **48**, 1656-62 (2008).

244. Jain, A.N. & Nicholls, A. Recommendations for evaluation of computational methods. *J Comput Aided Mol Des* **22**, 133-9 (2008).

245. Kikkawa, H., Kurose, H., Isogaya, M., Sato, Y. & Nagao, T. Differential contribution of two serine residues of wild type and constitutively active beta2-adrenoceptors to the interaction with beta2-selective agonists. *Br J Pharmacol* **121**, 1059-64 (1997).

246. Kikkawa, H., Isogaya, M., Nagao, T. & Kurose, H. The role of the seventh transmembrane region in high affinity binding of a beta 2-selective agonist TA-2005. *Mol Pharmacol* **53**, 128-34 (1998).

247. Williams, R.S. & Bishop, T. Selectivity of dobutamine for adrenergic receptor subtypes: in vitro analysis by radioligand binding. *J Clin Invest* **67**, 1703-11 (1981).

248. Pieper, U. et al. MODBASE, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **37**, D347-54 (2009).

249. Vroling, B. et al. GPCRDB: information system for G protein-coupled receptors. *Nucleic Acids Res* **39**, D309-19 (2010).

250. Kellenberger, E. et al. Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *J Med Chem* **50**, 1294-303 (2007).

251. Salo, O.M. et al. Virtual screening of novel CB2 ligands using a comparative model of the human cannabinoid CB2 receptor. *J Med Chem* **48**, 7166-71 (2005).

252. Tikhonova, I.G. et al. Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *J Med Chem* **51**, 625-33 (2008).

253. Dubocovich, M.L., Rivera-Bermudez, M.A., Gerdin, M.J. & Masana, M.I. Molecular pharmacology, regulation and function of mammalian melatonin receptors. *Front Biosci* **8**, d1093-108 (2003).

254. Perola, E., Walters, W.P. & Charifson, P.S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**, 235-49 (2004).

255. Kiss, R. et al. Binding mode analysis and enrichment studies on homology models of the human histamine H4 receptor. *Eur J Med Chem* (2007).

256. Jones, S.W. et al. The orphan G-protein coupled receptor RDC1: evidence for a role in chondrocyte hypertrophy and articular cartilage matrix turnover. *Osteoarthritis Cartilage* **14**, 597-608 (2006).

257. Tikhonova, I.G. et al. Bidirectional, iterative approach to the structural delineation of the functional "chemoprint" in GPR40 for agonist recognition. *J Med Chem* **50**, 2981-9 (2007).

258. Gan, L.L., Wang, M.W., Cheng, M.S. & Pan, L. Trachea relaxing effects and beta2-selectivity of SPFF, a newly developed bronchodilating agent, in guinea pigs and rabbits. *Biol Pharm Bull* **26**, 323-8 (2003).