VIRTUAL SCREENING OF MULTI-TARGET AGENTS BY COMBINATORIAL MACHINE LEARNING METHODS

SHI ZHE

(*B.Sc, Shandong University*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2011

Acknowledgements

As Jiddu Krishnamurti once said "The whole of life, from the moment you are born to the moment you die, is a process of learning." If there were any time when I would appreciate this more than any other point in my life so far, I would say it were the four-year PhD life of mine. The time I spent in National University of Singapore (NUS) and Singapore during the pursuit of the PhD degree is a precious gem in my life which has greatly expended the horizon of my minds through the process of learning, both in academic and personal aspects.

This learning process would not have become this meaningful without the encountering and interacting with the many wonderful people I have met during the past four years. Even millions of sincere thanks would not be enough to count for my gratefulness toward them.

First of all, I would like to express my foremost appreciation and thanks to Prof. Chen Yuzong who has been a great mentor throughout my four-year studying and research in NUS. He has been a very inspiring supervisor for my research work. His enthusiasm and dedication to research, his insight in science discovery, his critical thinking, his hard working spirit, and his humbleness has always been enlightening to me. He has provided for me invaluable guidance in bioinformatics and chemoinformatics research. I am especially grateful for his great patience and efforts in cultivating a good environment for my growth in research area with inspiring ideas and supervision. The great influence of him, however, is not limited to research area. He is also a wise person with insightful understanding of life who is ever so willing to share with others the principles and disciplines in life that can benefit a person to live a fulfilling life. I'd like to express my utmost gratefulness to Prof Chen Yuzong and wish him the very best to his work and life.

My many thanks also go to the wonderful BIDD group members. It has been a very pleasant time working with them. They have offered me great companionship and inspiration not only on research but also on personal life. I would like to thank each and every one of them for their collaboration and company in the past four years. I would like to thank Ms Hai Lei and Ms Wang Rong, even though they left BIDD not too long after I joined the group, for their kindness as seniors. My very gratefulness goes to Ms Ma Xiaohua, Mr Zhufeng and Ms Jia Jia. As seniors, they have been playing motivating role models for us juniors to look up to. Ms Ma Xiaohua has been amazingly helpful and supportive with my research work. She is quite knowledgeable and resourceful in chemoinformatcs and is always so patient in answering and discussing research questions. She tries her very best to help when someone turns to her. Ms Ma Xiaohua is also a wonderful person with a big heart. She cares for us as her friends. I couldn't thank her more for her supportiveness and kindness. She is one of the best persons one could have as a workmate and a friend. Ms Jia Jia is an inspiring figure with a strong fighting spirit. Her courage and efforts in pursing her goals in life has always inspired me. Mr Zhu Feng has been a wonderful collaborator in research. His great attitudes toward research and his never ending efforts to perfection in work have deeply impressed me to look up to. Meanwhile, he presents a strong sense of team-work spirit which has made the collaborations with him both very pleasant and fruitful. Ms Liu Xin has always inspired me with her great effort to present and make the best out of the tasks she has taken. I am very honored to have been able to work with them and learned so many valuable lessons from them. I would also like to thank my juniors, Ms Wei Xiaonai, Mr Zhang Jingxian, Mr Han Bucong, Mr Tao Lin, Ms Qin Chu and Mr Zhang Cheng for their assistance in research collaboration.

My learning process would not have been complete without the great lessons from life itself outside the academic research. I have always felt so fortunate to have met many wonderful and inspiring people from across the globe and become friends with some awesome individuals. The landscapes of my minds have become so much more extended and enlightened because of them. Their companies have made my time in Singapore a wonderful and interesting experience. To name a few, I would like to thank Ms Sit Wing Yee for her great friendship. I really appreciate her supportiveness in times of need. My very gratitude goes to Ms Laureline Josset, Ms Zhao Yangyang, Mr Zhang Yaoli, Mr Maximilian Klement, Mr Evan Conover and Mr Michael Stratil for believing in me and encouraging me to be who I am. I would also want to thank my awesome rock climbing friends, to name a few, Mr Remi Trichet, Mr Michael Stratil, Mr Siddharth Batra and Mr Hassan Arif. The climbing experiences with them have made me strong in body and mind.

Last but not least, my utmost gratefulness goes to my wonderful parents and families for their everlasting love and support. I could never thank my parents more for their love for me and their efforts in bringing the best out of me as a person. To my beloved parents, I dedicate this thesis.

Shi Zhe

September 2011

Table of Contents

Summary

Multi-target drugs have greatly attracted the attention and interest in drug discovery. Efforts that explore experimental and *in-silico* methods have been and are being made in search for the novel multi-target agents. As part of the collective efforts for developing the tools to facilitate discovery multi-target agents, I firstly participated in the updated the Kinetics database of bio-molecular interactions (KDBI) and the Therapeutic targets database (TTD). The information in the two databases can offer informative data in multi-target drug discovery.

Virtual screening (VS) is an increasingly used approach in the search for novel lead compounds. It is capable of providing valuable contributions in hit and lead compounds discovery. It has been intensively explored and various software tools have been developed for the application of VS. It would be very interesting to apply VS tools for the discovery of multi-target agents. However, many of the conventional VS tools encounter the issues of the insufficient coverage of compound diversity, high false positive, high false negative prediction and lower speed in screening large libraries. These issues would hinder the practical applications of conventional VS approaches in search of multi-target agents. Therefore, in order to identify multi-target agents that are more sparsely distributed in the chemical space than single-target agents, it is important to address these issues and develop the methods that are capable of searching large compound libraries at good yields and low false-hit rates.

In this work, I explored a machine learning method, support vector machines (SVM), to develop the combinatorial SVM (COMBI-SVM) VS tool for searching dual-target agents for the treatment of cancers and major depression. COMBI-SVMs models were preliminarily tested for searching dual-inhibitors of 4 combinations (EGFR-FGFR, EGFR-Src, VEGFR-Lck, and Src-Lck) of the 5 anticancer kinase targets (EGFR, VEGFR, Src, FGFR, Lck). COMBI-SVMs produced comparable dual-inhibitor yields and significantly lower false-hit rates for MDDR and PubChem dataset. There has been underpinning interest in discovery and developing selective multi-target serotonin reuptake inhibitors (SRIs) that can enhance antidepressant efficacy (1). The preliminary tests with the 4 kinase dual-inhibitors showed promising results and this encouraged me to develop and test COMBI-SVMs for VS multi-target serotonin reuptake inhibitors of 7 target pairs (serotonin transporter paired with noradrenaline transporter, H_3 receptor, 5-HT_{1A} receptor, 5-HT_{1B} receptor, 5-HT_{2C} receptor, Melanocortin 4 receptor and Neurokinin 1 receptor respectively) from large compound libraries. COMBI-SVMs showed moderate to good target selectivity in misidentifying individual-target inhibitors of the same target pair and inhibitors of the other target six pairs as dual-inhibitors; COMBI-SVMs also presented low dual-inhibitor false-hit rates in screening large compound databases MDDR and PubChem. Compared to the other three VS methods (similarity searching, k-NN and PNN), it produced comparable dual-inhibitor yields, similar to or slightly better target selectivity, and slightly to or substantially lower false-hit rate in screening MDDR compounds.

List of Tables

List of Figures

List of Acronyms

5HT1aAntags	5-HT$_{1A}$ receptor antagonists
5HT1aSRIs	Dual serotonin reuptake inhibitor and 5-HT$_{1A}$ receptor antagonists
5HT1bAntags	and 5-HT$_{1B}$ receptor antagonists
5HT1bSRIs	Dual serotonin reuptake inhibitor and 5-HT$_{1B}$ receptor antagonists
5HT2cAntags	5-HT$_{2C}$ receptor antagonists
5HT2cSRIs	Dual serotonin reuptake inhibitor and 5-HT$_{2C}$ receptor antagonists
CNS	Central nervous system
COMBI-SVM	Combinatorial support vector machines
DI	Diversity index
EGFR	Epidermal growth factor receptor
FDA	The Food and Drug Administration
FGFR	Fibroblast Growth Factor Receptor
FN	False negative
FP	False positive
H3Antags	H$_3$ receptor antagonists
H3SRIs	Dual serotonin reuptake inhibitor and H$_3$ receptor antagonists
HTS	High throughput screening
KDBI	Kinetics database of biomolecular interactions

k-NN	k-nearest neighbors
LBVS	Ligand-based Virtual Screening
Lck	Lymphocyte-specific protein tyrosine kinase
MC_4	Melanocortin 4
MC4Antags	MC_4 receptor antagonists
MC4SRIs	Dual serotonin reuptake inhibitor and MC_4 receptor antagonists
MCC	Matthews correlation coefficient
MDDR	MDL Drug Data Report
ML	Machine Learning
NCEs	Novel chemical entities
NET	Noradrenaline transporter
NETSRI	Dual serotonin reuptake and noradrenaline reuptake inhibitors
NK_1	Neurokinin 1
NK1Antags	NK_1 receptor antagonists
NK1SRIs	Dual serotonin reuptake inhibitor and NK_1 receptor antagonists
NRIs	Noradrenaline reuptake inhibitors
PNN	Probabilistic neural network
QSAR	Quantitative structure activity relationship
SAR	Structure-activity relationship
SERT	Serotonin transporter
SBML	System Biology Markup Language

SBVS	Structure-based Virtual Screening
SRI	Serotonin reuptake inhibitor
SSNI	Serotonin/noradrenaline reuptake inhibitor
SSRI	Serotonin reuptake inhibitor
SVM	Support vector machine
TN	True negative
TP	True positive
TK	Tyrosine kinase
TKI	Tyrosine kinase inhibitor
TTD	Therapeutic targets database
VEGFR	Vascular endothelial growth factor receptor
VS	Virtual Screening

List of Publications

1. Combinatorial Support Vector Machines Approach for Virtual Screening of Selective Multi-Target Serotonin Reuptake Inhibitors from Large Compound Libraries. **Z.Shi**, X.H.Ma, C.Qin, J.Jia, Y.Y.Jiang, C.Y.Tan, Y.Z.Chen. ***Journal of Molecular Graphics and Modelling.*** (Impact Factor: 2.033) Accepted, **(2011)**.

2. Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. F. Zhu, C. Qin, L. Tao, X. Liu, **Z. Shi**, X.H. Ma, J. Jia, Y. Tan, C. Cui, J.S. Lin, C.Y. Tan, Y.Y. Jiang and Y.Z. Chen. ***PNAS*****.** (Impact Factor: 9.771) 108(31):12943-8 **(2011)**.

3. Therapeutic Target Database Update 2012: A Resource for Facilitating Target-Oriented Drug Discovery. Zhu, Feng; **Shi, Zhe**; Qin, Chu; Tao, Lin; han, bucong ,; Zhang, Peng; Chen, Yuzong. ***Nucleic Acids Res.*** **Submitted** (Impact factor: 7.836) **(2011) (submitted)**

4. In-Silico Approaches to Multi-Target Drug Discovery. H.X. Ma, **Z. Shi**, C.Y. Tan, Y.Y. Jiang, M.L. Go, B.C. Low and Y.Z. Chen.***Pharm Res.***(Impact factor: 4.456) 27(5):2101-10 **(2010)**.

5. Update of KDBI: Kinetic Data of Bio-molecular Interaction Database. P. Kumar, Z.L. Ji, B.C. Han, **Z. Shi**, J. Jia, Y.P, Wang, Y.T. Zhang, L. Liang, and Y. Z. Chen. ***Nucleic Acids Res***._37(Database issue): D636-41**(2009)**.

Chapter 1 Introduction

Considerable efforts have been put into drug design; however, the number of successful drugs did not increase appreciably during the past decade. Recent evidence suggests that the main causes of failure of compounds in the clinic are lack of efficacy and poor safety. Agents that modulate multiple targets simultaneously have the potential to enhance efficacy or improve safety relative to drugs that modulate only a single target. As a result, multi-target agents have been gaining increasing interest of researchers and drug discovery teams. To assist the research of multi-target discovery, I participated in the further development of two pharmainformatics databases, i.e., the update of KDBI and BIDD. As a complementary approach to the traditional chemical and biological methods, virtual screening has aroused increasing attention in the pharmaceutical industry as a productive and cost-effective technology (2). Various computational screening tools, such as docking, quantitative structure activity relationship (QSAR), support vector machines (SVM), k-NN, PNN etc, are being developed and refined to effectively employ fast screening methods to yield potent lead hits. In my work, the combinatorial SVM (COMBI-SVM) virtual screening (VS) tool was developed for searching multi-target agents. This method was firstly tested with four anticancer kinase target pairs and then was applied to seven antidepressants target pairs. Compared with the other three VS methods, i.e., similarity searching, k-NN and PNN, COMBI-SVM produced comparable dual-inhibitor yields, similar to or slightly better target selectivity, and slightly to or substantially lower false-hit rate in screening MDDR compounds.

The following sections present a brief introduction to development of pharmainformatics databases (Section 1.1), an overview of methods in virtual screening (Section 1.2) and in-silico approaches to multi-target drug discovery (Section 1.3). In addition, the outline of this thesis (Section 1.4) is introduced.

1.1 Pharmainformatics Database Development and Updates

With the exponential increase in pharma-information, it is becoming increasingly necessary and important to collect and curate the information to provide informative sources to effectively assist the studies of disease mechanisms and the discovery of new drugs. Pharmainformatics databases can provide up-to-date information and data that relate to disease mechanism studies, pharmaceutical research and drug development. They offer various types of information for a number of interdisciplinary areas such as bioinformatics, chemoinformatics, drug data, bioactive compound data, interaction and kinetics data, *in- silico* ADME-Tox prediction and molecular modeling.

The process of a database construction consists of two major steps. The first step is data collection and quality control. The quantity and quality of the data are decisive to the usefulness and popularity of a database. The second step involves database interface design and maintenance. Well-designed databases usually share the following qualities: informative with a clear presentation; user-friendly with easy manipulation; fast and accurate search within the database; Continuous updates with new information, data and other features. Additional qualities include data download, inter links to other related databases and data processing functions for the personalized data.

In this work, I participated in the update of the Kinetics database of bio-molecular interactions (KDBI) http://xin.cz3.nus.edu.sg/group/kdbi/kdbi.asp (3) and the Therapeutic targets database (TTD) http://bidd.nus.edu.sg/group/ttd/ (4).

KDBI stores the kinetic information of bio-molecular interactions. This information is essential for quantitative studies of the interactions between bio-molecules of a given bio-system (3). Numerous improvements and updates have been added to KDBI, including new ways to access data by pathway and molecule names, data file in System Biology Markup Language (SBML) format. It can accommodate the increasing data demand in quantitative system biology studies which play an important role in understanding the mechanisms underlying many complex diseases.

TTD has been developed to provide comprehensive information about the known targets and the corresponding approved, clinical trial and investigative drugs. Since its last update in 2010, major improvements and updates have been made to TTD. These updates include a significant increase of data content, target validation information and quantitative structure activity relationship (QSAR) models.

1.2 Introduction to Virtual Screening in Drug Discovery

Traditionally, the progress in drug discovery has been made by a combination of random screening and rational design (5). Given the mounting competiveness of pharmaceutical industry, high throughput screening (HTS) has become a key tool in many pharmaceutical companies for its ability to test vast number of compounds quickly and efficiently. However, HTS offers no guarantee of success and over-reliance on random HTS are showing apparent problems. Additionally, establishing a robust assay is very costly: a single HTS programme without assay development could still cost approximately US $75,000 (6). Moreover, collections of synthesized compounds or natural products can only represent a limited space in the entire drug-like chemical space. The typical screening collection of a large pharmaceutical company is of the order of a few million compounds at most. This is a tiny fraction of the huge chemical space (7, 8), which is many orders of magnitude larger than this, even if only drug-like compounds are considered (9). Given these caveats, it is worth evaluating other technologies that may complement HTS assay and synthesis. The term 'virtual screening' first came into being in 1997; it has been used to describe a process of computationally analyzing large compound collections in order to prioritize compounds for synthesis or assay. During the last decade, a broad range of computational techniques have been applied to search for novel bioactive compounds for many targets. VS method does not require the physically synthesized compound libraries such greatly recedes the cost. This also potentially extends the exploration of the chemical space outside the in-house compound pools. There are around 10 million commercially available compounds that can be exploited with the VS approach. On top of it, virtual combinatorial libraries contain at least 1 million-fold larger libraries than those available for HTS. This adds a new dimension to the VS search space (**Figure 1-1**).

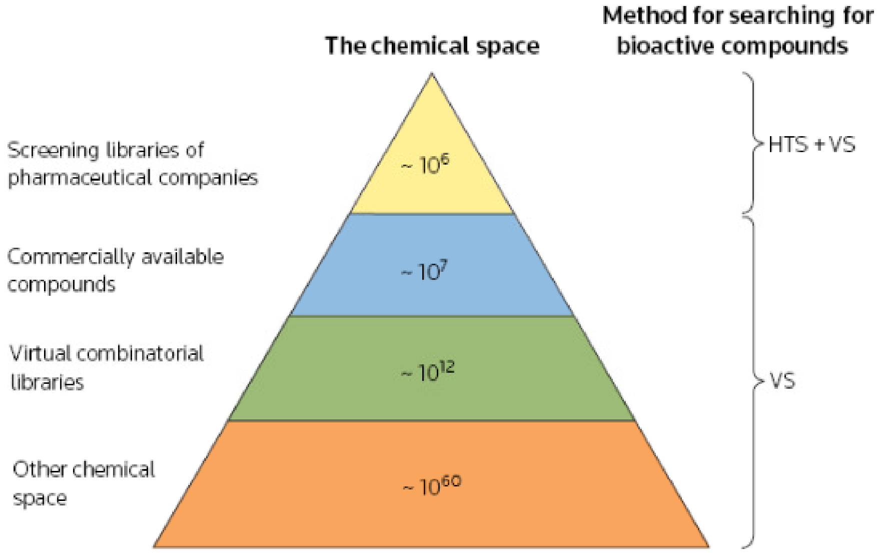

The chemical space
Method for searching for bioactive compounds
Screening libraries of pharmaceutical companies
~ 10^6
Commercially available compounds
~ 10^7
Virtual combinatorial libraries
~ 10^{12}
Other chemical space
~ 10^{60}
HTS + VS
VS

Figure 1-1 Typical numbers of compounds available in the chemical space

Based on the requirement of either the structure of a target or its ligands, virtual screening methods can be often classified into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) (10). SBVS consists of the virtual docking of candidate ligands into a protein target followed by the estimation of the probability of the high affinity binding between them calculated by a scoring function (11, 12). LBVS methods, such as pharmacophore methods (13) and chemical similarity analysis methods (14), require the ligand structure information, they focus on discoverying the new drug hits by analyzing the physical and chemical similarities of known compound pools by computational means.

Figure 1-2 shows the general procedure used in SBVS and LBVS.

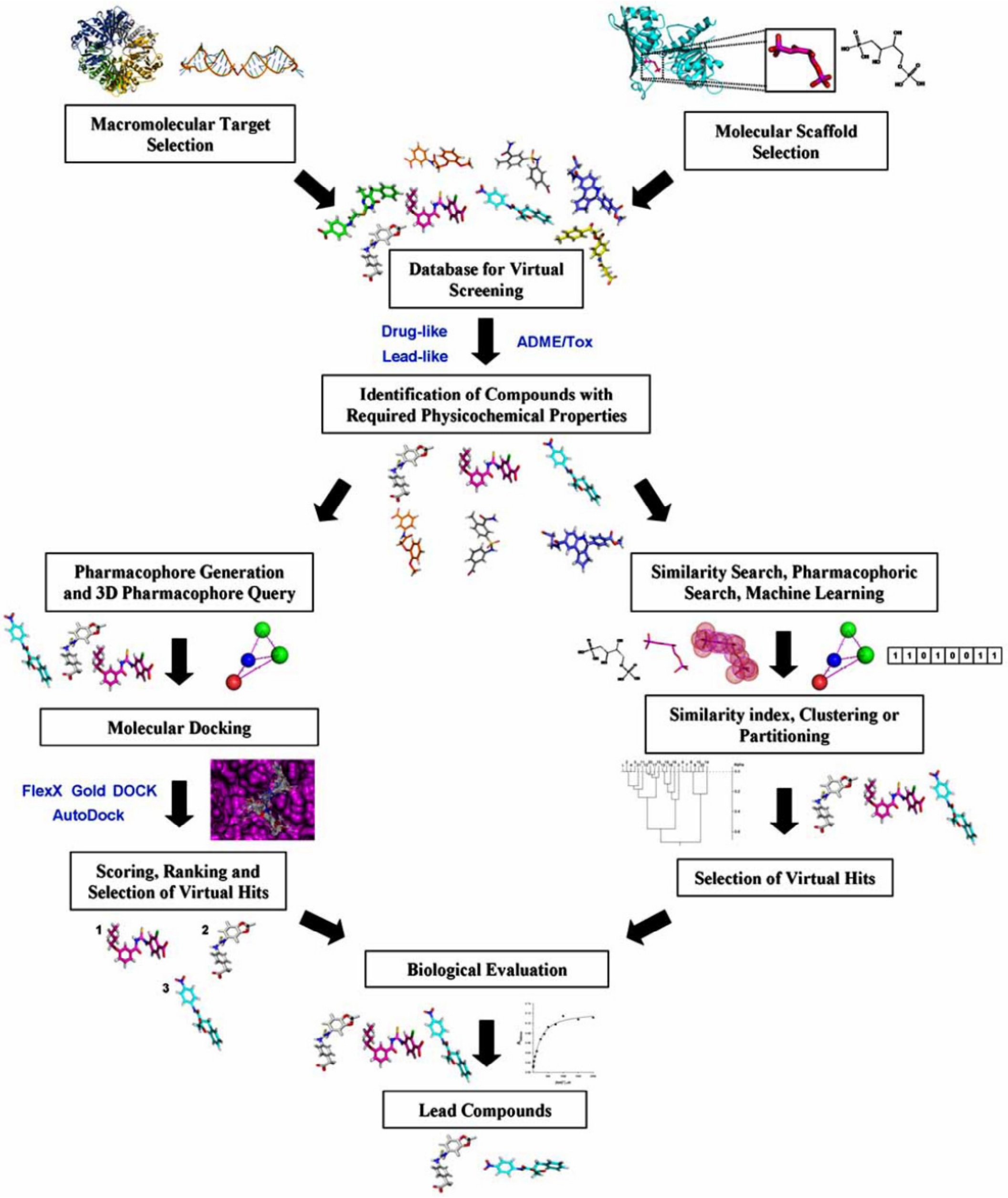

Macromolecular Target Selection
Molecular Scaffold Selection
Database for Virtual Screening
Drug-like
Lead-like
ADME/Tox
Identification of Compounds with Required Physicochemical Properties
Pharmacophore Generation and 3D Pharmacophore Query
Similarity Search, Pharmacophoric Search, Machine Learning
1 1 0 1 0 0 1 1
Molecular Docking
Similarity index, Clustering or Partitioning
FlexX Gold DOCK
AutoDock
Scoring, Ranking and Selection of Virtual Hits
Selection of Virtual Hits
1
2
3
Biological Evaluation
Lead Compounds

Figure 1-2 General procedure used in SBVS and LBVS (adopted from Rafael V.C. et al(10)).

1.2.1 Structure-based and ligand based virtual screening

Structure-based virtual screening (SBVS) starts with a 3-D structure of a target protein and a database of the 3-D structures of ligands as the screening pool. It is usually applied when the 3D structure of a protein target, derived either from experimental data (X-ray or NMR spectroscopy) or from homology modeling, is available. SBVS procedure consists of docking and scoring. The docking algorithms (11, 12) are designed to predict the ligand conformation and orientation within the targeted active site of the target. The scoring methods are empirically or semi-empirically derived to attempt (13) to estimate the binding tightness of the ligand and the protein in bound complexes. Docking and scoring algorithms are combined to detect the compounds with higher affinity against a target by predicting their binding mode (by docking) and affinity (by scoring), and retrieving those with the highest scores. To date, more than 60 docking programs and 30 scoring functions have been reported (14, 15). The major drawback with SBVS is the unavailability of appropriate scoring functions to differentiate between correct and incorrect poses of bound ligands and identifying false negative and positive hits. Some of the key challenges encountered by SBVS include the appropriate treatment of ionization, tautomerization of ligand and protein residues, target/ligand flexibility, choice of force fields, solvation effects, dielectric constants, exploration of multiple binding modes and, most importantly, the approximations in the scoring functions that lead to false-positives and missed true-hits. Moreover, most docking algorithms and scoring functions are tuned towards high throughput, which requires a compromise between the speed and accuracy of binding mode and energy prediction. Despite the successful drug discovery cases, currently there has not been a single docking program that outperforms all others with regard to either docking accuracy or hit enrichment. The hit enrichment is defined as the fraction of true active compounds in, for example, the upper 1% of the ranked VS hit list compared with the average fraction of active compounds in the search space. The performance of a docking program is difficult to evaluate in advance, and depends on the nature and quality of the target structure (14-16). Despite all optimization efforts, the currently available scoring functions do not provide reliable estimates of free binding energies, and are not able to rank compounds according to affinity (15, 17). The published comparisons of docking programs have been critically reviewed (18-20).

Ligand-based virtual screening (LBVS) does not require the target structure information. Instead, it uses the structure(s) of one or more active compounds as template(s) to indentify a new compound pool by chemical and physical similarities. In general, the application of LBVS methods employ the computational descriptors of molecular structure, properties, or pharmacophore features and analyze relationships between the active compounds and test compounds. Complex descriptors are designed to detect similarities in molecular shape and shape-related properties in order to find new hits. LBVS is computationally efficient and can scan very large databases in reasonably short time. As a result, it is often applied to sequentially filter large compound sets before more complex tools are applied. A considerable number of types of different methods have been reported with literally thousands of different descriptors. These descriptors are derived from the 2D or 3D distribution of atomic properties of the known compounds, or from the presence of specific structural elements. Many methods designed for the comparison of the similarity of compounds based on these descriptors. Shape comparison (21) and pharmacophore searches are frequently-used long-established techniques (22, 23). Other methods apply molecular fields to define the similarity of structures (24, 25). When large sets of active and inactive compounds are known, machine learning techniques, such as artificial neural nets, decision trees, support vector machines or Bayesian classifiers, can be used to train models that can distinguish active from inactive compounds based on their specific structural features. Comprehensive overviews of ligand-based VS have been presented in a number of reviews (26, 27). **Table 1-2, 1-3, 1-4, 1-5** show the performances of some frequently applied SBVS and LBVS methods for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance.

1.2.2 Conventional approaches of virtual screening methods

Conventional VS approaches such as docking have been widely studied for facilitating lead discovery against individual targets (28-30). Among the various conventional methods, molecular docking (31), pharmacophore (32), structure-activity relationship (SAR) and quantitative structure activity relationship

(QSAR) (33), similarity searching (34) have been extensively used for searching and designing active compounds against individual targets.

1.2.3 Machine learning methods for virtual screening

Machine learning classification methods use binary, categorical or continuous descriptors to estimate the probability of a molecule to be active on the basis of learning sets. Machine learning methods can be classified as supervised or unsupervised. If instances are given with known labels then the learning is called supervised (**Table 1-1**) whereas instances are unlabeled in unsupervised learning.

Data in standard descriptor format					
Case	**Feature 1**	**Feature 2**	**…**	**Feature n**	**Class**
1	Charge: 0	Benzene ring: 1		Nitrogen: 2	Active
2	Charge:+1	Benzene ring: 2		Nitrogen: 3	Active
3	Charge:-1	Benzene ring: 3		Nitrogen: 1	Inactive
…					…

Table 1-1 Instances of supervised machine learning methods

Commonly utilized supervised machine learning methods include Support Vector Machine (SVM), Artificial Neural Network, Decision tree learning, Inductive logic programming, Boosting, Gaussian process regression etc. Unsupervised machine learning with the unlabeled training aims at finding the internal organization of the data. Examples of unsupervised machine learning include Clustering, Adaptive Resonance Theory, and Self Organized Map.

Compared to SBVS and other LBVS methods such as QSAR, pharmacophore and clustering methods (35-42), machine learning methods are more capable of working with a more diverse spectrum of compounds and more complex structure-activity relationships. This is because machine learning methods apply complex nonlinear mappings from molecular descriptors to activity classes without restriction on structural frameworks, and they do not require prior knowledge of relevant molecular descriptors and functional form of structure-activity relationships (43-47). Additionally, machine learning methods can overcome several problems that have obstructed some conventional virtual screening tools (28, 44). These obstacles include the extensiveness and discreteness natures of the chemical space, the absence of protein target structures (current statistics shows that the known protein sequences (~1,000,000)(48) vastly outnumber the available protein structures (~20,000)(49)), complexity and flexibility of target structures, limited diversity caused by the biased training molecules, and difficulties in computing binding affinity and solvation effects.

The performance report of machine learning methods in screening pharmacodynamically active compounds from libraries of >25,000 compounds is summarized in **Table 1-2**. These reported studies (50-57) primarily focused on the prediction of compounds that inhibit, antagonize, block, agonize, or activate specific therapeutic target proteins. The majority of the reported screening tasks by machine learning methods are found to demonstrate good performances. The yields, hit rates, and enrichment factors of machine learning methods are in the range of 50%~94%, 10%~98%, and 30~108 respectively.

Tentative comparisons are presented in **Table 1-3**, **Table 1-4** and **Table 1-5** for the reported performances of structure-based VS methods and two classes of ligand-based VS methods, pharmacophore and clustering. The majority of the yields, hit rates, and enrichment factors lay in the range of 7%~95%, 1%~32%, and 5~1189 for structure-based, 11%~76%, ~0.33%, and 3~41 for pharmacophore, and 20%~63%, 2%~10%, and 6~54 for clustering methods respectively. Therefore, the general performance of machine learning methods appears to be comparable to or in some cases better than the reported performances of the conventional VS studies such as pharmacophore and clustering methods. In screening extremely-large libraries, the reported yields, hit-rates and enrichment factors of machine learning VS tools are in the range of 55%~81%, 0.2%~0.7% and 110~795 respectively, compared to those of 62%~95%, 0.65%~35% and 20~1,200 by structure-based VS tools. The reported hit-rates of some machine learning VS tools are comparable to those of structure-based VS tools in screening libraries of ~98,000 compounds, but their enrichment factors are substantially smaller. Therefore, while exhibiting equally good yield, in screening extremely-large (≥1 million) and large (130,000~400,000) libraries, the currently developed machine learning VS tools appear to show lower hit-rates and, in some cases, lower enrichment factors than the best performing structure-based VS tools.

The machine learning methods employed in this work are SVM, Probabilistic Neural Network (PNN) and k nearest neighbor (k-NN). They are explained below in subsequent sub sections. For a comparative study, Tanimoto similarity searching method is also introduced.

Table 1-2 Performance of machine learning methods in virtual screening test for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance. The relevant literature references are given in the method column.

Screening task	**Compounds screened**		**Method and reference of reported study**	**Molecular descriptors**	**Compounds in training set (No of positives / No of negatives)**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**				**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
COX2 inhibitors	2.5M	22	SVM (58)	Molecular fingerprints	94/200K	2,500	0.1%	18	81%	0.7%	795
	25,300	25	SVM+ BKD (59)	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5
COX inhibitors	102,514	536	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	76	14.3%	1.4%	2.7
	98,435	536	CKD (47)	Pipeline pilot	100/4000	984	1%	232	43.4%	23.7%	43.1
				ECFP4	100/4000	984	1%	365	68.1%	37.2%	67.7
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	240	44.7%	24.4%	44.5
Thrombin inhibitors	2.5M	46	SVM (58)	Molecular fingerprints	188/200K	11,250	0.45%	25	55%	0.2%	108.7
	102,514	703	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	367	52,3%	7.1%	10.3
	98,435	703	CKD (47)	Pipeline pilot	100/4000	984	1%	435	61.9%	44.4%	61.7
				ECFP4	100/4000	984	1%	603	85.8%	61.5%	85.5
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	381	54.2%	38.9%	54.0
Protease inhibitors	171,726	118	SVM (62)	Extended connectivity fingerprints	228/4200	1717	1%	26	22%	1.5%	21.8
			LMNB					19	16%	1%	14.5

			(62)								
Chemokine receptor antagonists	171,560	128	SVM (62)	Extended connectivity fingerprints	258/4199	1716	1%	70	55%	4.1%	54.9
			LMNB (60, 63)					68	53%	3.9%	52.3
5HT3 antagonists	102,514	652	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	236	36.3%	4.6%	7.2
	98,435	852	CKD (47)	Pipeline pilot	100/4000	984	1%	480	56.4%	49.0%	56.3
				ECFP4	100/4000	984	1%	680	79.8%	69.4%	79.8
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	529	62.1%	54.0%	62.1
5HT1A antagonists	102,514	727	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	224	30.9%	4.3%	6.1
	98,435	727	CKD (47)	Pipeline pilot	100/4000	984	1%	268	36.9%	27.3%	36.9
				ECFP4	100/4000	984	1%	426	58.6%	43.5%	58.7
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	319	43.9%	32.6%	44.0
5HT reuptake inhibitors	102,514	259	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	65	25%	1.2%	4.7
	98,435	259	CKD (47)	Pipeline pilot	100/4000	984	1%	131	50.7%	13.4%	51.5
				ECFP4	100/4000	984	1%	194	75.6%	19.7%	75.9
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	137	52.9%	14.0%	53.8
D2 antagonists	102,514	295	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	90	30.6%	1.7%	5.9
	98,435	295	CKD (47)	Pipeline pilot	100/4000	984	1%	132	44.7%	13.5%	44.9
				ECFP4	100/4000	984	1%	219	74.4%	22.4%	74.7
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	137	46.4%	14.0%	53.8

Rennin inhibitors	102,514	1030	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	972	94.4%	18.9%	18.9
	98,435	1030	CKD (47)	Pipeline pilot	100/4000	984	1%	842	81.8%	86.0%	81.9
				ECFP4	100/4000	984	1%	960	93.2%	98.0%	93.3
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	710	68.9%	72.4%	69.0
Angiotesin II AT1 antagonists	102,514	843	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	776	92.1%	15.1%	18.4
	98,435	843	CKD (47)	Pipeline pilot	100/4000	984	1%	393	46.6%	40.1%	46.6
				ECFP4	100/4000	984	1%	593	70.4%	60.6%	70.4
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	384	45.6%	39.2%	45.6
Substance P antagonists	102,514	1146	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	378	33%	7.3%	6.5
	98,435	1146	CKD (47)	Pipeline pilot	100/4000	984	1%	705	61.5%	71.9%	61.5
				ECFP4	100/4000	984	1%	942	82.2%	96.1%	82.2
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	509	44.4%	51.9%	44.4
HIV protease inhibitors	102,514	650	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	377	58%	7.3%	11.5
	98,435	650	CKD (47)	Pipeline pilot	100/4000	984	1%	436	67.1%	44.5%	67.4
				ECFP4	100/4000	984	1%	574	88.3%	58.6%	88.7
			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	355	54.6%	36.2%	54.9
Protein kinase C inhibitors	102,514	353	BKD (60, 61)	Extended connectivity fingerprints	100/400	5125	5%	81	23.1%	1.5%	4.4
	98,435	353	CKD (47)	Pipeline pilot	100/4000	984	1%	238	67.3%	24.2%	67.3
				ECFP4	100/4000	984	1%	291	82.5%	29.7%	82.5

			SVM-RBF (47)	Pipeline pilot	100/4000	984	1%	206	58.3%	21.0%	58.3
MAO inhibitors	101,437	1166	BKD (64)	Atom pairs and topological torsions APTT descriptors	1166/3834	6000	5.9%	600	51.4%	10%	11.5
Muscarinic M1 agonists	98,435	748	CKD (47)	Pipeline pilot	100/4000	984	1%	467	62.4%	47.4%	62.4
				ECFP4	100/4000	984	1%	597	79.8%	60.7%	79.8
NMDA receptor antagonists	98,435	1211	CKD (47)	Pipeline pilot	100/4000	984	1%	604	49.9%	61.4%	49.9
				ECFP4	100/4000	984	1%	889	73.4%	90.3%	73.4
Nitric oxide synthase inhibitors	98,435	277	CKD (47)	Pipeline pilot	100/4000	984	1%	192	69.3%	19.5%	69.7
				ECFP4	100/4000	984	1%	244	88.2%	27.3%	97.6
Aldose reductase inhibitors	98,435	782	CKD (47)	Pipeline pilot	100/4000	984	1%	436	55.8%	44.3%	56.1
				ECFP4	100/4000	984	1%	665	85.0%	67.6%	85.5
Reverse transcriptase inhibitors	98,435	419	CKD (47)	Pipeline pilot	100/4000	984	1%	238	56.9%	24.2%	56.3
				ECFP4	100/4000	984	1%	337	80.4%	34.2%	79.6
Aromatase inhibitors	98,435	413	CKD (47)	Pipeline pilot	100/4000	984	1%	284	68.7%	28.8%	68.6
				ECFP4	100/4000	984	1%	389	94.1%	39.5%	94.0
Phospholipase A2 inhibitors	98,435	604	CKD (47)	Pipeline pilot	100/4000	984	1%	297	49.2%	30.2%	49.5
				ECFP4	100/4000	984	1%	447	74.0%	45.4%	74.5
CDK2 inhibitors	25,300	25	SVM+ BKD (59)	DRAGON descriptors	125/5035	506	2%	18	72%	3.5%	35.4
FXa inhibitors	25,300	25	SVM+ BKD (59)	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	N/A
PDE5 inhibitors	50,000	19	RO5+ DS (65)	Pharmacophore and macroscopic descriptors	130/10K	1821	3.6%	11	57.8%	0.6%	15.8

	25,300	25	SVM+ BKD (59)	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	41.5
Alpha1A AR antagonists	25,300	25	SVM+ BKD (59)	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5

BKD – binary kernel discrimination; **CKD** – Continuous kernel discrimination; **DS** – decision tree; **LMNB** – laplacian modified naive Bayesian; **SVM** – support vector machine; **DRAGON** – (an application for the calculation of molecular descriptors); **AR** – androgen receptor; **PDE 5** – phosphodiesterase type 5; **FXa** – factor Xa; **CDK2** – cyclin-dependent kinase 2; **MAO** – mono amino oxidase; **HIV** – human immunodeficiency virus; **COX** – cycloocygenase;

Table 1-3 Performance of docking methods in virtual screening test for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance; the relevant literature references are given in the method column.

Screening task	**Compounds screened**		**Method and reference of reported study**	**No of pre-docking selected compounds**	**Docking cut-off**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**				**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
Factor Xa inhibitors	2M	630	AUTODOCK + pre-docking RO5 and EA screen (66)	60,000	Binding energy < -10.5 kcal/mol	60,000	3%	392	62%	0.65 %	20
COX2 inhibitors	1.2M	355	DOCK+ pre-docking chemical group screen (67)	13,711	DOCK scores < -35	959	0.08% for all 7% for actually docked	337	95%	35.2 %	1189.2 for all 13.6 for actually docked
Human casein kinase II	400K	>4	DOCK4 + H-bond and hinge segment screen (68)	<400K	N/A	35	0.0087%	4	N/A	11.4 %	N/A
Thyroid hormone receptor antagonists	250K	>14	ICM VLS module (Molsoft) (69) + pre-docking RO5	190K	Selected 75 from top-100 dock scores	75	0.03% for all 0.039% for actually docked	14	N/A	18.7 %	N/A
PTP1B inhibitors	235K	>127	DOCK3.5 + atom count (17~60) screen (70)	165,581	Top-500 + Top-500	889	0.38%	127	N/A	14.3 %	N/A
	141K	10	GOLD + elements and chemical group screen (71)	<141K	Top-2%	<2820	<2.5%	8	80%	<0.28 %	39.4

BCL-2 inhibitors	206,876	>1	DOCK3.5 + non-peptidic screen (72)	<206,876	Top-500	35	0.017%	1	N/A	2.9%	N/A
HIV-1 protease inhibitors	141K	5	GLIDE + elements and chemical group screen (71)	<141K	Top-5%	<7050	<5%	1	20%	<0.014%	4.6
HDM2 inhibitors	141K	14	DOCK + elements and chemical group screen (71)	<141K	Top-5%	<7050	<5%	4	28.6%	<0.056%	5.7
UPA inhibitors	141K	10	GOLD + elements and chemical group screen (71)	<141K	Top-2%	<2820	<2.5%	9	90%	<0.32%	45.1
Alpha 1A adrenergic receptor antagonists	141K	>38	GOLD on homology model + pharmacophore screen (73)	22,950	Top-300	300	0.21%	38	N/A	N/A	N/A
Thrombin inhibitors	141K	10	GLIDE + elements and chemical group screen (71)	<141K	Top-2%	<2820	<2.5%	3	30%	<0.11%	15.5
	133.8K	760	FlexX + Similarity (74)	<133.8K	Top-1%	1338	1%	231	29.3%	17.3%	30.5
DHFR inhibitors	135K	165	DOCK3.5.54 applied to holo form (75)	135K	Top-1% of 50k docked	1350	1%	47	25%	3.4%	27.8
			DOCK3.5.54 applied to appo form (75)	135K	Top-1% of 100k docked	1000	1%	16	9.7%	1.6%	13.1
Neutral endopeptidase inhibitors	135K	356	DOCK3.5.54 (75)	135K	Top-1% of 125.5K	1255	0.74%	3	0.8%	0.24%	~1

					docked						
Thrombin inhibitors	135K	788	DOCK3.5.54 (75)	135K	Top-1% of 121.5K docked	1215	0.9%	61	7.7%	5.0%	8.6
Thymidylate synthase inhibitors	135K	185	DOCK3.5.54 (75)	135K	Top-1% of 54K docked	540	0.4%	49	26.5%	9.1%	66.4
Phospholipase C inhibitors	135K	25	DOCK3.5.54 (75)	135K	Top-1% of 123K docked	1230	0.9%	5	20%	0.4%	21.6
Adenosine kinase inhibitors	135K	356	DOCK3.5.54 applied to holo form (75)	135K	Top-5% of database	4500	3.3%	10	2.8%	0.22%	~1
			DOCK3.5.54 applied to appo form (75)	135K	Top-5% of database	4500	3.3%	5	1.4%	0.11%	<1
	133.8K	59	FlexX + Similarity (74)	<133.8K	Top-1%	1338	1%	13	22%	0.97%	22.0
Acetylcholinesterase inhibitors	135K	637	DOCK3.5.54 applied to holo form (75)	135K	Top-1% of 77K docked	770	0.57%	49	7.7%	6.4%	13.6
			DOCK3.5.54 applied to appo form (75)	135K	Top-1% of 37.5K docked	375	0.28%	25	3.9%	6.7%	14.2
HMG-CoA reductase inhibitors	133.8K	1016	FlexX + Similarity (74)	<133.8K	Top-1%	1338	1%	35	3.4%	2.6%	3.4

Table 1-4 Performance of pharmacophore methods in virtual screening test for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance. The relevant literature references are given in the method column.

Screening task	**Compounds screened**		**Method and reference of reported study**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**		**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
ACE inhibitors	3.8M	55	Pharmacophore (76)	1M	26%	39	70.1%	0.0039%	2.8
	3.8M	55	Structure-based pharmacophore (77)	91K	2.4%	6	10.9%	0.0066%	4.6
11β-hydroxysteroid dehydrogenase 1 inhibitors	1.77M	144	Pharmacophore (41)	20.3K	1.15%	17	11.8%	0.084%	10.3
Rhinovirus 3C protease inhibitors	380K	30	Pharmacophore (42)	6,917	1.82%	23	76.7%	0.33%	41.8

Table 1-5 Performance of clustering methods in virtual screening test for identifying inhibitors, agonists and substrates of proteins of pharmaceutical relevance; the relevant literature references are given in the method column.

Screening task	**Compounds screened**		**Method and reference of reported study**	**Compounds selected**		**Known hits selected**			
	No of compounds	**No of known hits included**		**No of compounds selected**	**Percentage of screened compounds selected**	**No of hits selected**	**Yield**	**Hit rates**	**Enrichment factor**
ACE inhibitors	344.5K	490	Hierachical k-means (40)	5590	1.6%	246	50.2%	4.4%	31.2
			NIPALSTREE (40)	8174	2.4%	188	38.4%	2.3%	16.2
			Hierachical k-means + NIPALSTREE disjunction (40)	12240	3.6%	306	62.4%	2.5%	17.6
			Hierachical k-means + NIPALSTREE conjunction (40)	1662	0.48%	128	26.1%	7.7%	54
COX inhibitors	344.5K	1556	Hierachical k-means (40)	15322	4.4%	761	48.9%	5.0%	11
			NIPALSTREE (40)	22321	6.5%	625	40.2%	2.8%	6.16
			Hierachical k-means + NIPALSTREE disjunction (40)	33793	9.8%	980	63.0%	2.9%	6.42
			Hierachical k-means + NIPALSTREE conjunction (40)	3980	1.2%	406	26.1%	10.2%	22.6
Adrenoceptor ligand	344.5K	542	Hierachical k-means (40)	21285	6.2%	298	55.0%	1.4%	8.99
			NIPALSTREE (40)	28125	8.2%	270	49.8%	0.96%	6.14
			Hierachical k-means + NIPALSTREE disjunction (40)	42365	12.3%	394	72.7%	0.93%	5.93

			Hierachical k-means + NIPALSTREE conjunction (40)	6692	1.9%	174	32.1%	2.6%	16..3
Glucocorticoid receptor ligand	344.5K	91	Hierachical k-means (40)	3750	1.1%	27	29.7%	0.72%	27..3
			NIPALSTREE (40)	3469	1.0%	17	18.7%	0.49%	18.7
			Hierachical k-means + NIPALSTREE disjunction (40)	7317	2.1%	30	33.0%	0.41%	15.6
			Hierachical k-means + NIPALSTREE conjunction (40)	538	0.16%	14	15.4%	2.6%	98
GABA receptor ligand	344.5K	478	Hierachical k-means (40)	10000	2.9%	110	23%	1.1%	7.97
			NIPALSTREE (40)	17143	5.0%	84	17.6%	0.49%	3.51
			Hierachical k-means + NIPALSTREE disjunction (40)	24265	7.0%	165	34.5%	0.68%	4.86
			Hierachical k-means + NIPALSTREE conjunction (40)	2636	0.77%	29	6.1%	1.1%	7.77

1.3 *In-silico* Approaches to Multi-target Drug Discovery

1.3.1 Introduction

Therapeutic agents directed at an individual target frequently show reduced efficacies, undesired safety profiles and drug resistances due to network robustness (78), redundancy (79), crosstalk (80), compensatory and neutralizing actions (81), anti-target and counter-target activities (82), and on-target and off-target toxicities (83). It is being increasingly recognized that a balanced modulation of several targets can provide a superior therapeutic effect and side effect profile compared to the modulation of a single selective ligand. It is not surprising that the search of multi-target agents is constantly attracting the attention of increasing number of drug discoverers (1, 84, 85).

With the extensive exploration of *in sillico* tools in pharmaceutical research, these computational approaches can greatly assist and complement the traditional biological and chemical methods in the discovery of new drug hits and leads. This is especially helpful with the increasingly abundant pharmainformatics data being published and shared across the globe which offers a strong foundation for *in-sillico* approaches.

Some of these methods have recently been applied for searching and designing multi-target agents that are more sparsely distributed in the chemical space than agents against a single target. **Figure 1-3**, **Figure 1-4**, **Figure 1-5**, and **Figure 1-6** summarize the schemes of using molecular docking, combined molecular docking and pharmacophore, framework combination, and fragment-based approaches for multi-target drug discovery using dual-inhibitor discovery as examples. These methods can be categorized into combinatorial approaches and fragment–based approaches. Combinatorial approaches (**Figure 1-3 and Figure 1-4**) are firstly straightforwardly conducted as parallel search against each individual targets and then the virtual hits that interact with individual multiple targets are detected as multi-target agents. Combinatorial approaches are practically useful when the retrieval rates against individual targets are sufficiently high and the false-hit rates are sufficiently low. High retrieval rates can compensate for the reduced collective retrieval rates (if the retrieval rate against individual target is 50%~70%, the collective retrieval rate for multi-target agents against two targets may be statistically reduced to 25%~49%). Low false-hit rates are needed for high enrichment in searching multi-target agents that are significantly fewer in numbers and more sparsely distributed in the chemical space than agents against a single target.

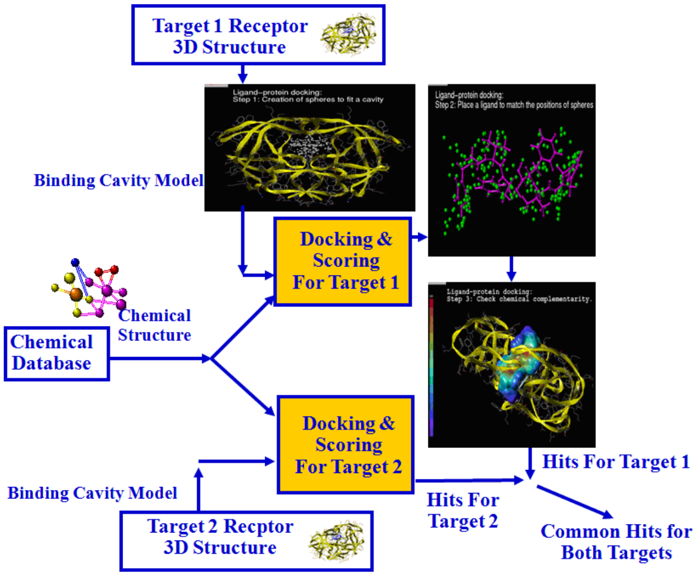


Figure 1-3 Molecular docking strategy for multi-target inhibitor discovery

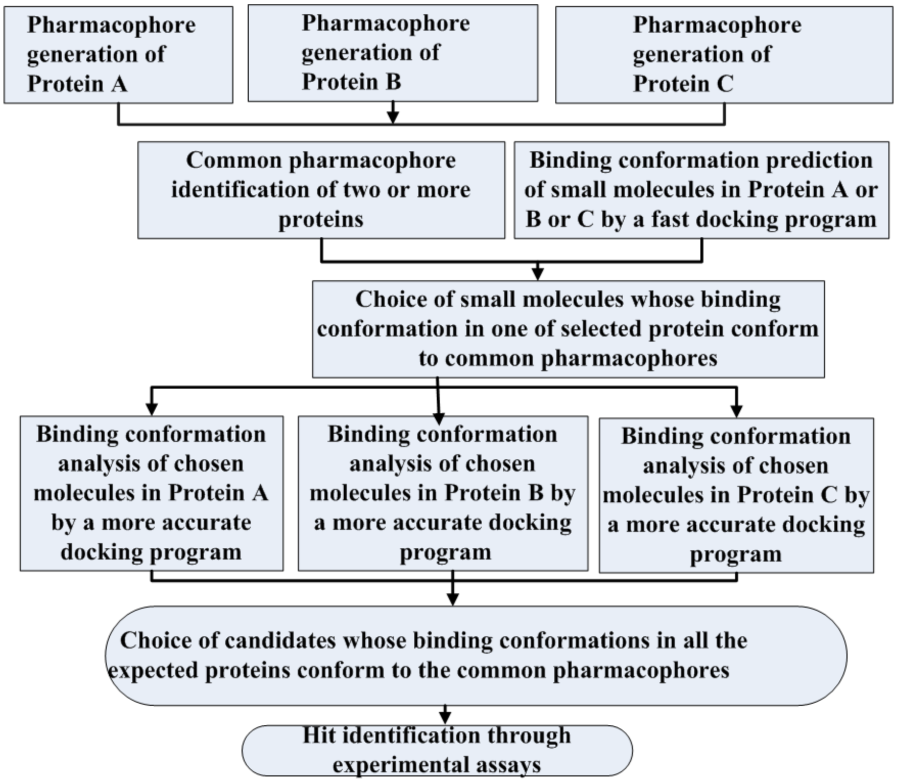


Figure 1-4 Combined pharmacophore and molecular docking strategy of multi-target inhibitor discovery

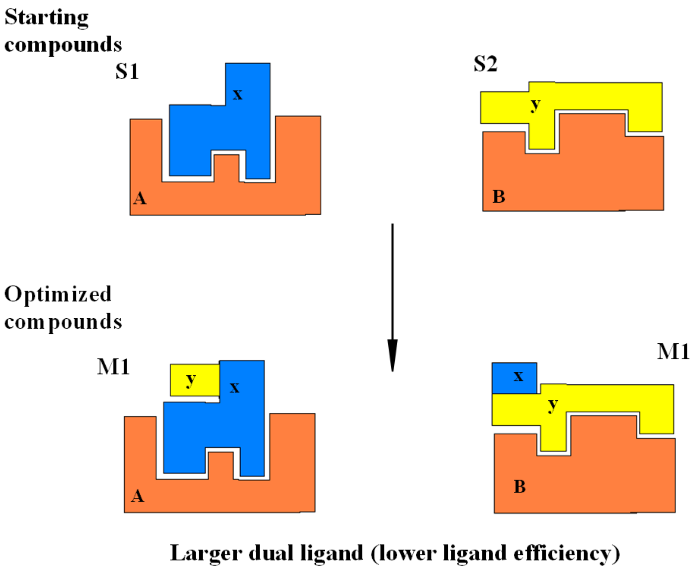


Figure 1-5 Illustration of framework combination approach to multi-target drug discovery

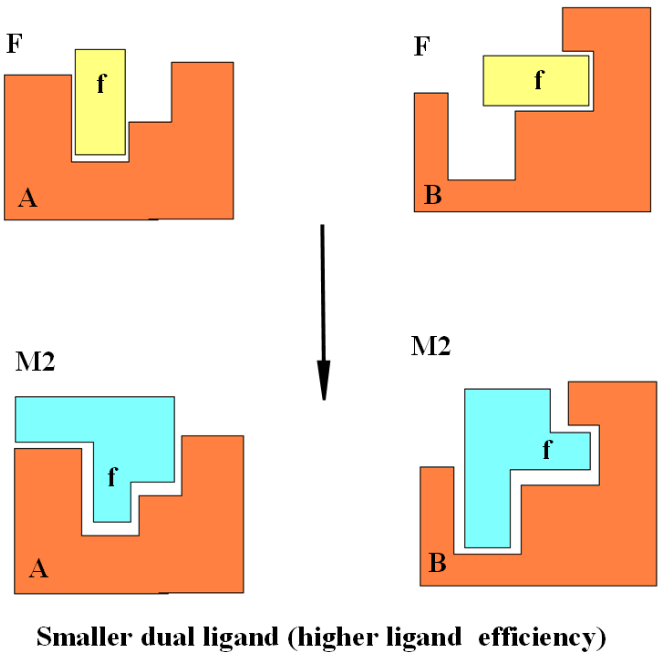


Figure 1-6 Illustration of fragment-based approach to multi-target drug discovery

Fragment-based approaches (**Figure 1-5** and **Figure 1-6**) are designed to combine multiple elements of structural frameworks or multiple fragments that bind to each individual target to produce compounds that bind to multiple targets. They have been introduced as tools for the design of multi-target agents (86). In one approach, molecular fragment libraries are searched to find the fragments with certain level of activities against selected multiple targets, and the identified fragments are further optimized into more potent bigger-sized multi-target active agents (86). The other approach aims at the analysis of the structure-activity relationships against individual targets for the search of molecular fragments and essential binding features which are either combined or incorporated into active agents against selected multiple targets (86). However, fragment combinations often produce larger and non-drug like molecules with more complex structures. Drug-like features may be retained if the degree of framework overlap is maximized and the size of the selected fragments is minimized. Targets sharing a conserved binding makes it relatively more easily for the optimizing fragments with weak multiple activities into potent multi-target drug-like agents (87). But with the increased similarities among binding sites, it is becoming more difficult to improve and adequately balance the high binding affinities for acceptable *in-vivo* efficacy and safety. One way to resolve this problem relies on synergistic targets for their modest activities at one or more of the relevant targets. This may still produce similar or better *in-vivo* effects compared with higher-affinity target-selective compounds (88).

Multi-target QSAR models for identification of multi-target agents (89) and active agents against multiple bacterial (90), fungal (91, 92), and viral (90), species have been developed. These models incorporate multi-target or species variations of binding-site features into the multi-target dependent molecular descriptors or species-dependent molecular descriptors, and stochastic Markov drug-binding process models. They can achieve high retrieval rates of 72%~85% and moderately low false-hit rates of 15%~28%. However, the application of multi-target QSAR models may be limited by the inadequate number of drug data for some of the targets or species. Moreover, the molecular size of the testing drugs is limited within a certain range for accurate computation of multi-target dependent or species-dependent molecular descriptors. This in some cases may also affect the development of multi-target QSAR models (92).

1.3.2 Machine learning methods for searching multi-target agents

Cancer is known to be a fatal disorder which has been threatening lives of millions of people per year. In the last decades , untangling the molecular mechanisms underlying malignant transformation have been the center of efforts in basic and clinical cancer research to discover molecules that play a crucial and specific role in tumor progression (93). Protein kinases play important roles in regulating most cellular functions: proliferation/cell cycle, cell metabolism, survival/apoptosis, DNA damage repair, cell motility, response to the microenvironment. Therefore, it is no surprise that they are often themselves oncogenes. Kinases take the second most popular drug target class in the pharmaceutical and biotech industries, after G-protein-coupled receptors (94). However, clinical experience confronts us with the fact that many tumors are multi-factorial and interlinked by more than one signaling pathway which makes the inhibition of a single molecule not sufficient to interfere efficiently with disease progression. For these reasons, monotherapy by means of single-target drugs may need to be reassessed in favor of a multi- target approach. In this work, I examined a SVM based combinatorial machine learning method COMBI-SVM for its performance for detecting multi-target kinase inhibitors for 4 kinase target pairs, i.e., EGFR-FGFR,VEGFR-Lck, and Src-Lck. This is a preliminary test for the performances of COMBI-SVM in searching multi-target agents.

Major depression is an enervating and recurrent disorder. It has become prevailing due to the fastened pace and enhanced stress levels in the modern societies. It affects patients with a substantial lifetime risk. A primary anti-depression strategy is to inhibit monoamine oxidase; Second-generation drugs launched in the 1980s and 1990s, such as the selective serotonin reuptake inhibitors (SSRI) and the mixed serotonin/noradrenaline reuptake inhibitors (SNRI). They present the dominant treatment strategy for major depression (95). However, single-target drugs (78, 85) frequently encounter the drug resistance problems caused by the network robustness (78), redundancy (79), crosstalk (80), compensatory and neutralizing actions (81), anti-target and counter-target activities (82), and on-target and off-target toxicities (83). Multi-target drugs are particularly useful for solving these drug resistance problems. After the previous performance tests of COMBI-SVM in dual kinase inhibitors, I applied this approach to study the dual inhibitor SSRIs: SNRIs, dual serotonin reuptake inhibitor (SRI) / 5-HT$_{1A}$ receptor antagonists, dual SRI/ 5HT$_{1B}$ receptor antagonists, dual SRI/ H3 histamine receptor antagonists, dual SRI/5-HT$_{2C}$ receptor antagonists, dual SRI/Melanocortin 4 receptor antagonists, dual SRI/neurokinin 1 receptor antagonists. **Figure 1-7** shows the work flow for detecting multi-target agents by machine learning (ML) methods.

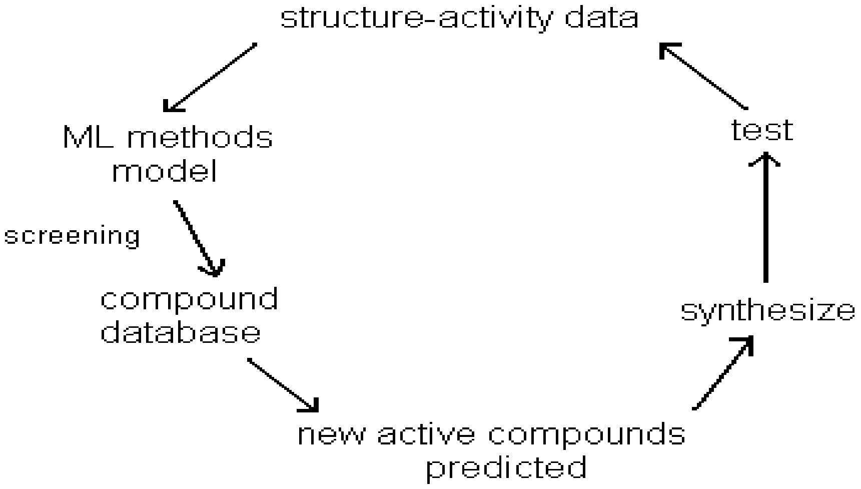


Figure 1-7 Work flow for detecting multi-target agents by machine learning (ML) methods; Structure-activity data are collected by literature mining. Then the ML method is applied to build a screening model which will be used to scan the compound database (e.g. PubChem); After the screening, positive dual-inhibitors will be selected for further synthesis and test. If they prove to have promising

pharmacological profiles, they can be used into the training data for new predictions.

1.4 Objectives and Outline

Over all, I want to achieve four major objectives.

1. To update and construct pharmainformatics databases to provide resourceful and informative platforms for researchers in various bio-and chemo-informatics.
2. To test combinatorial machine learning methods for virtual screening of multi-target agents for cancer treatment involving kinase targets EGFR, FGFR, Src, Lck and VEGFR.
3. To apply combinatorial machine learning methods for virtual screening of Selective multi-target serotonin reuptake inhibitors.
4. To compare the virtual screening performances of the machine learning methods SVM, k-NN, PNN and similarity searching in search of multi-target agents.

In summary, this work aims at contributing to the current multi-target strategy in novel drug hits and leads discovery. This study employs two approaches to reach this goal. The first approach targets at optimizing the benefit of the increasingly abundant pharmaceutical data information. Pharmainformatics database is an efficient and resourceful means to achieve this goal. They dramatically accelerate the accumulation of data thus enhance the opportunity in new discoveries. A collective pharmaceutical data and discoveries have been presented online in the past ten years. These valuable data would benefit the discovery of multi-target agents. Additionally, combinatorial virtual screening methods can also assist the discovery of novel multi-target anticancer and antidepressant agents.

This work is presented in the following manner. Chapter 1 firstly introduces the development of pharmainformatic databases and the different approaches in virtual screening. Secondly, it describes the ML approaches for multi-target discovery.

In Chapter 2, methods used in this work are described. It introduces the data collection and processing before the application of VS tools. Theoretical backgrounds of machine learning methods discussed in the work are provided. VS model validation and performance measurements are described in details.

Chapter 3 elaborates the updates and development work of two pharmainformatics databases, i.e. Kinetic database of bio-molecular interactions (KDBI) and Therapeutic targets database (TTD).

Chapter 4 describes the preliminary tests with 4 kinase target pairs of combinatorial SVM (COMB-SVM) and Chapter 5 elaborates the studies and application of COMBI-SVM as virtual screening tools for multi-target antidepressants agents.

At last, Chapter 6 summarizes the findings of this work and discusses the limitations.

Chapter 2 Methods

Virtual screening for multi-target agents by combinatorial machine learning methods is usually consisted of the following 4 components: (I) data collection, analysis and processing from pharmaceutical datasets and chemical compound libraries of known single and multi-target agents (section 2.1), (II) physicochemical and structural descriptions of the compounds in the dataset (section 2.2) and (III) a statistical learning method (section 2.3 and 2.4) to analyze the pharmaceutical datasets (component (I)) in the form of descriptors (component (II)), and (IV)the evaluation of the virtual screening models. This chapter describes in details the four components and elaborates all the methods used in this work for developing combinatorial virtual screening tools and their evaluation measurements.

2.1 Data Collection and Processing

Sufficient, good quality data are critical for drug discovery and especially essential for *in-silico* approaches which rely on the quantity and quality of the available data. Enormous amount of data about small molecules and their related information have been accumulated in various scientific literatures and databases. **Table 2-1** lists some of the important small molecule databases.

The datasets used in this work mainly come from the following two types of sources. We collected data from credible journals such as Bioorganic & Medicinal

Chemistry Letters, Bioorganic & Medicinal Chemistry, European Journal of Medicinal Chemistry, European Journal of Organic Chemistry and Journal of Medicinal Chemistry, etc. Additionally, I use databases that contain accurate and reliable data such as PubChem and ChEMBL (96).

Table 2-1 Examples of small molecule databases available online

Database Name	**URL**
BindingDB	http://www.bindingdb.org/bind/index.jsp
MDDR	http://www.symyx.com/products/databases/bioactivity/mddr/index.jsp
PubChem	http://nihroadmap.nih.gov
ZINC	http://zinc.docking.org/
ChEMBL	http://www.ebi.ac.uk/chembl/
DrugBank	http://www.drugbank.ca/
eMolecules	http://www.emolecules.com/
WOMBAT	http://www.sunsetmolecular.com

2.1.1 Analysis of data quality and diversity

The reliability of *in-silico* approaches of pharmacological properties classification depends on the availability of high quality pharmacological data with low experimental errors (97). Ideally, the measurements of pharmacological data properties should be conducted with a same protocol so that there is a common ground to compare different compounds with each other. However, some pharmacological properties measurements have been used only for a limited number of compounds and most pharmacological properties measurements are rarely determined by the same protocol. Thus the collected data consist of compound data measured by different protocols and the incorporation of additional experimental information. To maintain the stability of data quality, in this work, several methods are adopted to ensure that inter-laboratory variations caused by different experimental protocols do not significantly affect the quality of the training sets. The pharmacological property measurements for data were investigated and the ones that contain large variations in experimental protocols compared to the majority of the data are filtered out. It is estimated that the most common range of the pharmacological properties measurements for the compounds investigated in more than one source was used to select compounds for the different classes (98).

Diversity Index (DI) is employed to evaluate the structural diversity of a collection of compounds. It is defined as the average value of the similarity between pairs of compounds in a dataset (99),

$$DI = \frac{\sum_{i,j \in D \wedge i \neq j} sim(i,j)}{|D|\left(|D|-1\right)} \tag{1}$$

where $sim(i,j)$ is a measure of similarity between compounds i and j, D is the dataset and |D| is set cardinality (number of elements of the set). The dataset is more diverse when DI approaches 0. Tanimoto coefficient (34) was used to compute $sim(i,j)$ in this study,

$$sim(i,j) = \frac{\sum_{d=1}^{k} x_{d_i} x_{d_j}}{\sum_{d=1}^{k} \left(x_{d_i}\right)^2 + \sum_{d=1}^{k} \left(x_{dj}\right)^2 - \sum_{d=1}^{k} x_{d_i} x_{d_j}} \tag{2}$$

where k is the number of molecular descriptors calculated for the compounds in the datasets. The concept of molecular descriptors will be introduced in chapter 2.2.

2.1.2 Redundancy within the datasets

In this study, the data were collected from varied sources. This approach can enrich the diversity in the datasets and reduce the potential bias that may arise from a monotonic due to the preferences of the researchers. However, since the data are presented by independent researchers who don't share pre-existing agreement on their individual data collection. It is likely that there is a certain level of redundancy between the datasets from different sources. The redundancy could contrarily deduce diversity in the datasets. Therefore, compounds are checked for redundancy by comparing exact match of chemical descriptors. In this work, scripts are written in Perl to find exact match of chemical descriptors to remove redundancy from dataset.

2.2 Molecular descriptors

Molecular descriptors are generated by a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment. They quantitatively represent structural and physicochemical features of molecules, and have been extensively used in deriving structure-activity relationships (100), quantitative structure activity relationships (101) and VS tools (102, 103) (104) including the multi-target VS tools (105). They represent compounds in the form of mathematical vectors. This transformation enables the statistical analysis of chemical compounds.

2.2.1 Definition and calculation of molecular descriptors

A number of programs e.g. PaDEL-descriptor (106), DRAGON (107), Molconn-Z (108), MODEL (109), Chemistry Development Kit (CDK) (110, 111), JOELib (112) and Xue descriptor set (113), are available to calculate chemical descriptors. These methods can be applied to derive >3,000 molecular descriptors. These descriptors include constitutional descriptors, topological descriptors, RDF descriptors (114), molecular walk counts (115), 3D-MoRSE descriptors (116), BCUT descriptors (117), WHIM descriptors (118), Galvez topological charge indices and charge descriptors (119), GETAWAY descriptors (120), 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices (121), Randic molecular profiles (122), electrotopological state descriptors (123), linear solvation energy relationship descriptors (124), and other empirical and molecular properties. However, not all of the available descriptors are needed to fully represent the features of a particular class of compounds. Contrarily, without appropriate descriptors, the performance of a developed ML VS tool may be affected to some degrees. This is caused by the noise arising from the high redundancy and overlapping of the available descriptors. In this work, the Xue descriptor set and 98 1D and 2D descriptors were used. These 98 descriptors were selected from the descriptors derived from MODEL program by discarding those that were redundant and unrelated to the problem studied here. The Xue descriptor set and these 98 descriptors are listed in **Table 2-2** and **Table 2-3**.

Table 2-2 Xue descriptor set

Descriptor Class	**Number of descriptor in class**	**Descriptors**
Simple molecular properties	18	Molecular weight, Number of rings, rotatable bonds, H-bond donors, and H-bond acceptors, Element counts
Molecular connectivity and shape	28	Molecular connectivity indices, Valence molecular connectivity indices, Molecular shape Kappa indices, Kappa alpha indices, flexibility index
Electro-topological state	97	Electrotopological state indices, and Atom type electrotopological state indices, Weiner Index, Centric Index, Altenburg Index, Balaban Index, Harary Number, Schultz Index, PetitJohn R2 Index, PetitJohn D2 Index, Mean Distance Index, PetitJohn I2 Index, Information Weiner, Balaban RMSD Index, Graph Distance Index

Quantum chemical properties	31	Polarizability index, Hydrogen bond acceptor basicity (covalent HBAB), Hydrogen bond donor acidity (covalent HBDA), Molecular dipole moment, Absolute hardness, Softness, Ionization potential, Electron affinity, Chemical potential, Electronegativity index, Electrophilicity index, Most positive charge on H, C, N, O atoms, Most negative charge on H, C, N, O atoms, Most positive and negative charge in a molecule, Sum of squares of charges on H,C,N,O and all atoms, Mean of positive charges, Mean of negative charges, Mean absolute charge, Relative positive charge, Relative negative charge
Geometrical properties	25	Length vectors (longest distance, longest third atom, 4th atom), Molecular van der Waals volume, Solvent accessible surface area, Molecular surface area, van der Waals surface area, Polar molecular surface area, Sum of solvent accessible surface areas of positively charged atoms, Sum of solvent accessible surface areas of negatively charged atoms, Sum of charge weighted solvent accessible surface areas of positively charged atoms, Sum of charge weighted solvent accessible surface areas of negatively charged atoms, Sum of van der Waals surface areas of positively charged atoms, Sum of van der Waals surface areas of negatively charged atoms, Sum of charge weighted van der Waals surface areas of positively charged atoms, Sum of charge weighted van der Waals surface areas of negatively charged atoms, Molecular rugosity, Molecular globularity, Hydrophilic region, Hydrophobic region, Capacity factor, Hydrophilic-Hydrophobic balance, Hydrophilic Entry Moment, Hydrophobic Intery Moment, Amphiphilic Moment

Table 2-3 98 molecular descriptors used in this work

Descriptor Class	No of Descriptors in Class	Descriptors
Simple molecular properties	18	Number of C,N,O,P,S, Number of total atoms, Number of rings, Number of bonds, Number of non-H bonds, Molecular weight,, Number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Number of 5-member aromatic rings, Number of 6-member aromatic rings, Number of N heterocyclic rings, Number of O heterocyclic rings, Number of S heterocyclic rings.
Chemical properties	3	Sanderson electronegativity, Molecular polarizability, ALogp
Molecular Connectivity and shape	35	Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, Gravitational topological index, Molecular path count of length 1-6, Total path count, Balaban Index J, 0-2th valence connectivity index, 0-2th order delta chi index, Pogliani index, 0-2th Solvation connectivity index, 1-3th order Kier shape index, 1-3th order Kappa alpha shape index, Kier Molecular Flexibility Index, Topological radius, Graph-theoretical shape coefficient, Eccentricity, Centralization, Logp from connectivity.
Electro-topological state	42	Sum of E-state of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; Sum of E-state of all heavy atoms, all C atoms, all hetero atoms, Sum of E-state of H-bond acceptors, Sum of H E-state of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin, Sum of H E-state of H-bond donors

In my work, the 2D structure of each of the compounds was generated by using ChemDraw or downloaded from databases such as PubChem and BindingDB (125). Then they were subsequently converted into 3D structure by using CORINA (126). The 3D structure of each compound was manually inspected to ensure the proper chirality of each chiral agent. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation. The optimization of generated geometries was conducted without symmetry restrictions. The 3D structures of the compounds then were used to compute the molecular descriptors by the in-house programs and scripts.

2.2.2 Scaling of molecular descriptors

The scaling of molecular descriptors is normally required before they can be used in machine learning method. The scaling process of molecular descriptors ensures the unbiased contribution of each descriptor in constructing the prediction models (127). There are various types of scaling methods e.g. auto-scaling, range scaling, Pareto scaling (128), and feature weighting (127). In this work, range scaling is applied to scale the molecular descriptors. Range scaling is conducted by dividing the difference between the descriptor value and the minimum value of that descriptor with the range of that descriptor:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}}$$

2.3 Introduction to Machine Learning Methods

A machine learning (ML) method takes a training set of objects that have previously been classified into two or more classes as input. In the pharmainformatics context, it is a set of molecules that had previously been tested and shown to be either active or inactive. The training samples are represented by vectors which can be binary, categorical or continuous and then they are analyzed to develop a decision rule that can be used to classify new molecules (the test set) into one of the two classes (129). Machine learning can be divided into supervised and unsupervised categories. Supervised machine learning labels the training data as a predefined class (130). Example of supervised machine learning includes Support Vector Machine, Artificial Neural Network, Decision tree learning, Inductive logic programming, Boosting, Gaussian process regression etc. Unsupervised machine learning methods use unlabeled training data and the learning task involves finding the organization of data (131). Clustering, Adaptive Resonance Theory, and Self Organized Map are some of the commonly applied unsupervised machine learning methods. In this work, I used SVM, Probabilistic Neural Network (PNN) and k nearest neighbor (k-NN). The theories behind these methods are described below in subsequent sections. For a comparative study, Tanimoto similarity searching method was also introduced.

2.3.1 Support vector machine (SVM) method

Support vector machine (SVM) is designed on the basis of the structural risk minimization principle of statistical learning theory (132, 133). It consistently shows outstanding classification performance and is less penalized by sample redundancy. SVM also has lower risk for over-fitting problem (134, 135).

In linearly separable cases, SVM constructs a hyper-plane to separate active and inactive classes of compounds with a maximum margin. A compound is represented by a vector $\boldsymbol{x}_i$ composed of its molecular descriptors. The hyper-plane is constructed by finding another vector **w** and a parameter b that minimizes $\|\mathbf{w}\|^2$ and satisfies the following conditions:

[illegible] Class 1 (active) (1)

[illegible] Class 2 (inactive) (2)

where y_i is the class index, **w** is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin and $\|\mathbf{w}\|^2$ is the Euclidean norm of **w**. Based on **w** and b, a given vector $\boldsymbol{x}$ can be classified by $f(\boldsymbol{x})$ = [illegible]. A positive or negative $f(\boldsymbol{x})$ value indicates that the vector **x** belongs to the active or inactive class respectively. Linear SVM can then be applied to this feature space based on the following decision function:

$f(\mathbf{x}) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b)$, where the coefficients α_i^0 and b are determined by maximizing the following Lagrangian expression:

$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ under the conditions $\alpha_i \geq 0$ and $\sum_{i=1}^{l}\alpha_i y_i = 0$. A positive or negative *f(**x**)* value indicates that the vector ***x*** belongs to the active or inactive class respectively. However, in classifying compounds of diverse structures, nonlinearly separable situation is frequently found to occur (59, 60, 62, 136-139). In this case, SVM maps the input vectors into a higher dimensional feature space by using a kernel function K(**x**$_i$, **x**$_j$). We used RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2/2\sigma^2}$ where σ is the kernel parameter. RBF kernel has been extensively used and consistently shown better performance than other kernel functions (35, 140, 141). For a given training set of instance-label pairs (x_i, y_i), i=1, …,l where $x_i \in R^n$ and $y_i \in \{ 1, -1 \}$in^l , in SVM, the task of finding the hyper-plane which is able to separate active and inactive classes with a maximum margin ,in essence, is to look for the solution of the following optimization problem:

$$\min_{w,b,\xi}\frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i$$

$$\textbf{subject to} \quad y_i(w^T\Phi(x_i)) + b \geq 1 - \xi_i,$$
$$\xi_i \geq 0.$$

C>0 is the penalty parameter of the error term. The process of training and using a SVM VS model for screening compounds based on their molecular descriptors is schematically illustrated in **Figure 2-1**.

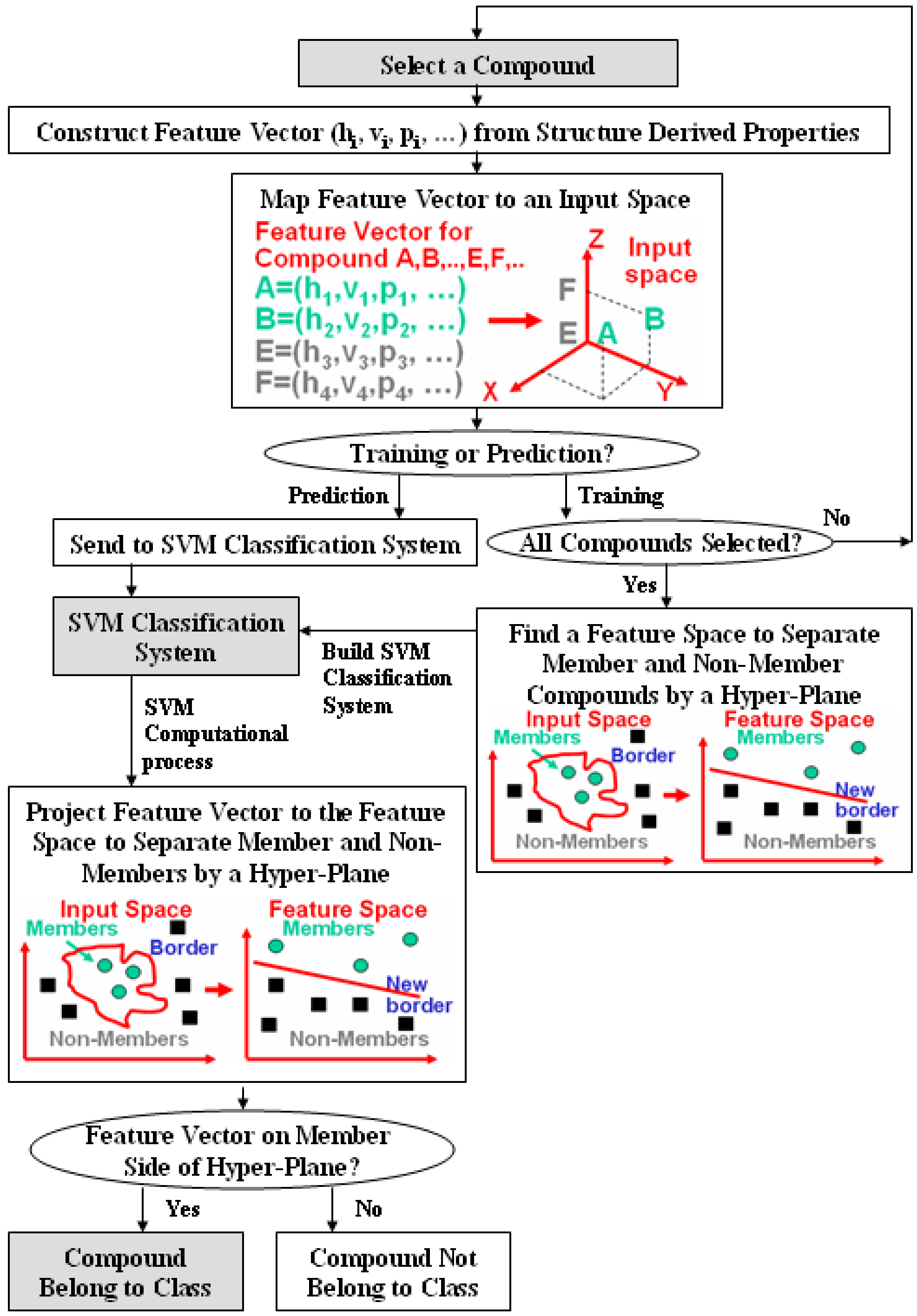

Select a Compound
Construct Feature Vector (h_i, v_i, p_i, ...) from Structure Derived Properties
Map Feature Vector to an Input Space
Feature Vector for Compound A,B,..,E,F,..
$A=(h_1,v_1,p_1, \ldots)$
$B=(h_2,v_2,p_2, \ldots)$
$E=(h_3,v_3,p_3, \ldots)$
$F=(h_4,v_4,p_4, \ldots)$
Z
Input space
F
E
A
B
X
Y
Training or Prediction?
Prediction
Training
Send to SVM Classification System
All Compounds Selected?
No
Yes
SVM Classification System
Build SVM Classification System
Find a Feature Space to Separate Member and Non-Member Compounds by a Hyper-Plane
Input Space
Members
Border
Non-Members
Feature Space
Members
New border
Non-Members
SVM Computational process
Project Feature Vector to the Feature Space to Separate Member and Non-Members by a Hyper-Plane
Input Space
Members
Border
Non-Members
Feature Space
Members
New border
Non-Members
Feature Vector on Member Side of Hyper-Plane?
Yes
No
Compound Belong to Class
Compound Not Belong to Class

Figure 2-1 Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines; A, B, E, F and (hj, pj, vj,…) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

2.3.2 K-nearest neighbor method (k-NN)

k-NN measures the Euclidean distance $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$ between a compound ***x*** and each individual compound $\boldsymbol{x}_i$ in the training set(142, 143). A total of *k* number of vectors nearest to the vector ***x*** are used to determine the decision function *f(**x**)*:

$$\hat{f}(\mathbf{x}) \leftarrow \arg\max_{v \in V} \sum_{i=1}^{k} \delta(v, f(\mathbf{x}_i)) \qquad (6)$$

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, arg max refers to the maximum value of the function, *V* is a finite set of vectors $\{\text{v1},\ldots,\text{vs}\}$ and $\hat{f}(\mathbf{x})$ is an estimate of *f(**x**)*. Here estimate refers to the class of the majority compound group (i.e. inhibitors or non-inhibitors) of the *k* nearest neighbors. The procedure of k-NN is illustrated in **Figure 2-2**.

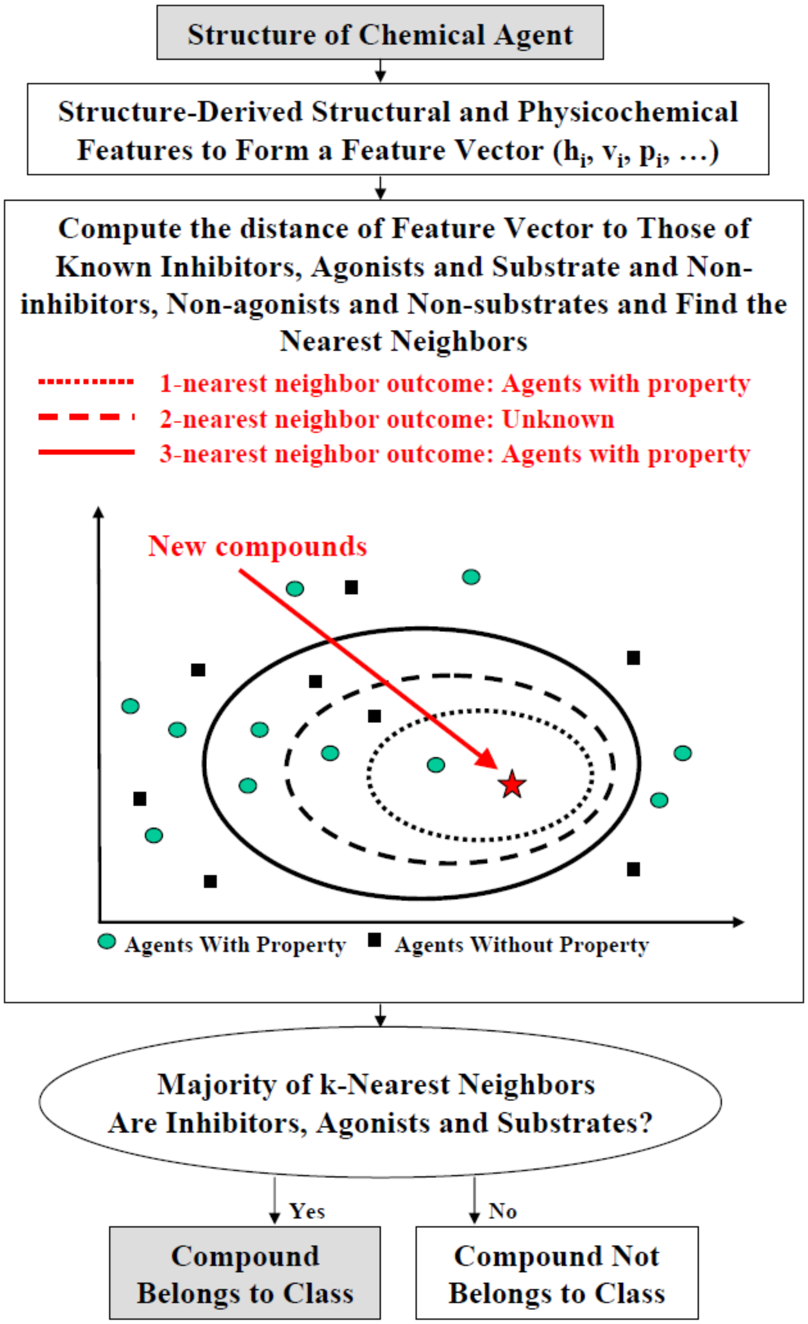

Structure of Chemical Agent
Structure-Derived Structural and Physicochemical Features to Form a Feature Vector (h_i, v_i, p_i, …)
Compute the distance of Feature Vector to Those of Known Inhibitors, Agonists and Substrate and Non-inhibitors, Non-agonists and Non-substrates and Find the Nearest Neighbors
1-nearest neighbor outcome: Agents with property
2-nearest neighbor outcome: Unknown
3-nearest neighbor outcome: Agents with property
New compounds
Agents With Property
Agents Without Property
Majority of k-Nearest Neighbors Are Inhibitors, Agonists and Substrates?
Yes
No
Compound Belongs to Class
Compound Not Belongs to Class

Figure 2-2 Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using a machine learning method – k-nearest neighbors (k-NN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property; feature vector (hj, pj, vj,…) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

2.3.3 Probabilistic neural network method

Probabilistic Neural Network (PNN) belongs to the neural network methods. It is designed for classification through the use of Bayes' optimal decision rule (98): $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$, where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification and $f_i(x)$ and $f_j(x)$ are the probability density function for class i and j respectively. An unclassified vector $\boldsymbol{x}$ is classified into population i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator(144),

$$g(\mathbf{x}) = \frac{1}{n\sigma}\sum_{i=1}^{n} W(\frac{\mathbf{x}-\mathbf{x}_i}{\sigma}) \tag{7}$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\boldsymbol{x} - \boldsymbol{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos for the multivariate case.

$$g(x_1,\ldots,x_p) = \frac{1}{n\sigma_1\ldots\sigma_p}\sum_{i=1}^{n} W(\frac{x_1 - x_{1,i}}{\sigma_1},\ldots,\frac{x_p - x_{p,i}}{\sigma_p}) \tag{8}$$

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated and satisfies the conditions required by Parzen's estimator. Thus the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\exp(-\sum_{j=1}^{p}\left(\frac{x_j - x_{ij}}{\sigma_j}\right)^2) \qquad (9)$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. PNN are constituted of four layers, the input layer, the pattern layer, the summation layer and the output layer. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparametric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. Finally, the single neuron in the output layer then estimates the class of the unknown vector ***x*** by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function. **Figure 2-3** illustrates the procedure of PNN method.

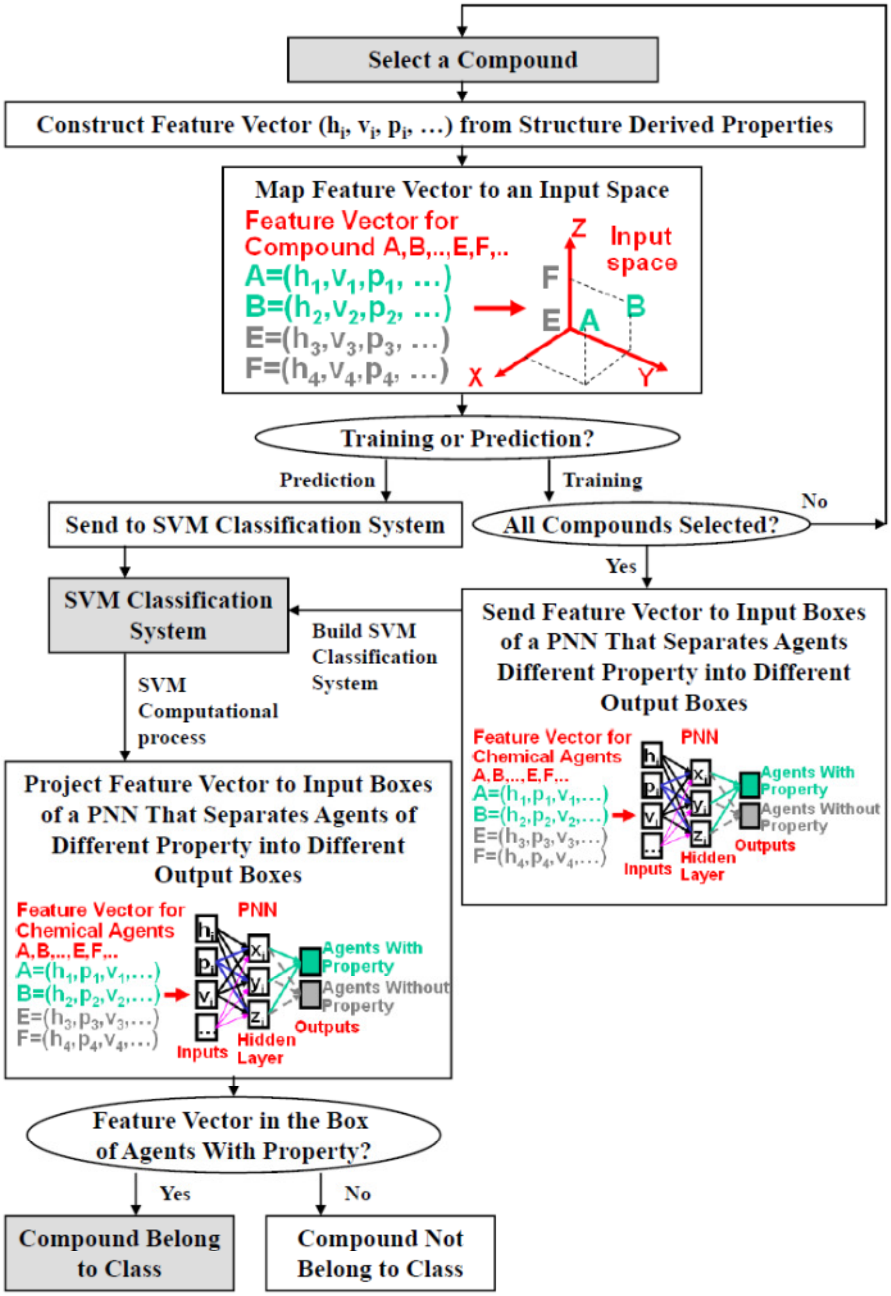

Select a Compound
Construct Feature Vector (h_i, v_i, p_i, …) from Structure Derived Properties
Map Feature Vector to an Input Space
Feature Vector for Compound A,B,..,E,F,..
A=(h_1,v_1,p_1, …)
B=(h_2,v_2,p_2, …)
E=(h_3,v_3,p_3, …)
F=(h_4,v_4,p_4, …)
Input space
Z
F
E
A
B
X
Y
Training or Prediction?
Prediction
Training
Send to SVM Classification System
All Compounds Selected?
No
Yes
SVM Classification System
Build SVM Classification System
Send Feature Vector to Input Boxes of a PNN That Separates Agents Different Property into Different Output Boxes
Feature Vector for Chemical Agents A,B,..,E,F,..
A=(h_1,p_1,v_1,…)
B=(h_2,p_2,v_2,…)
E=(h_3,p_3,v_3,…)
F=(h_4,p_4,v_4,…)
PNN
h_i
p_i
v_i
..
x_i
y_i
z_i
Inputs
Hidden Layer
Outputs
Agents With Property
Agents Without Property
SVM Computational process
Project Feature Vector to Input Boxes of a PNN That Separates Agents of Different Property into Different Output Boxes
Feature Vector for Chemical Agents A,B,..,E,F,..
A=(h_1,p_1,v_1,…)
B=(h_2,p_2,v_2,…)
E=(h_3,p_3,v_3,…)
F=(h_4,p_4,v_4,…)
PNN
h_i
p_i
v_i
..
x_i
y_i
z_i
Inputs
Hidden Layer
Outputs
Agents With Property
Agents Without Property
Feature Vector in the Box of Agents With Property?
Yes
No
Compound Belong to Class
Compound Not Belong to Class

Figure 2-3 Schematic diagram illustrating the process of the prediction of the prediction of compounds of a particular property from their structure by using a machine learning method –probabilistic neural networks (PNN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the

property; feature vector (h_j, p_j, v_j,…) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

2.3.4 Tanimoto similarity searching method

Determining if two compounds are similar to each other or not in a training dataset can be conducted by using the Tanimoto coefficient *sim(i,j)* (34)

$$sim(i,j) = \frac{\sum_{d=1}^{l} x_{di}x_{dj}}{\sum_{d=1}^{l}(x_{di})^2 + \sum_{d=1}^{l}(x_{dj})^2 - \sum_{d=1}^{l} x_{di}x_{dj}} \qquad (10)$$

where *l* is the number of molecular descriptors. A compound *i* is considered to be similar to a known active j in the active dataset if the corresponding *sim(i,j)* value is greater than a cut-off value. In this work, in computing *sim(i,j)*, the molecular descriptor vectors $\mathbf{x}_i$s were scaled with respect to all of the MDDR. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9 (145, 146). A stricter cut-off value of 0.9 was used in this work.

2.3.5 Generation of putative inactive compounds

The construction of machine learning prediction models requires both positive data (e.g. active compounds) and negative data (e.g. inactive compounds). Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds (47, 58-60, 62-64, 136)**,** a new approach extensively used for generating inactive proteins in ML classification of various classes of proteins (147-149) is considered to be used for generating putative inactive compounds. An advantage of this approach lies in its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which can expend the coverage of the "inactive" chemical space when only limited knowledge of inactive compounds and compounds of other biological classes can be found. A drawback of this approach is the possible inclusion of some undiscovered active compounds in the "inactive" class, which may affect the capability of ML methods for identifying novel active compounds. Such an adverse effect is expected to be relatively small for many biological target classes as explained below.

In applying this approach to proteins, all known proteins are clustered into ~8,900 protein families based on the clustering of their amino acid sequences (112), and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. Undiscovered active proteins of a specific functional class typically cover no more than a few hundred families, which gives a maximum possible "wrong" family representation rate of <10% even when all of the undiscovered active proteins are misplaced into the inactive class (150). Importantly, the inclusion of the representative of a "wrong" family into the inactive class does not preclude other active family members from being classified as active. Statistically, a substantial percentage of active members can be classified by ML methods as active even if its family representative is in the inactive class (150). Therefore, in principle, a reasonably good ML model can be derived from these putative inactive samples, which has been confirmed by a number of studies (147-150).

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors (40, 151). As ML methods predict compound activities based on their molecular descriptors, it is reasonable to represent compounds in terms of molecular descriptors and cluster them in the similar manner. We applied K-means classification method (40, 151) and used molecular descriptors computed from our own software to represent the compounds (152). Then 7,990 cluster families were generated from the available compounds in PubChem database, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms (8), and the 2,851 clusters for 171,045 natural products (153). Analogue groups such as steroids and catecholamines are distributed in a few families. Active compounds in extensively studied target classes such as those of HIV-1 protease inhibitors, DHFR inhibitors, and dopamine antagonists are distributed in 770, 135, and 799 families respectively. The number of undiscovered "active" families in PubChem database is expected to be relatively small after then extensive effort in searching the known compound libraries for identifying active compounds in these target classes, most likely no more than several hundred families. The ratio of the undiscovered "active" families (hundreds on less) and the families that contain no known active compound (6,000~7,000 based on current version of

PubChem) for these and possibly many other target classes is expected to be <15%. Therefore, it is reasonable to generate putative inactive compounds by extracting a few representative compounds of those families that contain no known active compound, with a maximum possible “wrong” family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class.

CNS active agents are distributed in numerous biological target classes such as agonists, antagonists, regulators of G-protein coupled receptors and nuclear receptors, blockers and regulators of ion channels, substrates, inhibitors, activators, and regulators of transporters, and inhibitors and regulators of enzymes involved in the synthesis and metabolism of signaling molecules in the CNS system (154). Therefore, agents in this multi-target class are expected to cover a significantly larger portion of the chemical space than those of a single target class, leading to a possibly higher “wrong” family representation rate because of the possibility of higher number of undiscovered active families in the limited chemical space covered by the currently available compounds in existing databases. As a result, the quality of the putative non-CNS active compounds generated by the new approach may be affected to some extent. The new approach is expected to become more and more useful for multi-target classes when the coverage of chemical space can be significantly expanded as a result of increasing volume of the chemical databases.

There are 7,220, 7,855, 7,191, 3,440 families that contain no known HIV-1 protease inhibitor, DHFR inhibitor, dopamine antagonist, and CNS active agent respectively. Thus datasets of 41,254 putative non- HIV-1 protease inhibitors, 44,856 putative non-DHFR inhibitors, 42,804 putative non-dopamine antagonists, and 20,465 putative non-CAN active compounds were generated by random selection of 5~6 representative compounds from each of these families respectively.

2.4 Virtual Screening Model Validation and Performance Measurements

2.4.1 Model validation

In-silico modeling offers the prediction of the pharmacological properties of compounds which have not been clinically or biologically tested. Therefore it is important to estimate and validate the predicting ability of the pharmacological-data-derived models by their performances with the compounds that are not present in the training set. In this work, I used 5-fold cross-validation and independent validation datasets for this purpose. In 5-fold cross-validation, compounds are randomly divided into five subsets of approximately equal size. Four subsets are used as the training set for developing a model; the remaining one is used as a testing set for evaluating the prediction performance of the model. This procedure is repeated five times such that every subset is used as a testing set once. The average accuracy of the five time models is seen as the accuracy predicting capability of the model constructed with the machine learning method. 5-fold cross-validation can reflect the average performance of a model, however, it has the tendency of underestimating the prediction capability of a classification model, especially if important molecular features happen to be contained only in a minority of the compounds in the training set (155, 156). Hence if a model has relatively low cross-validation accuracy, it can still be predictive (155). Therefore, cross-validation alone is not decisive to the performance of a model. To complement cross-validation, independent validation datasets are used. They may provide a more reliable estimation of the prediction capability of a pharmacological property prediction model (157, 158). The independent validation dataset should be strictly independent from the training.

2.4.2 Performance evaluation

Measurements such as sensitivity, specificity and the overall prediction accuracy are employed to quantitatively assess the performance of virtual screening models. They are defined in terms of true positives TP (pharmaceutical agents possessing a specific pharmacological property), true negatives TN (pharmaceutical agents not possessing a specific pharmacological property), false positives FP (pharmaceutical agents not possessing a specific pharmacological property but predicted as agents possessing the specific pharmacological property) and false negatives FN (pharmaceutical agents possessing a specific pharmacological property but predicted as agents not possessing the specific pharmacological property). Sensitivity and specificity are the measurement of prediction accuracy for pharmaceutical agents possessing a specific pharmacological property and agents not possessing that pharmacological property respectively. The overall prediction accuracy (Q) and Matthews correlation coefficient (MCC) (159) are used to measure the overall prediction performance. They are defined as follows:

$$SE = \frac{TP}{TP + FN} \qquad (11)$$

$$SP = \frac{TN}{TN + FP} \qquad (12)$$

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \qquad (13)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \qquad (14)$$

The typical measurements of a model performance in screening large libraries include (44) yield (percentage of known positives predicted as virtual hits), hit-rate (percentage of virtual hits that are known positives), false hit-rate (percentage of virtual hits that are known negatives) and enrichment factor EF (magnitude of hit-rate improvement over random selection):

Yield = SE (15)

Hit-rate = $TP/(TP+FP)$ (16)

False hit-rate = $FP/(TP+FP)$ (17)

Enrichment factor EF = $\textit{hit-rate} / (TP+FN)/(TP+FN+TN+FP)$ (18)

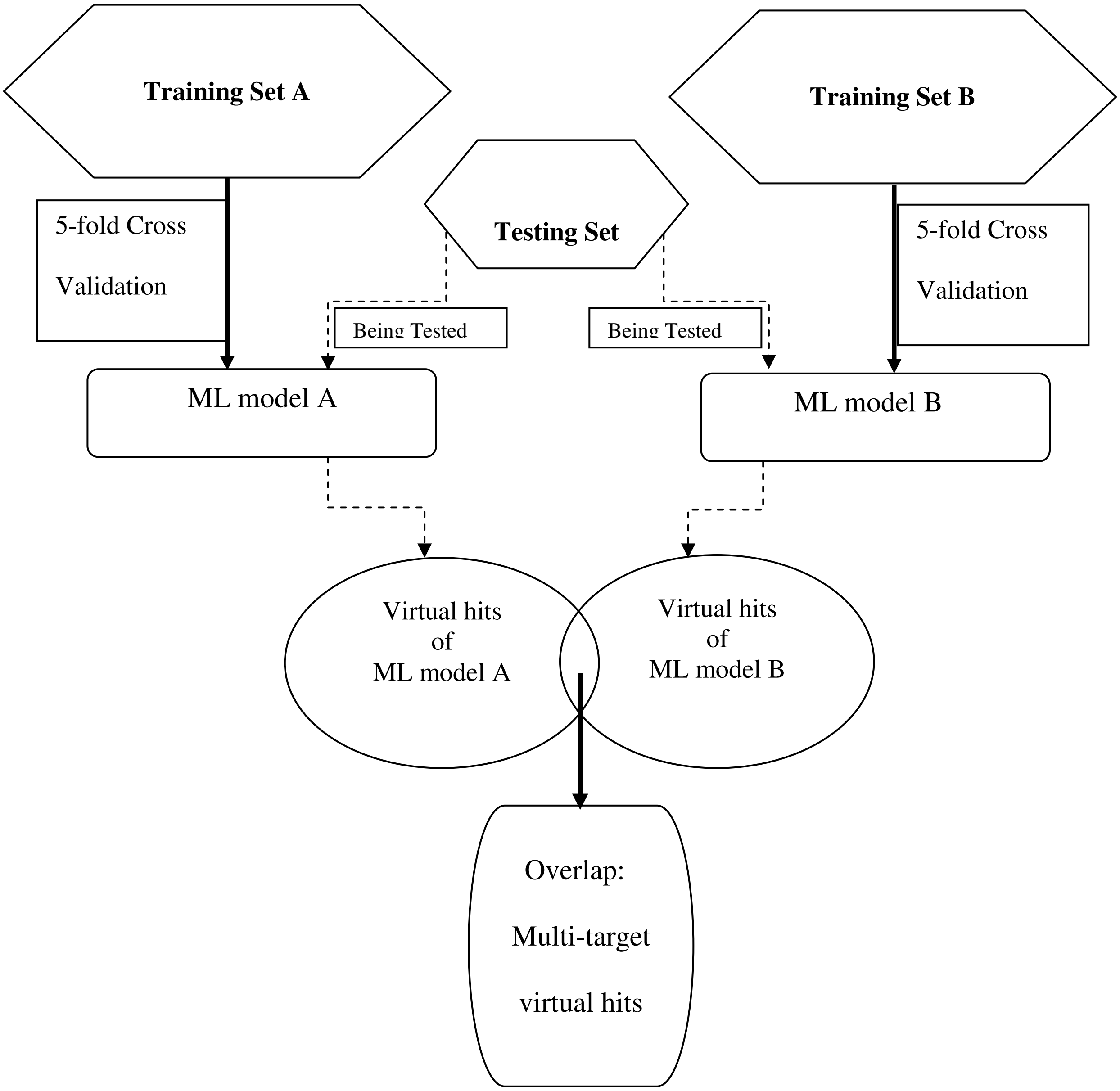
2.4.3 Overfitting problem and its detection

Overfitting is a major concern in machine learning classification methods. It happens when a model that agrees well with the observed data but has no predictive ability, which means it does not have any value to unseen or future data. There are two main types of overfitting situations: (1) a model more flexible than it needs to be and (2) a model including irrelevant descriptors (156). An overfitted classification system tends to obtain much higher prediction accuracies in the cross-validation sets than in the independent validation sets. Hence frequently used method for checking whether a model is overfitted is to compare the prediction accuracies in the cross-validation procedure with those found in testing independent validation sets (156).

2.5 Combinatorial Machine Learning Methods

Combinatorial machine learning methods are designed for the discovery of multi-target agents. They are composed of the machine learning methods applied for individual target agents. Virtual hits simultaneously selected by all individual VS tools are considered as multi-target virtual hits (160). The multi-target agents search capability of combinatorial machine learning methods is rigorously tested by excluding all known multi-target inhibitors from the training datasets and only those compounds known to be active against only one target in the target pair (these are tentatively referred to as individual-target inhibitors regardless of their possible activity against other targets outside the target pair) are used; The purpose of this exclusiveness is to test to what extent these individual-target based VS tools can identify multi-target inhibitors without explicit knowledge of known multi-target inhibitors (105). Target selectivity of combinatorial machine learning methods is assessed by using the known individual-target inhibitors of each target pair and those in the other target pairs used for the same disease treatment. This can reflect the selectivity of combinatorial machine learning methods against random selection. **Figure 2-4** shows the procedure of combinatorial machine learning methods in predicting dual-target agents.

Figure 2-4 The illustration of the procedure for combinatorial machine learning methods to predict dual-target agents



Chapter 3 Pharmainformatics Database Construction and Update

With the exponential increase in pharmaceutical information, it is becoming increasingly necessary and important to collect and curate the information to provide informative databases to greatly assist the studies of disease mechanisms and the discovery of new drugs. Pharmainformatics databases can provide up-to-date information and data that relate to disease mechanism studies, pharmaceutical research and drug development.

3.1 The update of Kinetic Database of Bio-molecular Interaction

3.1.1 Introduction to bio-molecular interactions

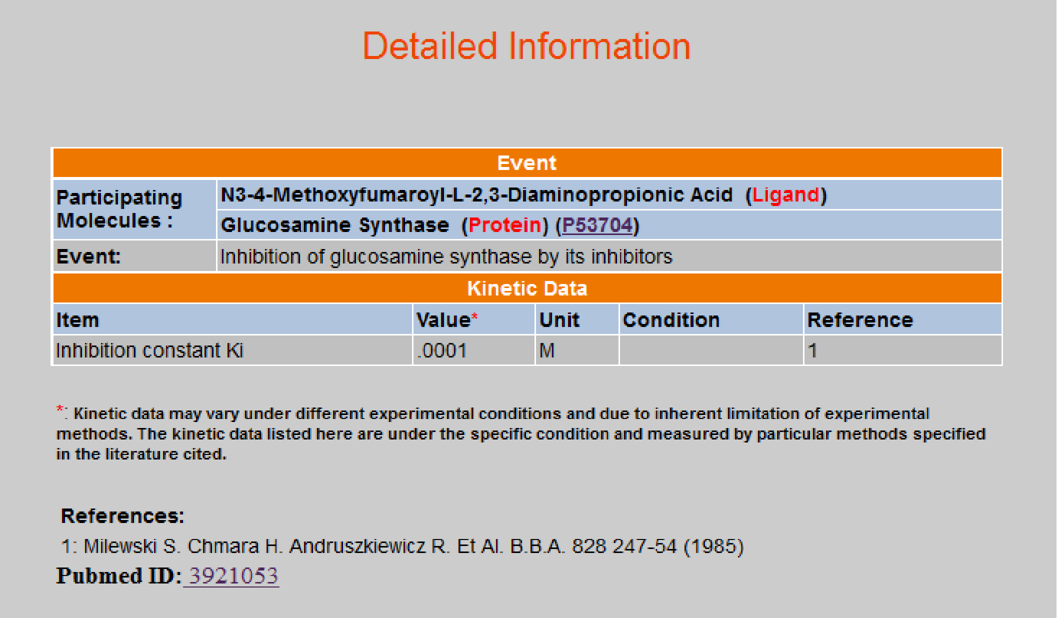
Via individual and network actions, bio-molecular interactions participate fundamentally in biological, disease, and therapeutic processes (161-164). Over the past 20 years, the understanding of the characteristics, organization, evolution and complexity of bio-molecular interaction networks in biological systems have significantly advanced thanks to the extensive experimental and computational studies (165-168) which also enabled the generation of genome-scale protein-protein interactions and the development prediction tools (166, 167, 169-172) . Quite a number of databases have been developed for providing information about bio-molecular interactions (e.g. MIPS(173), DIP (174), BIND (175) , Biocyc (176), MINT (177), Biomodels (178), STRING (179), and IntAct (180)), and biological networks and pathways (KEGG (181), BioGRID (182), NetworKIN (183), STITCH (184), DOMINE (185), CellCircuits (186), Reactome (187) and enzyme reactions (188)).

3.1.2 New features of updated KDBI

The updated KDBI includes a 2.3 fold increase of experimental kinetic data and four new features. The first new feature enables the access of KDBI entries via the list of nucleic acid and pathway names. The second new feature is added for facilitating the applications, assessments, and further development of the pathway models, it includes the literature-reported kinetic parameter sets of 63 pathway simulation models (189-198). The third new feature enables the facility for collectively accessing the available kinetic data of multi-step processes (e.g. metabolism, pathway segments) collected in KDBI. The fourth new feature provides the user with the SBML (199) files for all records of the kinetic parameter sets of pathway simulation models. This format can facilitate the use of the relevant data in such software tools as Celldesigner (200), Copasi (201), cPath (202), PaVESy (203), and SBMLeditor (204).

3.1.2.1 New Feature 1: nucleic acid and pathway names as KDBI entries

The additional sets of the experimentally determined kinetic data of bio-molecular interactions were collected from published literatures. The updated KDBI now contains 2635 protein-protein, 1711 protein-nucleic acid, 11873 protein-small molecule, and 1995 nucleic acid-small molecule interactions. Each entry provides detailed description about binding or reaction event, participating molecules, binding or reaction equation, kinetic data, and related references. Compared to the last version of KDBI, the number of entries in the updated KDBI is increased by 2.3 fold to 19263. As shown in **Figure 3-1**, kinetic data for protein-protein, small molecule-nucleic acid and protein-small molecule interactions is provided in terms of one or a combination of kinetic quantities as given in the literature of a particular event. These quantities include association/dissociation rate constant, on/off rate constant, first/second/third/… order rate constant, catalytic rate constant, equilibrium association/dissociation constant, inhibition constant, and binding affinity constant, IC50, etc. and experimental conditions (pH value and temperature).



Detailed Information

Event	
Participating Molecules :	**N3-4-Methoxyfumaroyl-L-2,3-Diaminopropionic Acid (Ligand)**
	Glucosamine Synthase (Protein) (P53704)
Event:	Inhibition of glucosamine synthase by its inhibitors

Kinetic Data

Item	Value*	Unit	Condition	Reference
Inhibition constant Ki	.0001	M		1

*: Kinetic data may vary under different experimental conditions and due to inherent limitation of experimental methods. The kinetic data listed here are under the specific condition and measured by particular methods specified in the literature cited.

References:

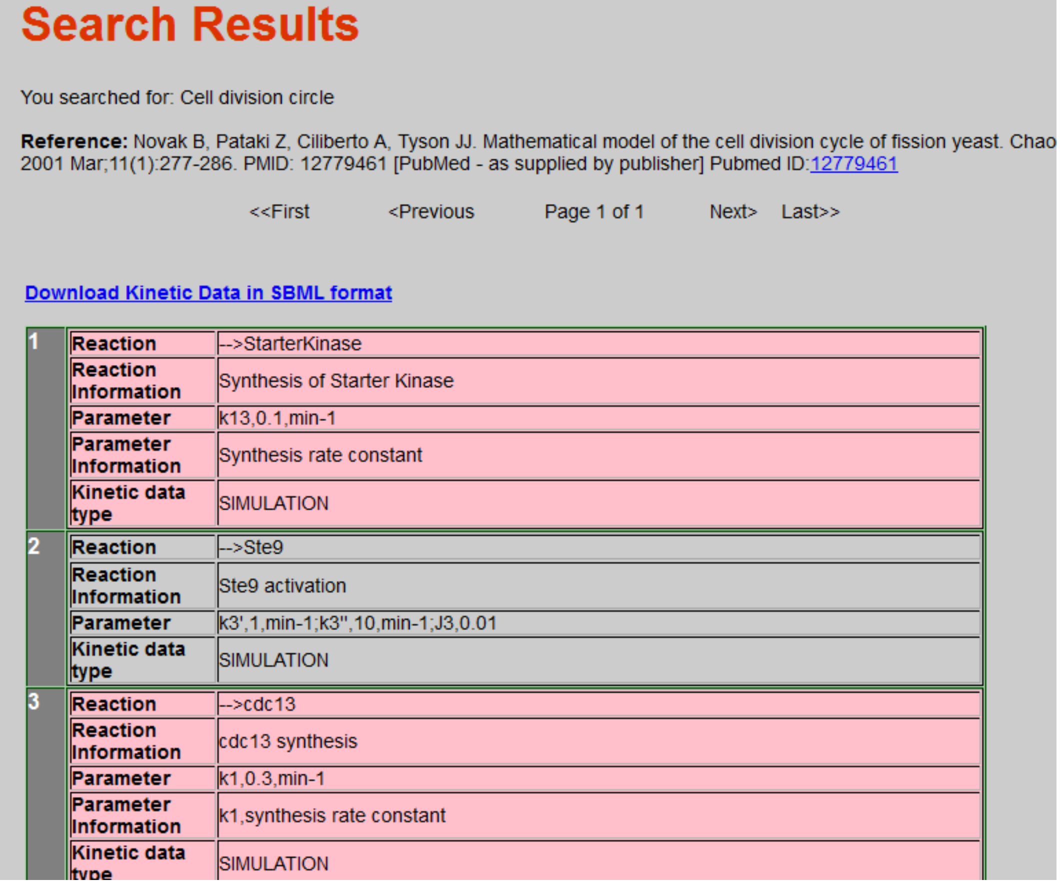
1: Milewski S. Chmara H. Andruszkiewicz R. Et Al. B.B.A. 828 247-54 (1985)

Pubmed ID: 3921053

Figure 3-1: Experimental kinetic data page showing protein–protein interaction; It includes kinetic data and reaction equation (while available) as well as the name of participating molecules and description of event.

3.1.2.2 New Feature 2: pathway simulation models

Mathematical simulation models of various pathways have been developed and extensively used for studying and quantitative understanding of signaling dynamics (189-193), signal specific sensing (194) and discrimination (198), feedback regulations and crosstalk (196, 197), and receptor cross-activation (195) and internalization (196). They have greatly assisted the understanding and quantitative analysis of complex biological processes and network responses. In the mathematical models, the temporal dynamic behaviors of molecular species in the pathway are typically described by ordinary differential equations (ODEs). The kinetic rate constants of protein–protein, protein-small molecule, protein-nucleic acid, and other interactions (e.g. binding association rate Kf, binding dissociation rate Kb, reaction rate K, reaction turnover rate Kcat, Michaelis–Menten constant Km) are needed to establish these ODEs. Those data have been primarily generated by combinations of experimental data, computed theoretical values, and empirically fitted values computational (193-198) . Therefore, in order to facilitate further applications, developments, and assessments of the published pathway models, in the KDBI update, I collected parameter sets of 63 published ODE-based models, which can be accessed from the pathway list in the "Pathway Simulation Parameters" field in KDBI webpage. Additionally, the data type of kinetic data was included to every entry to clearly distinguish its original source (experimental or simulation model). In particular, when the data come from a simulation model, the cross reference to the original source is provided. **Figure 3-2** demonstrates a pathway simulation model page.



Search Results

You searched for: Cell division circle

Reference: Novak B, Pataki Z, Ciliberto A, Tyson JJ. Mathematical model of the cell division cycle of fission yeast. Chao 2001 Mar;11(1):277-286. PMID: 12779461 [PubMed - as supplied by publisher] Pubmed ID:12779461

<<First <Previous Page 1 of 1 Next> Last>>

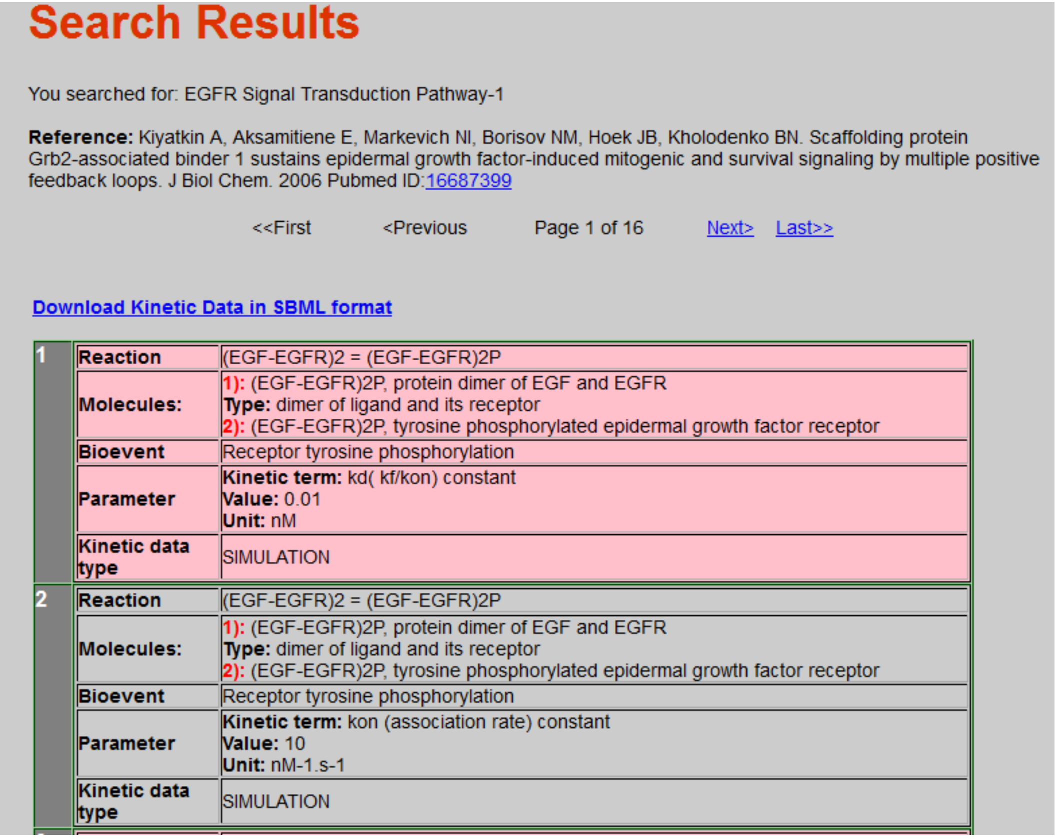
Download Kinetic Data in SBML format

1	**Reaction**	-->StarterKinase
	Reaction Information	Synthesis of Starter Kinase
	Parameter	k13,0.1,min-1
	Parameter Information	Synthesis rate constant
	Kinetic data type	SIMULATION
2	**Reaction**	-->Ste9
	Reaction Information	Ste9 activation
	Parameter	k3',1,min-1;k3",10,min-1;J3,0.01
	Kinetic data type	SIMULATION
3	**Reaction**	-->cdc13
	Reaction Information	cdc13 synthesis
	Parameter	k1,0.3,min-1
	Parameter Information	k1,synthesis rate constant
	Kinetic data type	SIMULATION

Figure 3-2 This page provides kinetic data and reaction equation (when available) as well as the name of participating molecules and description of event in the pathway simulation models.

3.1.2.3 New Feature 3: multi-step processes of kinetic data

Multi-step processes have caught the interest and attention and there have been published studies providing information about the experimental kinetic data for the multiple components involved in multi-step processes (205-207). Some examples of these processes include RNA binding activity to translation initiation factors eIF4G, 70-kDa Heat Shock Protein polymerization, control of platelet function by cyclic AMP, GroEL interaction with conformational states of horse cytochrome c, intermolecular catalysis by hairpin ribozymes, antisense RNA interaction with its complimentary RNA, nucleotide binding to actin. To facilitate the development of pathway simulation models based on these building blocks, direct access to the collection of the kinetic data for each of these processes are provided in the KDBI update, which can be accessed via a separate search field "Multi-step processes" in KDBI webpage. Figure 3-3 illustrates the multi-step process data displaying page.



Search Results

You searched for: EGFR Signal Transduction Pathway-1

Reference: Kiyatkin A, Aksamitiene E, Markevich NI, Borisov NM, Hoek JB, Kholodenko BN. Scaffolding protein Grb2-associated binder 1 sustains epidermal growth factor-induced mitogenic and survival signaling by multiple positive feedback loops. J Biol Chem. 2006 Pubmed ID:16687399

<<First <Previous Page 1 of 16 Next> Last>>

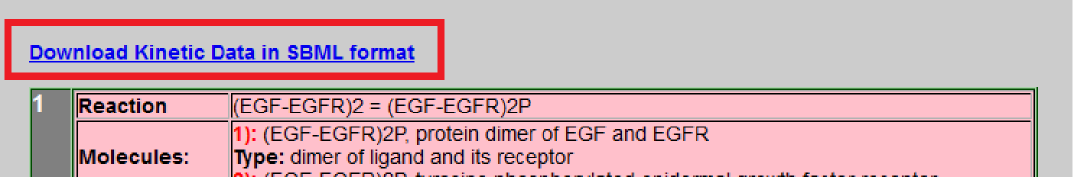
Download Kinetic Data in SBML format

1	**Reaction**	(EGF-EGFR)2 = (EGF-EGFR)2P
	Molecules:	1): (EGF-EGFR)2P, protein dimer of EGF and EGFR **Type:** dimer of ligand and its receptor 2): (EGF-EGFR)2P, tyrosine phosphorylated epidermal growth factor receptor
	Bioevent	Receptor tyrosine phosphorylation
	Parameter	**Kinetic term:** kd(kf/kon) constant **Value:** 0.01 **Unit:** nM
	Kinetic data type	SIMULATION
2	**Reaction**	(EGF-EGFR)2 = (EGF-EGFR)2P
	Molecules:	1): (EGF-EGFR)2P, protein dimer of EGF and EGFR **Type:** dimer of ligand and its receptor 2): (EGF-EGFR)2P, tyrosine phosphorylated epidermal growth factor receptor
	Bioevent	Receptor tyrosine phosphorylation
	Parameter	**Kinetic term:** kon (association rate) constant **Value:** 10 **Unit:** nM-1.s-1
	Kinetic data type	SIMULATION

Figure 3-3 Multi-process kinetic data page provides kinetic data and reaction equation (when available) as well as the name of participating molecules and description of event

3.1.2.4 New Feature 3: SBML availability

Incredibly bio-information is being offered by numerous laboratories and research groups across the globe and hundreds of software for bioinformatics and chemo-informatics are used to process those data. Therefore, it is critical to adopt a unified data format that is compatible across the different software platforms. To this end, systems Biology Markup Language (SBML) has been developed as a free, open, XML-based format for representing biochemical reaction networks, and it is a software-independent language for describing models common to computational biology research, including cell signaling pathways, metabolic pathways, gene regulation, and others (208). Many pathway simulation and analysis software tools have built-in SBML compatibility features to allow the input, manipulation, simulation and analysis of different pathway models and parameters (199, 208-212). To meet the demand for SBML formatted date, the SBML files for the parameter sets of all 63 pathway simulation models included in KDBI were also created. These files can be downloaded via the link provided on the top of the page that displays the relevant kinetic data. Moreover, the SMBL viewer is provided for the convenience of the users, which can be found in the home page of KDBI. The SBML formatted data are offered in each query result page. See **Figure 3-4**



Download Kinetic Data in SBML format

1	**Reaction**	(EGF-EGFR)2 = (EGF-EGFR)2P
	Molecules:	1): (EGF-EGFR)2P, protein dimer of EGF and EGFR **Type:** dimer of ligand and its receptor

Figure 3-4 The boxed part is link to where the SBML format data are offered. This link is presented in every query result page.

3.2 Update of Therapeutic Targets Database

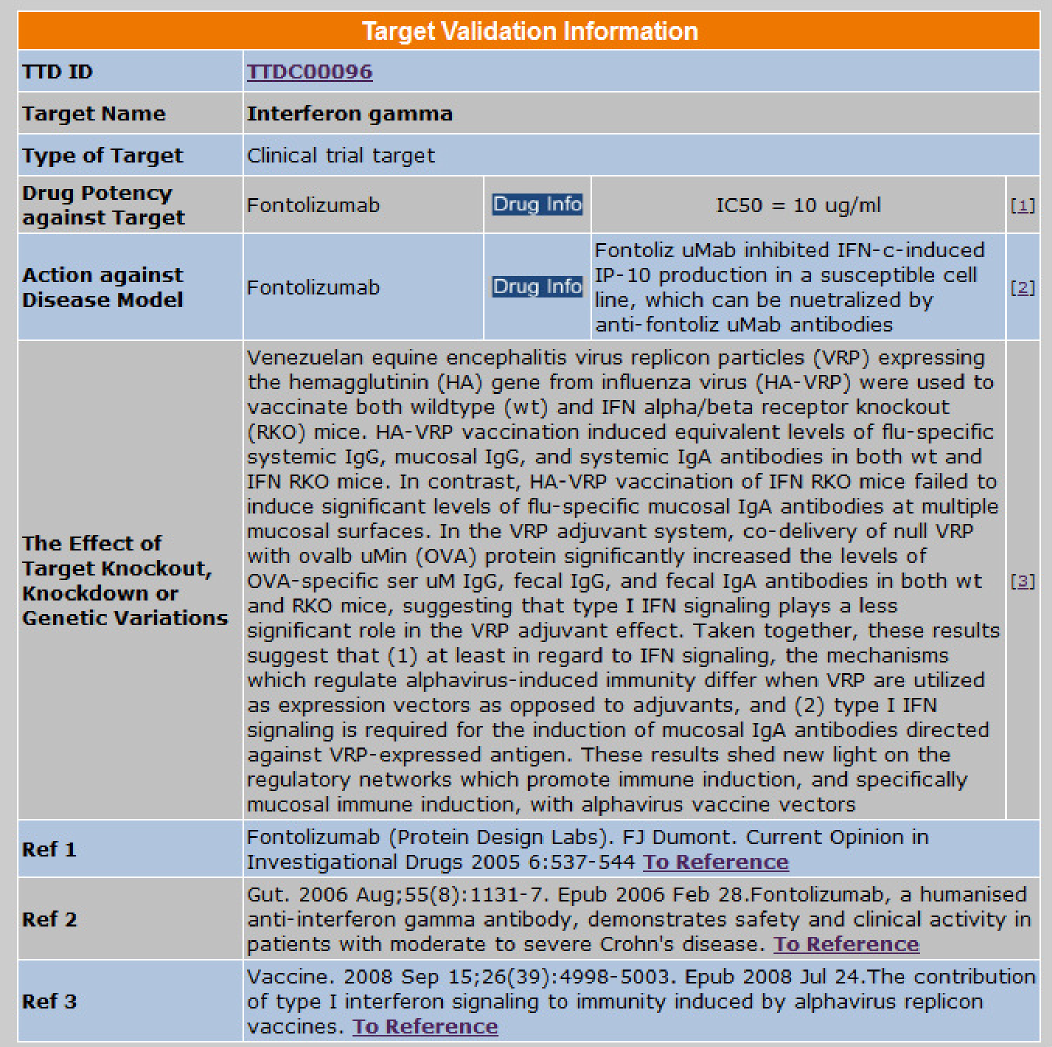
The studies of therapeutic targets (responsible for drug efficacy) and the targeted drugs can greatly facilitate both the discovery, validation of new targets and the development and optimization of new drug leads. Therapeutic Target Database (TTD) is developed to provide comprehensive information of the known efficacy targets and the corresponding approved, clinical trial and investigative drugs. Since its last update in 2010, major improvements and updates have been made to TTD. First of all, a significant increase of data content (from 1,894 targets and 5,028 drugs to 2,025 targets and 17,816 drugs plus 3,681 multi-target agents) has been made to the updated version. Besides, target validation information (drug potency against target, action against disease model, and the effect of target knockout, knockdown or genetic variations) for 932 targets (351 successful, 252 clincial trial, 34 discontinued and 295 research targets), and 841 quantitative structure activity relationship (QSAR) models for active compounds of 228 targets (71 of the targets are successful target, 20 are clinical trial and 30 are research targets) are added to the updates. Additionally, drug combinations and nature-derived drugs information are presented in the updated TTD. These updates are particularly useful for providing relevant information and facilitating target discovery and validation, drug lead discovery and optimization, and the development of multi-target drugs.

3.2.1 Target validation

The decision to develop a drug against a particular target is usually a considerable commitment in terms of time and money. Once a target enters a pharmaceutical company's pipeline, it can take about 12 years to develop a marketable drug. Each new drug that reaches the market represents research and development costs of close to US$1 billion (213). Despite the huge investment in time and money, the number of NDAs (New Drug Applications) approved each year by the Food and Drug Administration (FDA), however, has declined from 53 in 1996 to 35 in 1999 to 17 in 2002 to 15 in 2005.

After passing the pre-clinical trials, pre-drugs will go through clinical trials, which are the tests conducted on humans. There are usually three stages in the clinical trials: Phase I (screening for safety), Phase II (establishing the testing protocol), Phase III (final testing). And sometimes, Phase IV will be conducted for post-approval studies. Drugs fail in the clinic usually for two basic reasons: it is either that they do not work as they were expected or they prove to be unsafe for patients. The ultimately ideal way to be completely certain that a drug can affect a protein instrumentally in a given disease is to test the idea in humans. Obviously such clinical trials cannot be applied for initial drug development, which means that a potential target must undergo other validation processes to clearly define its role in disease before drugs are sought that act against it, or before it is used to screen large numbers of compounds for drug activity. There are several aspects that can reflect a target’s validation. Such measurements include the target’s association with disease pathology, its expression in cells linked to disease pathology and animal models. A relatively new route to target validation adopts the disrupting of gene expression. This is to reduce the amount of the corresponding protein, and so to identify the physiological role of the target. Examples of this technique include gene knockouts, antisense technology and RNA interference.

To facilitate the validation of current targets, TTD update includes the target validation data for 932 targets. The validation information collected is classified into three types: drug potency against target, action against disease model, and the effect of target knockout, knockdown or genetic variations. Currently, TTD provides complete or partial validation information for 932 targets (351 successful, 252 clinical trial, 34 discontinued and 295 research targets). This collection of target validation data can offer a good reference for drug development. **Figure 3-5** gives an example of the TTD update in target validation.



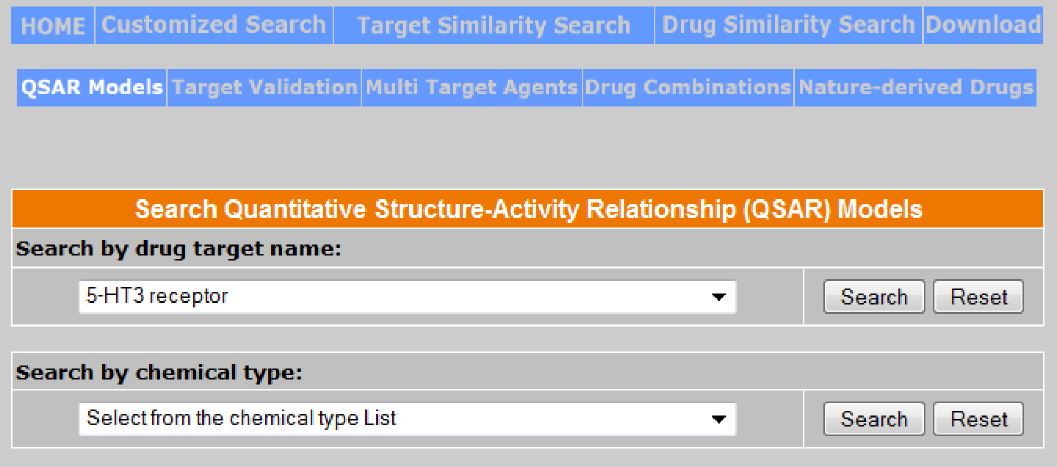
Target Validation Information				
TTD ID	TTDC00096			
Target Name	**Interferon gamma**			
Type of Target	Clinical trial target			
Drug Potency against Target	Fontolizumab	Drug Info	IC50 = 10 ug/ml	[1]
Action against Disease Model	Fontolizumab	Drug Info	Fontoliz uMab inhibited IFN-c-induced IP-10 production in a susceptible cell line, which can be nuetralized by anti-fontoliz uMab antibodies	[2]
The Effect of Target Knockout, Knockdown or Genetic Variations	Venezuelan equine encephalitis virus replicon particles (VRP) expressing the hemagglutinin (HA) gene from influenza virus (HA-VRP) were used to vaccinate both wildtype (wt) and IFN alpha/beta receptor knockout (RKO) mice. HA-VRP vaccination induced equivalent levels of flu-specific systemic IgG, mucosal IgG, and systemic IgA antibodies in both wt and IFN RKO mice. In contrast, HA-VRP vaccination of IFN RKO mice failed to induce significant levels of flu-specific mucosal IgA antibodies at multiple mucosal surfaces. In the VRP adjuvant system, co-delivery of null VRP with ovalb uMin (OVA) protein significantly increased the levels of OVA-specific ser uM IgG, fecal IgG, and fecal IgA antibodies in both wt and RKO mice, suggesting that type I IFN signaling plays a less significant role in the VRP adjuvant effect. Taken together, these results suggest that (1) at least in regard to IFN signaling, the mechanisms which regulate alphavirus-induced immunity differ when VRP are utilized as expression vectors as opposed to adjuvants, and (2) type I IFN signaling is required for the induction of mucosal IgA antibodies directed against VRP-expressed antigen. These results shed new light on the regulatory networks which promote immune induction, and specifically mucosal immune induction, with alphavirus vaccine vectors			[3]
Ref 1	Fontolizumab (Protein Design Labs). FJ Dumont. Current Opinion in Investigational Drugs 2005 6:537-544 To Reference			
Ref 2	Gut. 2006 Aug;55(8):1131-7. Epub 2006 Feb 28.Fontolizumab, a humanised anti-interferon gamma antibody, demonstrates safety and clinical activity in patients with moderate to severe Crohn's disease. To Reference			
Ref 3	Vaccine. 2008 Sep 15;26(39):4998-5003. Epub 2008 Jul 24.The contribution of type I interferon signaling to immunity induced by alphavirus replicon vaccines. To Reference			

Figure 3-5 An example for target validation information presented in the updated TTD

3.2.2 QSAR models

Quantitative structure-activity relationships (QSAR) attempts to correlate structural or property descriptors of compounds with their activities. QSAR use physicochemical descriptors to represent the features of the compounds for model construction. These physicochemical descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects, are determined empirically or, more recently, by computational methods. Activities used in QSAR include chemical measurements and biological assays. QSAR methods have been intensively applied in biomolecular discovery and drug design (214). A great number of QSAR models have been constructed for various targets with descriptors of all kinds. Therefore, it would be very helpful to summarize as many as possible already constructed models to provide an informative and convenient reference. TTD now includes 841 QSAR models for active compounds of 228 distinct chemical types against 121 targets (71 of the targets are successful target, 20 are clinical trial and 30 are research targets). It provides two means for search: search by drug target name and search by chemical type. And in the search result page, QSAR models are described in details in the pdf format files which can be downloaded in the result page by clicking “QSAR model page”. **Figure 3-6** shows the search page for the QSAR update of TTD and **Figure 3-7** is an example of the result page.

Figure 3-6 The QSAR model search page offers search by target and search by chemical type

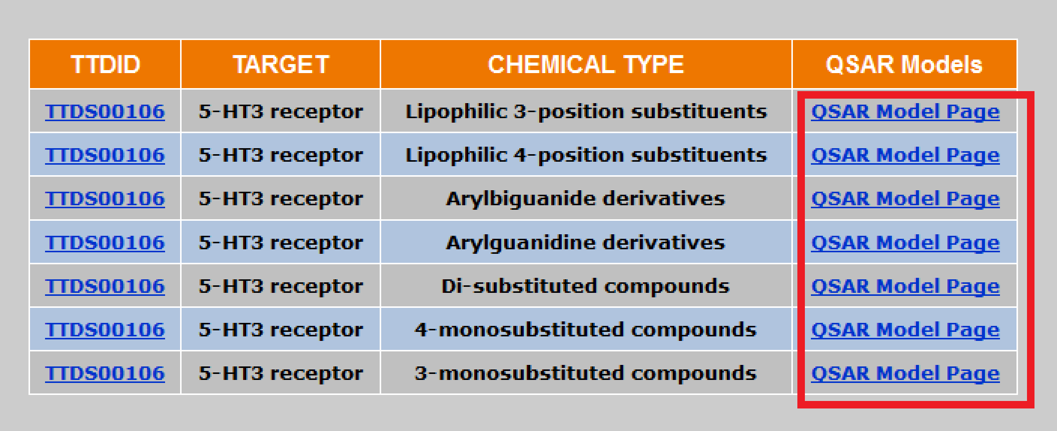


Search Quantitative Structure-Activity Relationship (QSAR) Models

Search by drug target name: 5-HT3 receptor

Search by chemical type: Select from the chemical type List

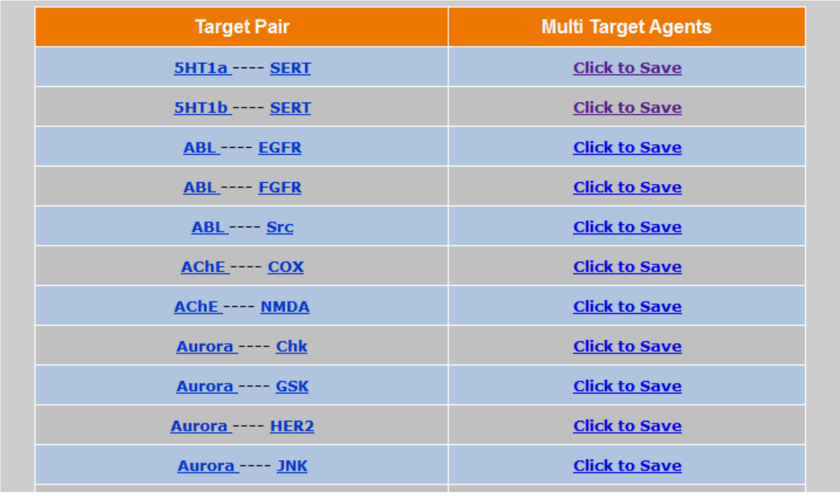
Figure 3-7 An example of the search page for QSAR models. Detailed description of QSAR models can be downloaded via the link "QSAR model page"



TTDID	TARGET	CHEMICAL TYPE	QSAR Models
TTDS00106	5-HT3 receptor	Lipophilic 3-position substituents	QSAR Model Page
TTDS00106	5-HT3 receptor	Lipophilic 4-position substituents	QSAR Model Page
TTDS00106	5-HT3 receptor	Arylbiguanide derivatives	QSAR Model Page
TTDS00106	5-HT3 receptor	Arylguanidine derivatives	QSAR Model Page
TTDS00106	5-HT3 receptor	Di-substituted compounds	QSAR Model Page
TTDS00106	5-HT3 receptor	4-monosubstituted compounds	QSAR Model Page
TTDS00106	5-HT3 receptor	3-monosubstituted compounds	QSAR Model Page

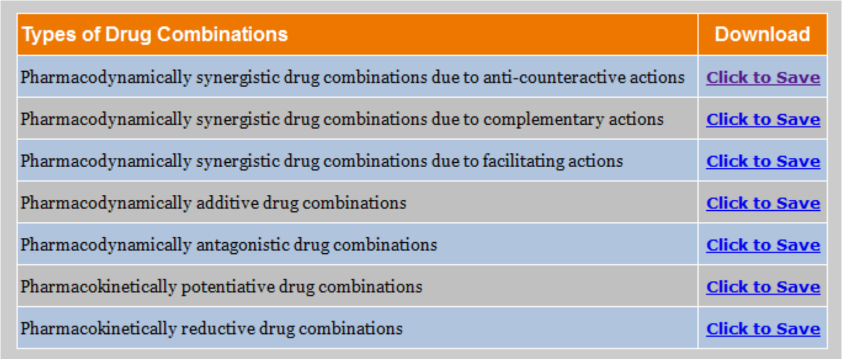
3.2.3 Other update features

Multi-target agents have become a new trend in drug design as this approach is showing the hope to conquer the issues confronting the traditional single-target drugs such as low efficacy and side effects (215). There have been intensive studies with rigorous analysis on the effect of drug combinations for which the combination effect has been evaluated by and for which relevant molecular interaction profiles of the drugs involved. These combinations are found to reveal general and specific modes of action (88). Nature derived drugs have always been widely applied by the traditional medicines of many cultures. Recently, Low drug productivity has renewed interest in natural products as drug-discovery sources (216). In order to keep pace with these new drug discovery trends, TTD updates provides structure and potency information of 3,681 multi target agents against 108 target pairs, drug-combination data of 72, 14 and 4 pharmacodynamically synergistic, additive, and antagonist combinations respectively, 19 and 7 pharmacokinetically potentiative and reductive combinations together with their mode of actions and combination mechanisms and 939, 369 and 119 nature-derived approved, clinical trial and preclinical drugs together with their species origin information. All data are available for user to download. **Figure 3-8**, **Figure 3-9** and **Figure 3-10** present the downloading page for these data. Right click “Click to save” and choose “Save link as” option, the data then can be saved to the users’ preferred destination.



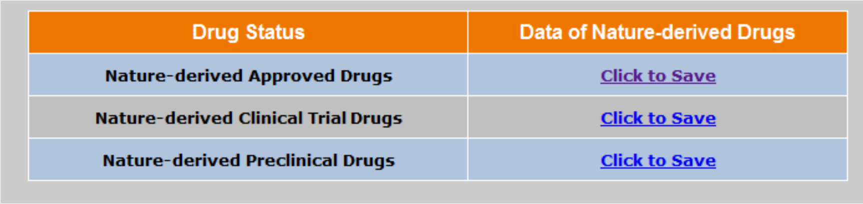
Target Pair	Multi Target Agents
5HT1a ---- SERT	Click to Save
5HT1b ---- SERT	Click to Save
ABL ---- EGFR	Click to Save
ABL ---- FGFR	Click to Save
ABL ---- Src	Click to Save
AChE ---- COX	Click to Save
AChE ---- NMDA	Click to Save
Aurora ---- Chk	Click to Save
Aurora ---- GSK	Click to Save
Aurora ---- HER2	Click to Save
Aurora ---- JNK	Click to Save

Figure 3-8



Types of Drug Combinations	Download
Pharmacodynamically synergistic drug combinations due to anti-counteractive actions	Click to Save
Pharmacodynamically synergistic drug combinations due to complementary actions	Click to Save
Pharmacodynamically synergistic drug combinations due to facilitating actions	Click to Save
Pharmacodynamically additive drug combinations	Click to Save
Pharmacodynamically antagonistic drug combinations	Click to Save
Pharmacokinetically potentiative drug combinations	Click to Save
Pharmacokinetically reductive drug combinations	Click to Save

Figure 3-9



Drug Status	Data of Nature-derived Drugs
Nature-derived Approved Drugs	Click to Save
Nature-derived Clinical Trial Drugs	Click to Save
Nature-derived Preclinical Drugs	Click to Save

Figure 3-10

Figure 3-8, Figure 3-9, Figure 3-10 Downloading pages for multi-target agents, Drug combination information and Nature-derived drugs

Chapter 4 Preliminary Tests of Combinatorial Machine Learning Methods in Screening Multi-target Agents

4.1 Introduction: Multi-target Kinase Inhibitor Therapeutics for Cancer Treatment

Tyrosine kinases (TKs) are usually classified into receptor tyrosine kinases (RTKs) (e.g. epidermal growth factor receptor (EGFR), vascular endothelial growth factor (VEGFR)) and cytoplasmic/non-receptor kinase (e.g. Src, Lck). They play pivotal roles in diverse cellular activities including growth, differentiation, metabolism, adhesion, motility, death (217). For example, EGFR is found to be overexpressed or aberrantly activated in the most common solid turmors, including non-small cell lung cancer, breast cancer, prostate cancer and colon cancer. It has been proven that tyrosine kinases have particularly important implications in development of cancers. Therefore, they have emerged as clinically useful drug target molecules for treating certain types of cancer (218). Several tyrosine kinase inhibitors (TKIs) (e.g. Gefitinib as EGFR inhibitor, Avastin as VEGFR inhibitor) have stumblingly survived the drug development stages and been applied for cancer treatment in clinical treatment of cancer (219). Despite the discovery and application of those kinase inhibitors, almost inevitably, cancer patients treated with single-target develop drug resistance and suffer a relapse. Moreover, many tumors are multi-factorial and are linked to defects in more than one signaling pathways, and the inhibition of a single molecule may not be sufficient to interfere efficiently with disease progression (220). As a result, a multi-target approach has become a prevailing idea and the new generation of TKIs is selected on the basis of their ability simultaneously to target different molecules. Based on their importance in cancerogenesis implications, I selected five TKs (EGFR, FGFR, VEGFR, Src and Lck) to test virtual screening (VS) for the studies of inhibitors of the four dual-target pair, EGFR-FGFR, EGFR-Src, VEGFR-Lck and Src-Lck. The strategy is to use experimentally obtained small-scale multi-target kinase inhibitors profiles to predict inhibitors in a larger kinase set (221) by means of virtual screening. In principle, single-target VS tools may be combined to collectively identify multi-target agents. This is practically useful when the individual VS tools have sufficiently high yields and low false-hit rates. High yields can compensate for the reduced collective yields of combinatorial VS tools (For two statistically-independent VS tools of 50%-70% yields, the collective yield of their combination is roughly the product of the yield of individual tools, which is 25%-49%). Low false-hit rates contribute to high enrichment factors in searching multi-target agents that are more sparsely distributed in the chemical space than non-dual inhibitors (**Table 4-1**).

Table 4-1 Datasets of dual-inhibitors and non-dual-inhibitors of the kinase-pairs used for developing and testing combinatorial SVM virtual screening tools; Additional sets of 13.56 million PubChem compounds and 168 thousand MDDR active compounds were also used for the test.

Kinase Pair	**Inhibitors in Training Sets**						**Inhibitors and Other Compounds in Testing Set**				
Kinase A – Kinase B	Training Set for Kinase A			Training Set for Kinase B			Dual Inhibitors of A and B				MDDR Compounds Similar to Dual Inhibitors of A and B
	No of inhibitors of A that are non-inhibitor of B (No of families)	No of these inhibitors that are in the B inhibitor families (No of families)	No of these inhibitors that are in the families of dual inhibitors of A and B (No of families)	No of inhibitors of B that are non-inhibitor of A (No of families)	No of these inhibitors that are in the A inhibitors families (No of families)	No of these inhibitors that are in the families of dual inhibitors of A and B (No of families)	No of dual inhibitors of A and B (No of families)	No (%) of dual inhibitors in the families that contain both A and B non-dual inhibitor in training sets	No (%) of dual-inhibitors of A and B as inhibitor of at least one of the other 5 kinases studied in this work	No (%) of dual-inhibitors of A and B as inhibitor of more than 2 of the other 5kinases studied in this work	No of Compounds
EGFR-FGFR	1303 (388)	284 (52)	160 (22)	392 (131)	154 (52)	124 (27)	71 (39)	37 (52.1%)	70 (98.6%)	2 (2.8%)	1001
EGFR-Src	1262 (372)	331 (73)	166 (31)	748 (216)	243 (73)	168 (38)	112 (64)	46 (41.1%)	46 (41.1%)	2 (1.8%)	1127
VEGFR-Lck	1232 (427)	220 (69)	102 (17)	445 (171)	206 (69)	52 (11)	61 (23)	29 (47.5%)	37 (60.7%)	0 (0.0%)	413
Src-Lck	804 (236)	222 (49)	98 (11)	450 (175)	160 (49)	23 (9)	56 (17)	23 (41.1%)	38 (67.9%)	0 (0.0%)	276

4.2 Materials and Methods

4.2.1 Compound collection, training and testing datasets, molecular descriptors

We collected a total of 233-1,316 non-dual inhibitors of EGFR, VEGFR, FGFR, Src, Lck and 56-188 dual inhibitors of EGFR-FGFR, EGFR-Src, VEGFR-Lck,, and Src-Lck, each with IC_{50}≤10μM, were collected from the litterature (222-231) and the BindingDB database (125). Here dual-inhibitors and non-dual inhibitors of a kinase-pair are defined as inhibitors of both and one of the two kinases respectively regardless of their activities against other kinases. **Table 4-1** summarizes the statistics of these inhibitors and MDDR compounds similar to at least one dual-inhibitor. The implication of machine learning methods requires both positive (e.g. the active compounds) and negative data (e.g. the inactive compounds). As few non-inhibitors have been reported, putative non-inhibitors of each kinase were generated by using our published method that requires no knowledge of inactive compounds or active compounds of other target classes and enables more expanded coverage of the "non-inhibitor" chemical space (85, 102). First, 13.56 million PubChem and 168 thousand MDDR compounds were clustered into 8,993 compound families of similar molecular descriptors (151). These are consistent with the reported 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms (8), and 2,851 clusters for 171,045 natural products (153). A total of

42,670- 44,115 compounds extracted from the 8,534-8,823 families (5 per family) that contain no known inhibitor were used as the putative non-inhibitors.

In this study, I used a total of 98 important descriptors calculated by the program MODEL. The details about molecular descriptors have been explained in **Chapter 2 Section 2.2.1**

4.2.2 Computational methods

Support Vector Machine (SVM) is based on the structural risk minimization principle of statistical learning theory (132). SVM can detect active compounds fast by differentiating physicochemical profiles rather than structural similarity to active compounds *per se*. It does not require knowledge of target structures and the computation of structural flexibility, activity-related features, solvation effects and binding affinities. It has shown outstanding classification performance, less chance being penalized by sample redundancy, low over-fitting risks. It is capable of accommodating large and structurally diverse training and testing datasets, and is fast in performing classification tasks (134, 135). Although the performance of SVM critically depends on the diversity of training datasets thus the limited knowledge of known inhibitors for many kinase targets may hinder the application of sufficiently good SVM VS, its high yields and low false-hit rates in searching single-target agents (103) sometimes even based on sparsely distributed active compounds (102) still make it a potentially good virtual screening tool for the exploration of multi-target agents. In this study, I derived multi-target SVM VS tool: combinatorial SVMs (COMBI-SMV), which combines the prediction of two separate SVM classifier for each the multiple kinases.

Figure 4-1 illustrates the application of COMBI-SVM for searching multi-target inhibitors. The SVM VS models were developed by using a hard margin c=100,000 and their σ values are in the range of 0.1-2. In terms of the numbers of true positives TP (true inhibitors), true negatives TN (true non-inhibitors), false positives FP (false inhibitors), and false negatives FN (false non-inhibitors), the yield and false-hit rate are given by TP/(TP+FN) and FP/(TP+FP) respectively.

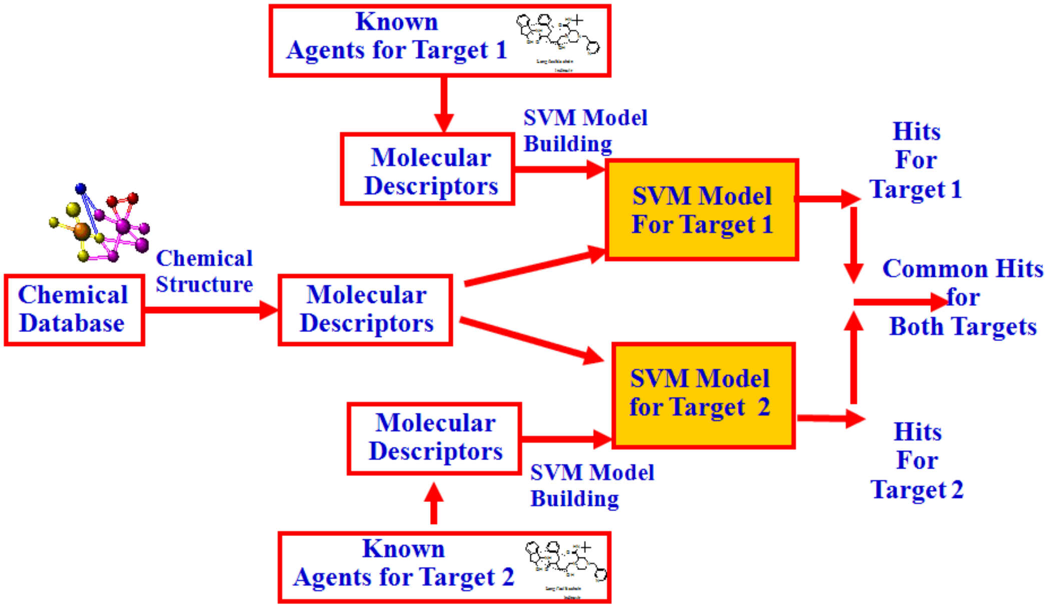


Figure 4-1 Illustration of combinatorial support vector machines method (COMBI-SVM) for searching multi-target inhibitors for searching multi-target inhibitors

4.3 Results and Discussion

4.3.1 Virtual screening performance of Combinatorial SVM in searching kinase dual-inhibitors from large libraries

The performance of combinatorial SVM (COMBI-SVM) for identifying the 4 dual-kinase inhibitors is summarized in **Table 4-2**. We used 5-fold cross-validation to select the COMBI-SVM model parameters for the evaluated kinases and they reside in the ranges of σ=0.5~0.8. The dual-inhibitor yields are 40.9% for EGFR-FGFR, 26.8% for EGFR-Src, 52.6% for VEGFR-Lck, and 48.2% for Src-Lck respectively. Besides, the yields for the intra-PTK group are comparable to the expected 25%-49% yields of combinations of good VS tools with individual yields of 50%-70%. This shows reasonably good capability for COMBI-SVM in identifying multi-target agents for kinase-pairs within a protein kinase group without requiring explicit knowledge of multi-target agents.

The target selectivity is conducted by two means. Firstly, I tested the target selectivity of COMBI-SVM by screening the 233-1,316 non-dual inhibitors of the 4 kinase pairs. The misidentifying rates are 10.1% and 8.7% of the non-dual inhibitors of the kinase pair as dual-inhibitors for EGFR-FGFR, 12.9% and 11.1% for EGFR-Src, 6.6% and 29.2% for VEGFR-Lck, 15.8% and 18.7% for Src-Lck (see **Table 4-2**). The misidentification of a substantial percentage of non-dual inhibitors as dual-inhibitors might be caused by the following two reasons. 1) SVMs were trained exclusively by non-dual inhibitors, which may make it difficult for a SVM model to fully distinguish dual and non-dual inhibitors. 2) Some of the misidentified non-dual inhibitors are probably true dual-inhibitors not yet experimentally tested for multi-target activities. Hence the “mistaken” selection of these non-dual inhibitors can still be utilized in tests as possible dual-inhibitor candidates. Therefore, COMBI-SVMs have reasonably good selectivity in distinguishing dual-inhibitors from non-dual inhibitors.

Virtual-hit rates and false-hit rates of COMBI-SVMs in detecting compounds similar in the structural and physicochemical properties to the training datasets were measured by using 276-1,127 MDDR compounds similar to a dual-inhibitor of each kinase-pair. Similarity was defined by Tanimoto similarity coefficient ≥0.9 between a MDDR compound and its closest dual-inhibitor (102). COMBI-SVMs identified 65 virtual hits from 1,001 MDDR compounds (6.5%) for EGFR-FGFR, 24 from 1,127 MDDR compounds (2.1%) for EGFR-Src, 21 from 413 MDDR compounds (5.1%) for VEGFR-Lck, and 26 from 276 MDDR compounds (9.4%) for Src-Lck.

The virtual-hit rates and thus false hit rates were tested by screening large libraries of 168 thousand MDDR and 13.56 million PubChem database. The numbers of virtual-hits and virtual-hit rates in screening 168 thousand MDDR compounds are 126 and 0.07% for EGFR-FGFR, 162 and 0.096% for EGFR-Src, 170 and 0.1% for VEGFR-Lck, and 131 and 0.078% for Src-Lck. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 2,200 and 0.015% for EGFR-FGFR, 4,471 and 0.033% for EGFR-Src, 4,817 and 0.036% for VEGFR-Lck, 2,674 and 0.02% for Src-Lck. COMBI-SVM hence showed significantly low false-hit rates in screening large libraries.

Analysis of the MDDR virtual hits showed that substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, and signal transduction inhibitors (**Table 4-3**). As some of these virtual-hits may be true dual-inhibitors, the actual number of true false-hits may be smaller than the total number of virtual-hits for each kinase-pair. Hence, the false-hit rates of the combinatorial SVMs are at most equal to and likely less than the virtual-hit rates. Hence the false-hit rates are ≤2.13%-9.4% in screening 276-1,127 MDDR similarity compounds, ≤0.078%-0.10% in screening 168 thousand MDDR compounds, and ≤0.002%-0.011% in screening 13.56 million PubChem compounds, which are comparable and in some cases better than single-target false-hit rates of 0.0054%-8.3% of single-target SVMs (29, 102), 0.08%-3% of structure-based methods, 0.1%-5% by other machine learning methods, 0.16%-8.2% by clustering methods, and 1.15%-26% by pharmacophore models (232).

Table 4-2 Virtual screening performance of combinatorial SVMs for identifying dual-inhibitors of 4 combinations of EGFR, VEGFR,FGFR, Src and Lck

Kinase	**Virtual Screening Performance**							
	Dual inhibitors		Non-dual inhibitors of the same kinase pair		MDDR compounds similar to dual inhibitors	All 168 thousand MDDR compounds	13.56 million PubChem comnds	1.02 million Zinc clean-leads dataset
	Yield	No (%) of identified true hits outside the common training active families of both kinases	False hit rate for inhibitors of kinase A	False hit rate for inhibitors of kinase B	Virtual hit rate (No of virtual hits)	Virtual hit rate (No of virtual hits)	Virtual hit rate (No of virtual hits)	Virtual hit rate (No of virtual hits)
EGFR-FGFR	40.90%	6 (8.5%)	10.10%	8.70%	6.5% (65)	0.07% (126)	0.016% (2200)	0.004% (36)
EGFR-Src	26.80%	13 (11.6%)	12.90%	11.10%	2.13% (24)	0.096% (162)	0.033% (4471)	0.007% (76)
VEGFR-Lck	52.60%	8 (13.1%)	6.60%	29.20%	5.1% (21)	0.10% (170)	0.036% (4817)	0.011% (113)
Src-Lck	48.20%	9 (16.1%)	15.80%	18.70%	9.4% (26)	0.078% (131)	0.020% (2674)	0.002% (25)

Table 4-3 MDDR classes that contain higher percentage (≥9%) of virtual-hits identified by combinatorial SVMs in screening 168 thousand MDDR compounds for dual-inhibitors of 4 combinations of EGFR, VEGFR, FGFR, Src and Lck

Kinase Pair	**No of SVM Identified Virtual Hits**	**MDDR Classes that Contain Higher Percentage of Virtual Hits**	**No of Virtual Hits in Class**	**Percentage of Class member as Virtual Hits**
EGFR-FGFR	126	Antineoplastic	78	0.40%
		Tyrosine-Specific Protein Kinase Inhibitor	47	4.00%
		Antiarthritic	37	0.30%
		Signal Transduction Inhibitor	23	1.10%
		Antiangiogenic	16	1.00%
EGFR-Src	162	Antineoplastic	95	0.40%
		Tyrosine-Specific Protein Kinase Inhibitor	42	3.60%
		Signal Transduction Inhibitor	39	1.90%
		Antiangiogenic	21	1.30%
		Antiarthritic	15	0.10%
VEGFR-Lck	170	Antineoplastic	87	0.40%
		Antiarthritic	42	0.40%
		Tyrosine-Specific Protein Kinase Inhibitor	36	3.00%
		Signal Transduction Inhibitor	31	1.50%
		Antiangiogenic	16	1.00%
		Atherosclerosis Therapy	10	0.90%
		Antiarthritic	10	0.10%
Src-Lck	131	Antineoplastic	65	0.30%
		Tyrosine-Specific Protein Kinase Inhibitor	34	2.90%
		Antiarthritic	23	0.20%
		Signal Transduction Inhibitor	17	0.80%
		Antineoplastic Enhancer	14	2.20%

4.3.2 Analysis of combinatorial sVM identified MDDR virtual hits

The virtual hits of MDDR detected by combinatorial SVM (COMBI-SVM) were analyzed based on the known biological or therapeutic target classes specified in MDDR. **Table 4-3** has listed the MDDR classes that contain higher percentage (≥9%) of COMBI-SVM virtual-hits and the percentage values. It is shown that 65-95 (41.6%-61.9%) of the 126-170 virtual-hits belong to the antineoplastic class, which represent 0.30%-0.40% of the 21,557 MDDR compounds in the class. In particular, 34-47 (21.2%-37.3%) of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represents 2.9%-4.00% of the 1,181 MDDR compounds in the class. Moreover, 17-39 (13.0%-24.1%) of the virtual-hits are the numbers of the signal transduction inhibitor representing 0.80%-1.9% of the 2,037 members in this class. Therefore, many of the COMBI-SVM virtual-hits are antineoplastic compounds that may also inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis and other cancer-related pathways. Although some of these kinase inhibitors might be true dual-inhibitors of specific kinase pairs, the majority of them are expected to be false selection of non-dual inhibitors of the same kinase-pairs (at 6.6%-29.2% false-hit rates).

Some of the COMBI-SVM virtual hits belong to the antiarthritic class. Four of the evaluated kinases have been linked to arthritis in the literature. EGFR-like receptor stimulates synovial cells and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis (29). VEGF has also been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis (233). FGFR may partly mediates osteoarthritis (234). Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma (235). Therefore, some of the COMBI-SVM virtual-hits in the antiarthritic class could actually be capable of producing antiarthritic activities.

Moreover, Multiple FGFRs are elevated in atherosclerotic lesions in apoE-/- micand and active FGFR-1 signalling promotes atherosclerosis development via increased SMC proliferation and by augmenting macrophage accumulation via increased expression of MCP-1 and factors promoting macrophage retention in lesions (236). Hence, some of the COMBI-SVM virtual hits in the atherosclerosis therapy could act as dual inhibitors of the two kinases.

4.4 Conclusion

In these preliminary tests for the search of dual-kinase inhibitors, combinatorial SVM (COMBI-SVM) VS tools developed by exclusively using non-dual inhibitors showed good yields for dual-inhibitors of several anticancer target kinase pairs and in many cases. The capability of the combinatorial SVMs and other VS tools in identifying multi-kinase inhibitors and other multi-target agents may be further enhanced by incorporating knowledge of multi-target agents into VS tool development processes. When more multi-target kinase inhibitors are found and tested to prove effective on their targets, it is possible to introduce more comprehensive elements of distinguished structural and physicochemical features of selective multi-target agents into the training of combinatorial VS tools. This could in turn enhance more effective identification of selective multi-target agents. In order to improve the target selectivity, the multi-target VS tools can be combined with structure-based filters. Because of their high computing speed and generalization capability, combinatorial SVM can be potentially studied and applied as useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of multi-kinase inhibitors and other multi-target agents.

Chapter 5 The Application of Combinatorial Machine Learning Methods in Virtual Screening of Selective Multi-target Antidepressant Agents

5.1 Introduction

Major depression is a prevailing, heterogeneous, and often incapacitating disorder. It is triggered by complex patterns such as genetic, epigenetic, developmental, and environmental factors (237). Major depression is characterized as an episode of change in mood that lasts for weeks or months. It is one of the most severe types of depression. It usually involves a low or irritable mood and/or a loss of interest or pleasure in usual activities. It interferes with one's normal functioning and often includes physical symptoms. A person may experience only one episode of major depressive disorder, but often there are repeated episodes over an individual's lifetime. The effects of major depression could be devastating and lead to suicide. Hence antidepressants have become one of the largest therapeutic areas of current drug market (63). Despite the efforts spent in the treatment of major depression, Antidepressant drug discovery has been a complex task due to incomplete understanding of neurobiological basis of depression. A primary anti-depression strategy is to inhibit monoamine reuptakes, such as serotonin reuptake, by both single-target and multi-target drugs (95). Commonly used antidepressants, such as the selective serotonin (5-HT) reuptake inhibitors (SSRIs) (e.g. Fluoxetine) are often effective, but the full efficacy takes several weeks to achieve and many patients only partially respond to the drugs while some others remain refractory (237). Single-target drugs (78, 85) frequently encounter reduced efficacy and drug resistance problems caused by network robustness (78), redundancy (79), crosstalk (80), compensatory and neutralizing actions (81), anti-target and counter-target activities (82), and on-target and off-target toxicities (83) . Multi-target drugs are particularly useful for avoiding these problems.

Multi-target monoamine inhibitor drugs achieve enhanced efficacies by several mechanisms. The first one involves the inhibition of multiple monoamine reuptakes (1). The simultaneous blockade of complementary monoamine reuptakes synergistically enhances the overall therapeutic efficacy (238). Specific types of monoamines in CNS are reduced both by a primary monoamine transporter and by alternative transporters (239, 240). For instance, 5-HT is reduced primarily by serotonin transporter (SERT), and secondarily by noradrenaline transporter (NET) and dopamine transporter (DAT) particularly at high levels of 5-HT and/or when SERT function/expression is compromised (240). Therefore, inhibition of one monoamine reduction route is complemented by the inhibition of the other routes to reduce their compensatory activities, leading to therapeutic synergy. This multi-target strategy is the basis for developing dual serotonin reuptake and noradrenaline reuptake inhibitors (NETSRIs) as antidepressant drugs of fast and enhanced therapeutic effects (241).

DES-VENLATAFINE and TESOFENSINE are good examples of NETSRI and dual SERT and DAT inhibitor respectively (**Figure 5-1**).

The second mechanism involves collective monoamine reuptake inhibition and receptor antagonism. For instance, it has been reported that increased release of 5-HT by SERT inhibition stimulates 5-HT$_{1A}$, 5-HT$_{1B}$, and 5-HT$_{1C}$ autoreceptors, which subsequently reduces 5-HT release, thereby delaying the therapeutic effect of serotonin reuptake inhibitors until the 5-HT$_{1A}$ and 5-HT$_{1B}$/$_{1C}$ autoreceptors become desensitized (242). This counteractive effect can be reduced by simultaneous targeting of serotonin transporters and 5-HT$_{1A}$, 5-HT$_{1B}$ or 5-HT$_{1C}$ receptors. Indeed, co-administration of a 5-HT$_{1A}$ receptor antagonist with a selective serotonin reuptake inhibitor leads to an immediate increase in CNS 5-HT levels (243) and shortened onset of anxiolytic activity (244). SSA-426 is an example of dual SERT and 5-HT$_{1A}$ receptor antagonist (**Figure 5-1**). Histamine H$_3$ receptor also promotes counteractive effect against serotonin reuptake inhibition by mediating the inhibition of serotonin release in the brain (245, 246). Therefore, in some circumstances, simultaneous targeting of serotonin reuptake transporter and histamine H$_3$ receptor achieves an improved antidepressant effect by more enhanced 5-HT release (247).

Another mechanism involves bimodal antidepressant actions. This approach aims at reducing the undesirable actions of selective serotonin reuptake inhibitors (SSRIs) through multi-targeted inhibition of other related receptors. Some of the undesirable actions of SSRIs, such as the short-term anxiety, arise from stimulation of 5-HT$_{2C}$ receptor, and 5-HT$_{2C}$ receptors also mediate the inhibitory effects of SSRIs on sleep, sexual function, and appetite (1). Therefore, serotonin reuptake inhibitors with antagonist activities against 5-HT$_{2C}$ receptor sites are expected to show a better tolerability than SSRIs (1). AGOMELATINE is a dual serotonin reuptake inhibitor and 5-HT$_{2C}$ receptor antagonist (5HT2cAntags) (Figure 1) with clinically proven activity against major depression. The blockade of neurokinin 1 (NK$_1$) receptors by NK$_1$ receptor antagonists (NK1Antags) not only complement the effects of serotonin reuptake inhibition but also accelerate the long-term facilitating influence of SSRIs on serotonergic transmission (237). Therefore, dual serotonin reuptake inhibitor and NK$_1$ receptor antagonist, such as UCB (**Figure 5-1**), is expected to be more efficacious and faster in achieving therapeutic effects than SSRIs. Moreover, dual serotonin reuptake inhibitor and melanocortin 4 (MC$_4$) receptor antagonist (MC4Antags), such as MCL10004 (Figure 1), has been found to interlink neuropeptide receptor antagonist activity with SRI activity to synergistically improve mood (237).

Extensive efforts have been directed at the development of multi-target serotonin reuptake inhibitors (e.g. dual serotonin reuptake and noradrenaline reuptake inhibitors (NETSRIs) (248, 249), dual serotonin reuptake inhibitor and 5-HT$_{1A}$ receptor antagonists (5HT1aSRIs) (250, 251), dual serotonin reuptake inhibitor and 5-HT$_{1B}$ receptor antagonists (5HT1bSRIs) (252), dual serotonin reuptake inhibitor and H_3 receptor antagonists (H3SRIs) (247), dual serotonin reuptake inhibitor and 5-HT_{2C} receptor antagonists (5HT2cSRIs) (253), dual serotonin reuptake inhibitor and MC_4 receptor antagonists (MC4SRIs) (254) and dual serotonin reuptake inhibitor and NK_1 receptor antagonists (NK1SRIs) (255)) based on the above mechanisms. While *in-silico* methods have been extensively used for searching selective serotonin reuptake inhibitors (256, 257), noradrenaline reuptake inhibitors (258),(259), $5HT_{1A}$ receptor antagonists (260, 261) and H_3 receptor antagonists (262, 263), these methods have been used in a few published works for searching NETSRIs, 5HT1aSRIs, 5HT1bSRIs, H3SRIs, 5HT2cSRIs, MC4SRIs and NK1SRIs (252) (264, 265). Therefore, in order to identify multi-target agents that are more sparsely distributed in the chemical space than single-target agents, there is a strong need to explore *in-silico* methods more extensively, particularly those methods capable of searching large compound libraries at good yields and low false-hit rates.

In this work, I used a machine learning method, support vector machines (SVM), to develop the combinatorial SVM (COMBI-SVM) virtual screening (VS) tool for searching dual-target agents NETSRIs, 5HT1aSRIs, 5HT1bSRIs, H3SRIs, 5HT2cSRIs, MC4SRIs and NK1SRIs. In Chapter 4, COMBI-SVM has been tested as dual-kinase inhibitor VS tools with reasonably good yields, target selectivity and low false-hit rates in searching large compound libraries (105). Hence, it is time to apply COMBI-SVM to search the dual-target antidepressant agents NETSRIs, H3SRIs, 5HT1aSRIs, 5HT1bSRIs, 5HT2cSRIs, MC4SRIs and NK1SRIs from large compound libraries.

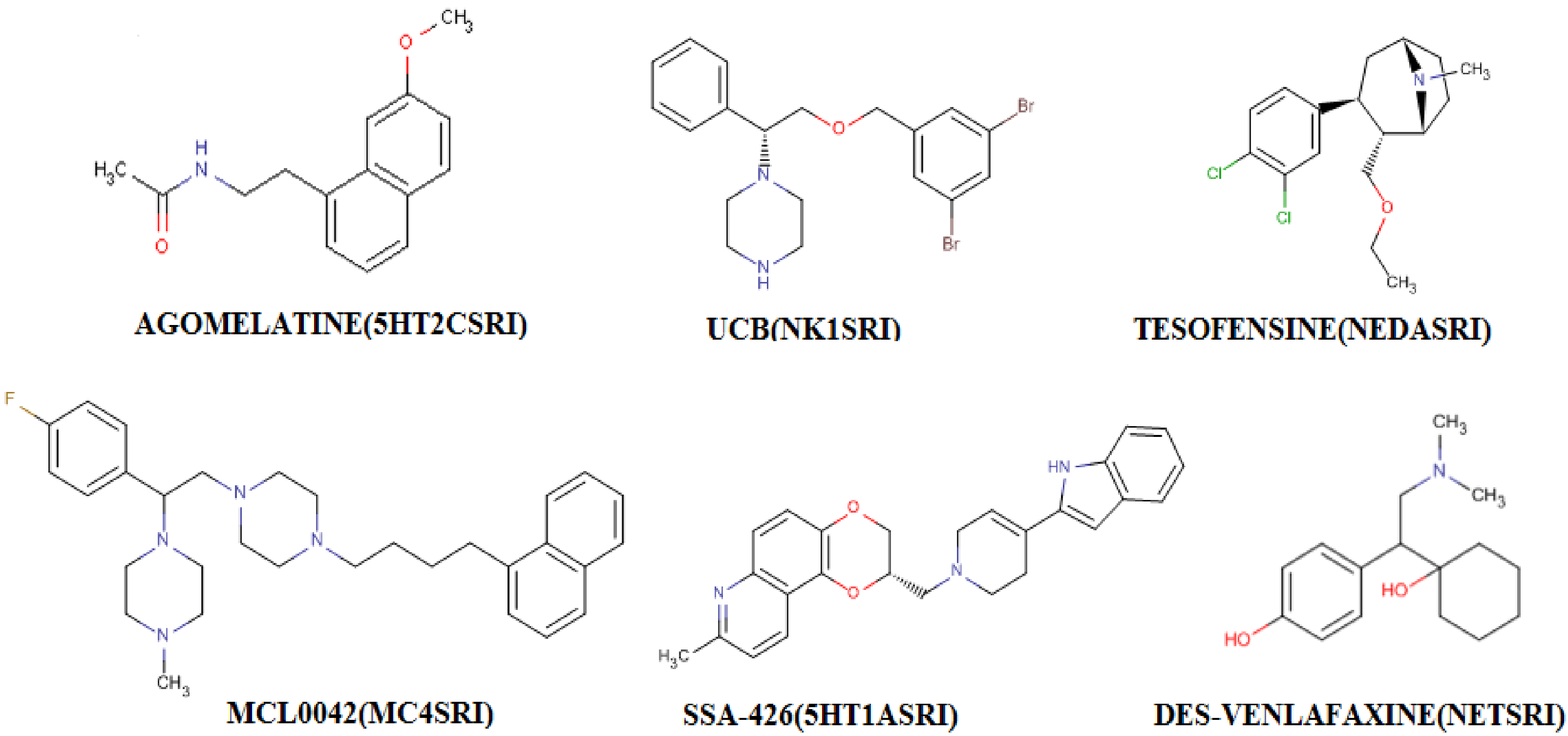


Figure 5-1 Examples of multi-target multi-target serotonin reuptake inhibitors; NETSRI=dual serotonin reuptake and noradrenaline reuptake inhibitor; NEDASRI= serotonin, dopamine, and noradrenaline reuptake inhibitor; 5HT1ASRI: dual serotonin reuptake inhibitor/5-HT_{1A} ; NK1SRI=dual serotonin reuptake inhibitor/neurokinin 1 receptor antagonist; MC4SRI=dual serotonin reuptake inhibitor/melanocortin 4 receptor antagonist.

5.2 Materials and Methods

5.2.1 Data collection and molecular descriptors

Individual-target and dual-target inhibitors, each with IC50 or K_i value ≤10μM, were collected from the literature (247, 248, 251), and the ChEMBL (266) and BindingDB (125) databases. The collected individual-target inhibitors include 1125-1951 SSRIs, 1410 noradrenaline reuptake inhibitors (NRIs), 1689 H_3 receptor antagonists (H3Antags), 1144 5-HT_{1A} receptor antagonists (5HT1aAntags), 917 5-HT_{1B} receptor antagonists (5HT1bAntags), 1234 5-HT_{2C} receptor antagonists (5HT2cAntags), 1721 melanocortin 4 receptor antagonists (MC4Antags) and 1787 neurokinin 1 receptor antagonists (NK1Antags). The collected dual inhibitors are 101 dual serotonin reuptake/noradrenaline reuptake inhibitors (NETSRIs), 147 dual serotonin reuptake inhibitor/H_3 receptor antagonists (H3SRIs), 216 dual serotonin reuptake inhibitor/5-HT_{1A} receptor antagonists (5HT1aSRIs), 57 dual serotonin reuptake inhibitor/5-HT_{1B} receptor antagonists (5HT1bSRIs), 27 dual serotonin reuptake inhibitor/5-HT_{2C} receptor antagonists (5HT2cSRIs), 6 dual serotonin reuptake inhibitor/melanocortin 4 receptor antagonists (MC4SRIs) and 45 dual serotonin reuptake inhibitor/neurokinin 1 receptor antagonists (NK1SRIs), Table 5-1 summarises the datasets of these individual-target inhibitors, dual-inhibitors and MDDR compounds similar to at least one dual-inhibitor for each the target pair used as the training and testing sets in this work. Figure 5-1 illustrates the composition of the collected dual-inhibitors of the seven studied.

As few non-inhibitors have been reported, putative non-inhibitors of each target were generated by using our published method that requires no knowledge of inactive compounds or active compounds of other target classes and enables more expanded coverage of the "non-inhibitor" chemical space (29, 102). First, 17 million PubChem and 168 thousand MDDR compounds were clustered into 8,993 compound families of similar molecular descriptors (267), which are consistent with the reported 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms (268), and 2,851 clusters for 171,045 natural products (269).

The putative non-inhibitors for each target were extracted from those families (5-8 per family) that contain no known individual-target inhibitors. The specific numbers of putative non-inhibitors are 60726-62593 from 7590-8018 families for SERT, 61957 from 7937 families for NET, 61960 from 7937 families for H_3 receptor, 62376 from 7991 families for 5-HT_{1A} receptor, 64790 from 8114 families for $5HT_{1B}$ receptor, 61912 from 7739 families for 5-HT_{2C} receptor, 63807 from 7976 families for MC_4 receptor and 62733 from 7842 families for NK_1 receptor. This approach has the risk of the wrong exclusion of the compound families that contain multi-target inhibitors and undiscovered individual-target inhibitors from the non-inhibitor training dataset. The maximum possible "wrong" classification rate arising from these mistakes has been estimated at <13% even in the extreme and unlikely cases that all of the undiscovered single-target and multi-target agents are misplaced into the non-inhibitor class (102), (103). The noise level generated by up to 13% “wrong” negative compound family representation is expected to be substantially smaller than the maximum 50% false-negative noise level tolerated by SVM (270).

We used the 98 descriptors which include 18 descriptors in the class of simple molecular properties, 3 descriptors in the class of chemical properties, 35 descriptors in the class of molecular connectivity and shape, 42 descriptors in the class of electro-topological state. These descriptors are described in **Chapter 2 Section 2.2.1.** This set of descriptors has been selected in our previous studies for representing diverse structural and physicochemical properties of both inhibitors of a specific target and non-inhibitors of that target distributed in large chemical space defined by 13.56 million Pubchem compounds (**Chapter 4**). Although the structures of inhibitors of one target can be very different from those of another target, each inhibitor set plus the representatives of the non-inhibitors cover the same chemical space defined by the 13.56 million Pubchem compounds. Therefore, the same set of molecular descriptors was used in this work as well as our previous works. The virtual screening models of vastly different biochemical classes (kinases, GPCR agonists/antagonists, peptidase inhibitors, DHFR inhibitors, and HDAC inhibitors) developed by this same descriptor set have shown equally good performance in screening large chemical libraries (102, 103, 105) (160).

Table 5-1 Datasets of individual-target inhibitors, dual inhibitors and MDDR compounds similar to at least one dual inhibitor used as the training and testingsets in this work

Target Pair	**Inhibitors in Training Sets**						**Inhibitors and Other Compounds in Testing Set**					
	Training Set for Target A			Training Set for Target B			Multi-Target Agents of Target A and B				Inhibitors of other six targets outside target-pair	MDDR Compounds Similar to Multi-Target Inhibitors of A and B
Target A – Target B	No of inhibitors of A that are non-inhibitor of B (No of families)	No of these inhibitors that are in the B inhibitor families (No of families)	No of these inhibitors that are in the families of multi-target agents of A and B (No of families)	No of inhibitors of B that are non-inhibitor of A (No of families)	No of these inhibitors that are in the A inhibitor families (No of families)	No of these inhibitors that are in the families of multi-target agents of A and B (No of families)	No of multi-target agents of A and B (No of families)	No (%) of multi-target agents in the families that contain single-target inhibitor of A in training sets	No (%) of multi-target agents in the families that contain single-target inhibitor of B in training sets	No (%) of multi-target agents outside the families that contain single-target inhibitor of A or B in training sets (No of families)	No of inhibitors	No of Compounds
SERT-NET	1125(405)	399(124)	113(33)	1410(486)	471(124)	176(42)	101(73)	65(64.3%)	46(45.5%)	25 (24.8%)	8389	8181
SERT-H$_3$	1804(604)	366(95)	39(16)	1689(486)	345(95)	124(28)	147(56)	97(65.9%)	53(36.1%)	27 (18.4%)	8191	1486
SERT-5HT$_{1A}$	1679(590)	512(130)	121(26)	1144(432)	421(130)	151(26)	216 (71)	130 (60.2%)	120 (55.6%)	52 (24.1%)	8354	7349
SERT-5HT$_{1B}$	1894(631)	514(108)	164(22)	917(309)	424(108)	93(11)	57(35)	21(41.2%)	42(73.7%)	14 (24.6%)	8688	7475
SERT-5HT$_{2C}$	1924(631)	689(145)	28(10)	1234(493)	405(145)	36(9)	27(23)	10(37.0%)	13(48.1%)	10(37.0%)	8426	1302
SERT-MC$_4$	1951(644)	175(61)	2(2)	1721(248)	557(61)	2(2)	6(2)	6(100%)	6(100%)	0	8164	7
SERT-NK$_1$	1910(631)	262(69)	39(8)	1787(358)	219(69)	62(8)	45(23)	29(64.4%)	9(20%)	9(20%)	8110	275

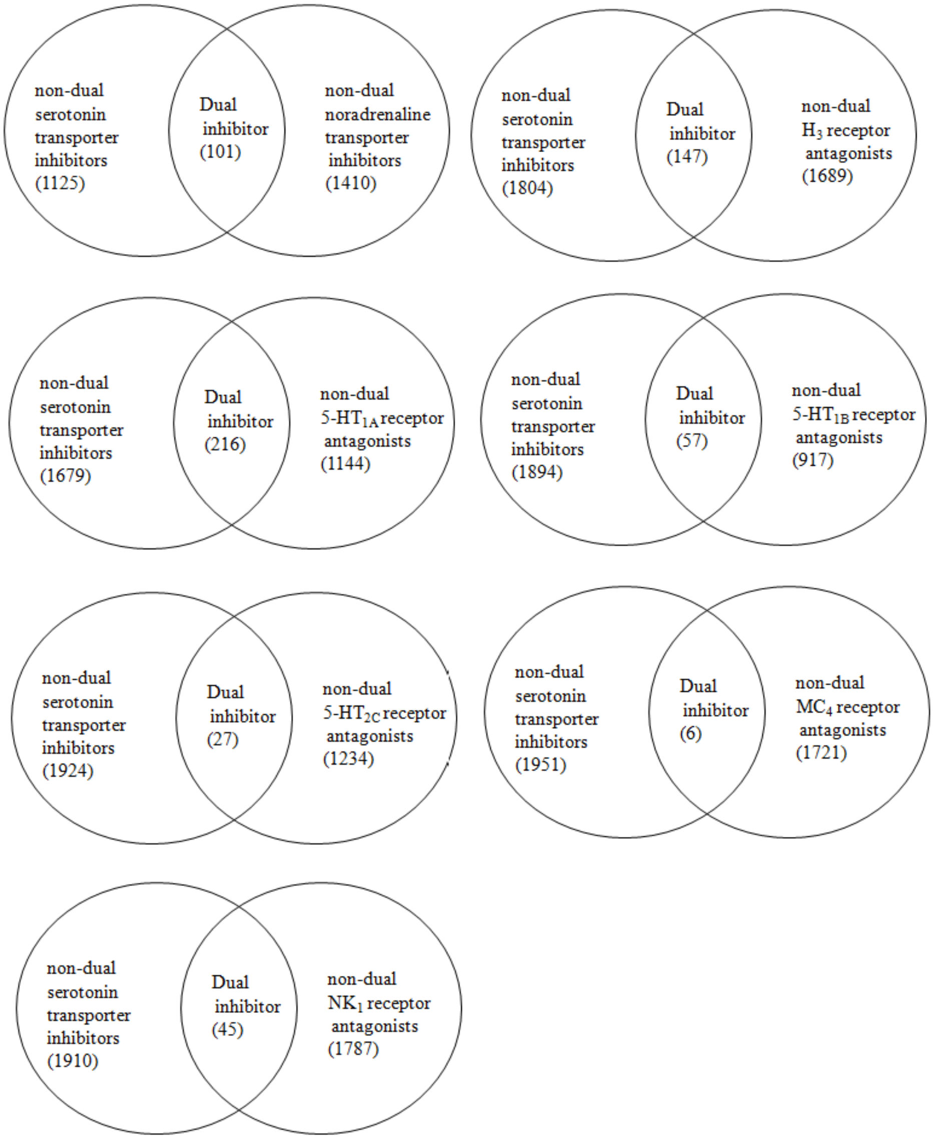

non-dual serotonin transporter inhibitors (1125)
Dual inhibitor (101)
non-dual noradrenaline transporter inhibitors (1410)
non-dual serotonin transporter inhibitors (1804)
Dual inhibitor (147)
non-dual H_3 receptor antagonists (1689)
non-dual serotonin transporter inhibitors (1679)
Dual inhibitor (216)
non-dual 5-HT_{1A} receptor antagonists (1144)
non-dual serotonin transporter inhibitors (1894)
Dual inhibitor (57)
non-dual 5-HT_{1B} receptor antagonists (917)
non-dual serotonin transporter inhibitors (1924)
Dual inhibitor (27)
non-dual 5-HT_{2C} receptor antagonists (1234)
non-dual serotonin transporter inhibitors (1951)
Dual inhibitor (6)
non-dual MC_4 receptor antagonists (1721)
non-dual serotonin transporter inhibitors (1910)
Dual inhibitor (45)
non-dual NK_1 receptor antagonists (1787)

Figure 5-2 The Venn graph of the collected 7 evaluated dual-inhibitors pairs and non-dual-inhibitors of the 8 evaluated targets

5.2.2 Computational models

SVM is based on the structural risk minimization principle of statistical learning theory (132). It consistently shows outstanding classification performance; It is less penalized by sample redundancy; It has lower risk for overfitting; It is capable of accommodating large and structurally diverse training and testing datasets, and is fast in performing classification tasks (134, 135). However, like all machine learning methods, the performance of SVM is critically dependent on the diversity of training datasets. Because of the limited knowledge of known inhibitors for many targets, sufficiently good SVM VS tools may not be readily developed for these targets. Nonetheless, SVM VS tools with comparable performances or partially improved performances in certain aspects (e.g. reduced false-hit rates at comparable inhibitor yield) are useful to complement other VS tools. The performance of SVM in predicting non-dual inhibitors was evaluated by 5-fold cross-validation test. For each target pair, non-dual inhibitors and non-inhibitors were randomly divided into 5 groups of approximately equal size, with 4 groups used for training a SVM VS tool and 1 group used for testing it, and the test process is repeated for all 5 possible compositions to derive an average VS performance. After the 5-fold cross-validation, the σ values are chosen in the range of 0.9-5 based on the average VS performance for the model development. **Table 5-2** shows the results of the 5-fold cross validation of SVM VS models for the target pairs SERT-NET, SERT-H_3, SERT-$5HT_{1A}$, SERT-$5HT_{1B}$, SERT-$5HT_{2C}$, SERT-MC_4 and SERT-NK_1. As for margin C, the SVM VS models were developed by using a hard margin c=100,000. A hard margin has been proven to provide well with a more sensitive and strict classification for unbalanced datasets in which the negative data outnumbered the positive ones (263)(36, 37) (43) (47). **Figure 5-3** illustrates the schematic diagram of COMBI-SVMs.

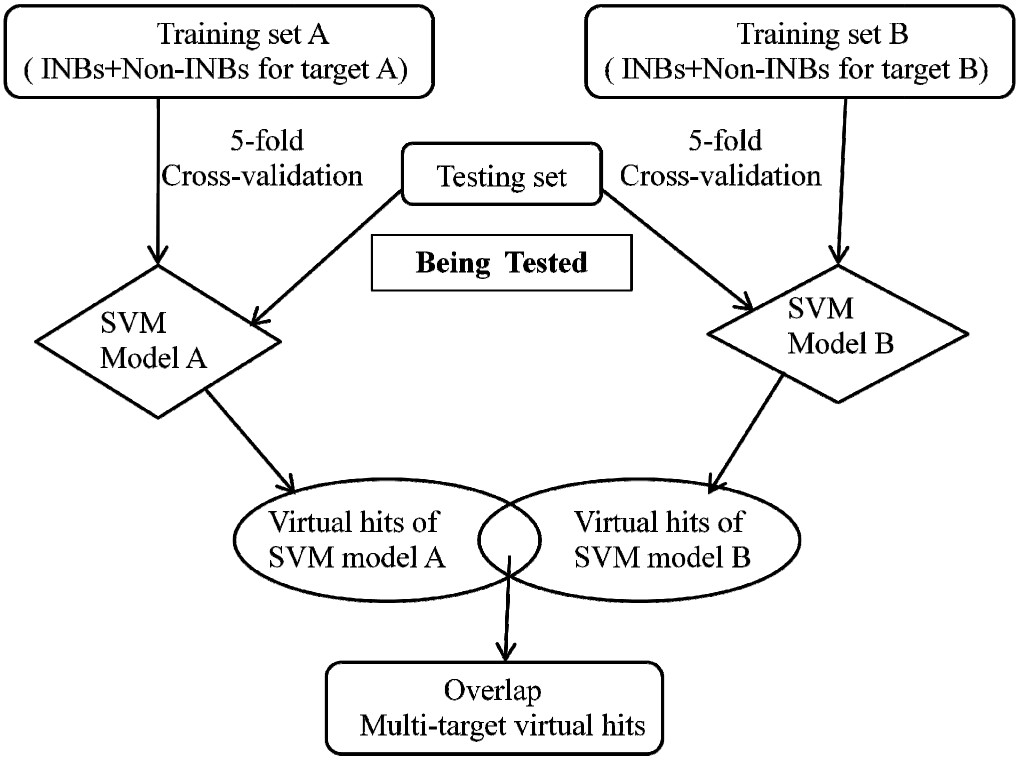
Table 5-2 5-fold cross-validation of SVM models for parameter selection and additional tests of these models for predicting dual-inhibitors and non-inhibitors; SEN: sensitivity, SPE: specificity, AC: overall accuracy, AVE: average; SD: standard deviation, SEM: standard error of means

Target Pair	**5-fold C.V.Performance for parameter selection**							**5-fold C.V. tests for dual and non-inhibitors**		
SERT-NET	**C.V. group**	**SERT**			**NET**			**NETSRIs**	**non-SSRIs**	**non-NRIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	84%	99.8%	99.5%	90%	99.8%	99.6%	48%	88%	81%
	2	91%	99.8%	99.7%	90%	99.7%	99.5%	48%	86%	77%
	3	89%	99.7%	99.5%	88%	99.7%	99.5%	45%	86%	81%
	4	85%	99.8%	99.5%	89%	99.7%	99.4%	43%	84%	83%
	5	87%	99.8%	99.6%	88%	99.8%	99.5%	48%	85%	82%
	AVE	**87%**	**99.8%**	**99.6%**	**89%**	**100%**	**99%**	**46%**	**86%**	**81%**
	S.D	**0.025**	**0.000**	**0.001**	**0.010**	**0.000**	**0.000**	**0.02**	**0.02**	**0.02**
	S.E.M	**0.011**	**0.000**	**0.000**	**0.004**	**0.000**	**0.000**	**0.01**	**0.01**	**0.01**
SERT-H_3	**C.V. group**	**SERT**			**H3**			**H3SRIs**	**non-SSRIs**	**non-H3Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	93%	99.5%	99.3%	92%	99.8%	99.6%	17%	85%	100%
	2	89%	99.6%	99.3%	93%	99.7%	99.5%	31%	80%	100%
	3	88%	99.6%	99.3%	93%	99.7%	99.5%	25%	77%	100%
	4	89%	99.6%	99.3%	93%	99.7%	99.5%	24%	84%	100%
	5	86%	99.5%	99.1%	92%	99.7%	99.5%	19%	87%	100%
	AVE	**89%**	**99.6%**	**99%**	**93%**	**99.7%**	**99.5%**	**23%**	**82%**	**100%**
	S.D	**0.025**	**0.001**	**0.001**	**0.006**	**0.000**	**0.000**	**0.06**	**0.04**	**1.00**
	S.E.M	**0.011**	**0.000**	**0.000**	**0.003**	**0.000**	**0.000**	**0.03**	**0.02**	**0.00**
SERT-$5HT_{1A}$	**C.V. group**	**SERT**			**5HT1a**			**5HT1aSRIs**	**non-SSRIs**	**non-5HT1aIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	99.7%	99.7%	99.4%	88%	99.8%	99.6%	48%	86%	74%
	2	99.6%	99.6%	99.3%	83%	99.8%	99.4%	45%	79%	74%

	3	99.5%	99.5%	99.2%	86%	99.7%	99.5%	44%	85%	74%
	4	99.7%	99.7%	99.3%	86%	99.8%	99.5%	45%	89%	77%
	5	99.7%	99.7%	99.4%	84%	99.8%	99.5%	45%	89%	82%
	AVE	**99.7%**	**99.7%**	**99%**	**85%**	**99.8%**	**99.5%**	**45%**	**86%**	**76%**
	S.D	**0.001**	**0.001**	**0.001**	**0.021**	**0.000**	**0.001**	**0.01**	**0.04**	**0.04**
	S.E.M	**0.000**	**0.000**	**0.000**	**0.009**	**0.000**	**0.000**	**0.01**	**0.02**	**0.02**
SERT-5HT$_{1B}$	**C.V. group**	**SERT**			**5HT1b**			**5HT1bSRIs**	**non-SSRIs**	**non-5HT1bIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	84%	99.98%	99.7%	85%	99.8%	99.6%	23%	98%	99%
	2	82%	99.98%	99.7%	85%	99.8%	99.6%	19%	98%	91%
	3	78%	99.95%	99.5%	80%	99.8%	99.6%	23%	98%	96%
	4	81%	99.96%	99.6%	85%	99.8%	99.6%	23%	97%	92%
	5	80%	99.97%	99.6%	83%	99.9%	99.6%	23%	97%	93%
	AVE	**81%**	**99.97%**	**99.6%**	**84%**	**99.8%**	**99.6%**	**22%**	**98%**	**94%**
	S.D	**0.024**	**0.0001**	**0.001**	**0.023**	**0.000**	**0.000**	**0.018**	**0.004**	**0.03**
	S.E.M	**0.011**	**0.00005**	**0.0003**	**0.010**	**0.000**	**0.000**	**0.008**	**0.002**	**0.02**
SERT-5HT$_{2C}$	**C.V. group**	**SERT**			**5HT2c**			**5HT2cSRIs**	**non-SSRIs**	**non-5HT2cIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	86.8%	99.5%	99.2%	91.4%	99.7%	99.3%	22%	75%	89%
	2	89.4%	99.6%	99.3%	90.6%	99.7%	99.3%	26%	84%	88%
	3	90.1%	99.6%	99.3%	90.6%	99.7%	99.4%	15%	81%	88%
	4	88.8%	99.6%	99.3%	94.1%	99.8%	99.3%	15%	82%	88%
	5	92.4%	99.6%	99.4%	98.0%	99.8%	99.4%	15%	81%	95%
	AVE	**90%**	**99.6%**	**99%**	**92.9%**	**99.7%**	**99.3%**	**20%**	**81%**	**90%**
	S.D	**0.021**	**0.000**	**0.001**	**0.032**	**0.0006**	**0.0003**	**0.05**	**0.03**	**0.028**
	S.E.M	**0.009**	**0.000**	**0.000**	**0.014**	**0.0003**	**0.0002**	**0.02**	**0.01**	**0.013**
SERT-MC$_4$	**C.V. group**	**SERT**			**MC4**			**MC4SRIs**	**non-SSRIs**	**non-MC4Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**

	1	89.3%	99.5%	99.2%	97.4%	100.0%	99.9%	83%	94%	95%
	2	92.6%	99.5%	99.3%	98.3%	99.9%	99.9%	83%	93%	91%
	3	92.1%	99.5%	99.3%	98.3%	99.9%	99.9%	67%	91%	92%
	4	92.8%	99.6%	99.4%	98.3%	99.9%	99.9%	67%	91%	94%
	5	87.7%	99.6%	99.2%	97.1%	99.9%	99.8%	67%	91%	93%
	AVE	**91%**	**99.5%**	**99%**	**98%**	**99.9%**	**99.9%**	**73%**	**92%**	**93%**
	S.D	**0.023**	**0.000**	**0.001**	**0.006**	**0.000**	**0.000**	**0.09**	**0.01**	**0.02**
	S.E.M	**0.010**	**0.000**	**0.000**	**0.003**	**0.000**	**0.000**	**0.04**	**0.01**	**0.01**
SERT-NK$_1$	**C.V. group**	**SERT**			**NK1**			**NK1SRIs**	**non-SSRIs**	**non-NK1Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	87.4%	99.5%	99.1%	93.6%	99.8%	93.6%	36%	88%	91%
	2	89.8%	99.6%	99.3%	95.5%	99.8%	95.5%	40%	88%	93%
	3	89.5%	99.5%	99.2%	96.6%	99.8%	96.6%	40%	90%	93%
	4	91.4%	99.5%	99.2%	95.0%	99.8%	95.0%	38%	86%	93%
	5	87.7%	99.6%	99.3%	95.2%	99.8%	95.2%	38%	88%	93%
	AVE	**89%**	**99.5%**	**99%**	**95%**	**99.8%**	**95.2%**	**38%**	**88%**	**93%**
	S.D	**0.016**	**0.001**	**0.001**	**0.011**	**0.000**	**0.011**	**0.02**	**0.02**	**0.01**
	S.E.M	**0.007**	**0.000**	**0.000**	**0.005**	**0.000**	**0.005**	**0.01**	**0.01**	**0.00**

Figure 5-3 The COMBI-SVMs diagram. Individual SVM models are built after 5-fold cross-validation where the training set is randomly divided into 5 sub sets and in turns 4 sets are used for training and 1 set for testing to choose the best parameters for model construction. Virtual hits simultaneously selected by all individual SVM models are considered as multi-target virtual hits. INBs=inhibitors; Non-INBs=non-inhibitors.

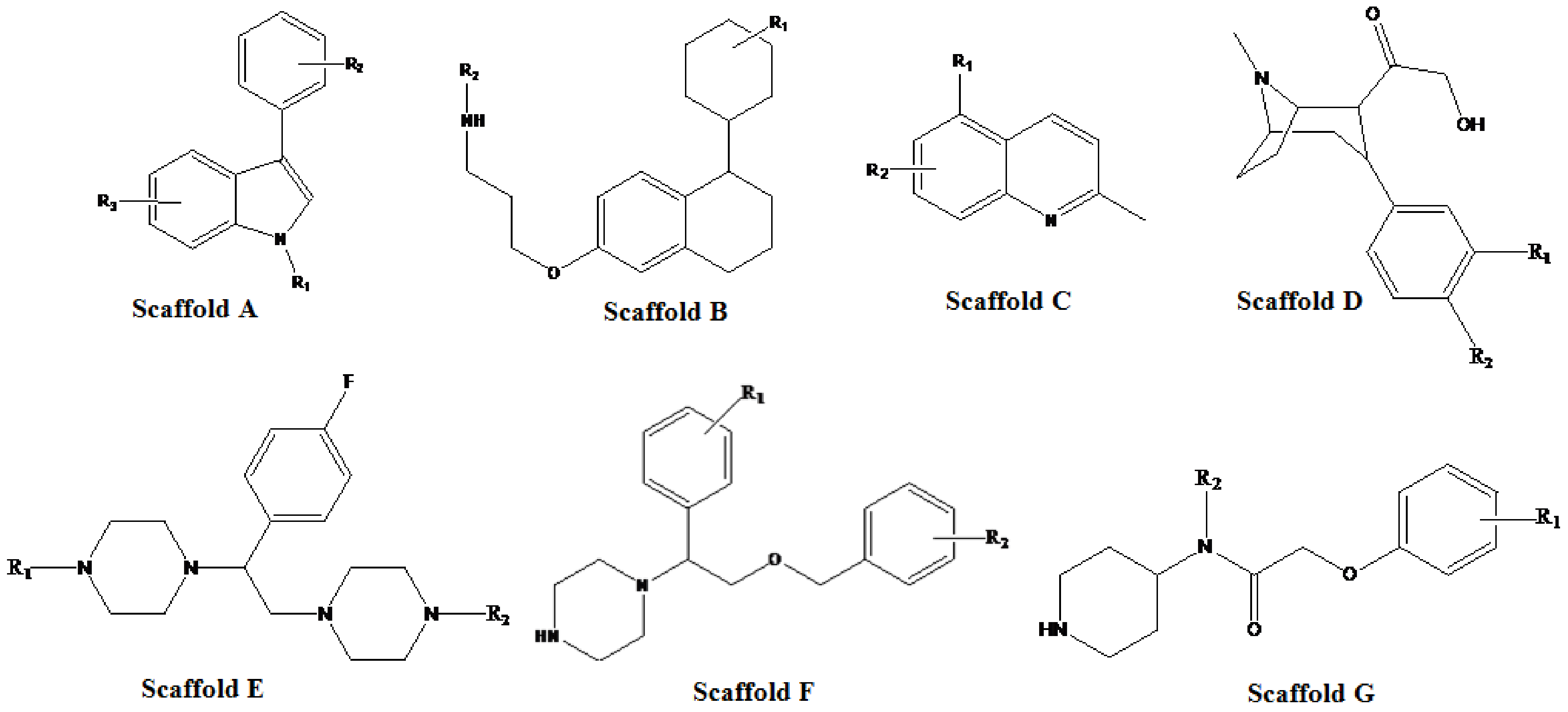


5.3 Results and Discussion

5.3.1 Individual target inhibitors and dual inhibitors of the studied target pairs

As shown in **Table 5-1**, high percentages of the known dual inhibitors of the seven studied target pairs are distributed in the compound families containing individual target inhibitor of at least one target in the target pair. Only 18.4%-37.0% of the known dual inhibitors are not in the compound families of the known individual target inhibitors. Nonetheless, dual inhibitors have some features distinguished from those of individual target inhibitors, which are partly exhibited from the top-ranked scaffolds contained in higher percentages of dual inhibitors of the studied target pairs (**Figure 5-3**). **Table 5-3** gives the distribution of some of these scaffolds in the dual inhibitors of the studied target pairs and inhibitors of individual targets of these target pairs. Scaffolds A, B, C, D, E, F and G are contained in high percentages of dual inhibitors. Specifically, scaffold A is contained in 21.8% of the 101 NETSRIs, scaffold B in 17.7% of the 147 H3SRIs, scaffold C in 14.8% of the 216 5HT1aSRIs, scaffold D in 14.8% of the 27 5HT2cSRIs, scaffold E in 100% of the 6 MC4SRIs, and scaffold F and G in 44.4% and 33.3% of the 45 NK1SRIs, whereas these scaffolds are contained in single-digit percentages or less of the inhibitors of other target pairs and the individual target inhibitors of the specific target-pairs. Known 5HT1bSRIs appear to be distributed in many scaffolds each containing no more than three compounds. Nonetheless, some specific variations of side-chain groups of these and other scaffolds found in the known 5HT1bSRIs as well as known NETSRIs, H3SRIs and 5HT1aSRIs appear to be sufficient to convert individual target inhibitors into dual inhibitors. Moreover, physicochemical properties as well as structural features are also important for distinguishing individual target inhibitors and dual inhibitors.

Figure 5-4 Top-ranked molecular scaffolds primarily found in known multi-target serotonin reuptake inhibitors; Scaffold A, B and C are distributed in significantly higher percentage of known multi-target NETSRIs, H3SRIs, and 5HT1aSRIs than known “individual-target”



Scaffold A

Scaffold B

Scaffold C

Scaffold D

Scaffold E

Scaffold F

Scaffold G

Table 5-3 Distribution of the top-ranked scaffolds in multi-target inhibitors of the 7 target pairs SERT-NET, SERT-H_3, SERT-$5HT_{1A}$, SERT-$5HT_{1B}$, SERT-$5HT_{2C}$, SERT-MC_4 and SERT-NK_1

Target pair	Datasets	Percent of inhibitors in dataset						
		containing scaffold						
		Scaffold A	**Scaffold B**	**Scaffold C**	**Scaffold D**	**Scaffold E**	**Scaffold F**	**Scaffold G**
SERT-NET	Multi-target NETSRIs	21.8(22/101)	0(0/101)	3(3/101)	1.0(1/101)	0(0/101)	0(0/101)	0(0/101)
	NET reuptake inhibitors inactive against SERT	0.07(1/1410)	0(0/1410)	0(0/1410)	3.0(43/1410)	0.6(8/1410)	0(0/1410)	0(0/1410)
	SERT reuptake inhibitors inactive against NET	2(23/1125)	2.1(24/1125)	1.7(24/1125)	0.4(5/1125)	0(0/1125)	0.2(2/1125)	1.3(15/1125)
SERT-H_3	Multi-target H3SRIs	0(0/147)	17.7(26/147)	0(0/147)	0(0/147)	0(0/147)	0(0/147)	0(0/147)
	H_3 receptor antagonists inactive against SERT	0(0/1689)	0(0/1689)	0.4(6/1689)	0(0/1689)	0(0/1689)	0(0/1689)	0(0/1689)
	SERT reuptake inhibitors inactive against H_3 receptors	1.5(27/1804)	0(0/1804)	1.3(24/1804)	1.8(32/1804)	0(0/1804)	1.0(18/1804)	0.9(16/1804)
SERT-$5HT_{1A}$	Multi-target 5HT1aSRIs	0(0/216)	0(0/216)	14.8(32/216)	0(0/216)	0(0/216)	0(0/216)	0(0/216)
	$5HT_{1A}$ receptor antagonists inactive against SERT	4.8(55/1144)	0(0/1144)	1.4(16/1144)	0(0/1144)	0(0/1144)	0(0/1144)	0(0/1144)
	SERT reuptake inhibitors inactive against $5HT_{1A}$ receptors	1.3(21/1679)	1.5(26/1678)	0.8(13/1679)	1.8(31/1679)	0(0/1679)	1.1(18/1679)	0.9(15/1679)
SERT-$5HT_{1B}$	Multi-target 5HT1bSRIs	1.8(1/57)	0(0/57)	7(4/57)	5.3(3/57)	0(0/57)	0(0/57)	0(0/57)
	$5HT_{1B}$ receptor antagonists inactive against SERT	0(0/917)	0(0/917)	0.3(3/917)	0(0/917)	0(0/917)	0(0/917)	0(0/917)
	SERT reuptake inhibitors inactive against $5HT_{1B}$ receptors	1.4(26/1894)	1.4(26/1894)	1.1(20/1894)	1.4(26/1894)	0(0/1894)	1.0(18/1894)	0.8(15/1894)
SERT-$5HT_{2C}$	Multi-target 5HT2cSRIs	3.7(1/27)	0(0/27)	3.7(1/27)	14.8(4/27)	0(0/27)	0(0/27)	0(0/27)
	$5HT_{2C}$ receptor antagonists inactive against SERT	1.6(20/1234)	0(0/1234)	(3/1234)	0(0/1234)	0(0/1234)	0(0/1234)	0(0/1234)
	SERT reuptake inhibitors inactive against $5HT_{2C}$ receptors	0(0/1924)	0(0/1924)	1.2(23/1924)	1.5(29/1924)	0(0/1924)	0.9(18/1924)	0.7(13/1924)
SERT-MC_4	Multi-target MC4SRIs	0(0/6)	0(0/6)	0(0/6)	0(0/6)	100%(6/6)	0(0/6)	0(0/6)
	MC_4 receptor antagonists inactive against SERT	0(0/1721)	0(0/1721)	0(0/1721)	0(0/1721)	2.5(43/1721)	0(0/1721)	0(0/1721)
	SERT reuptake inhibitors inactive against MC_4 receptors	0(0/1951)	0(0/1951)	1.2(23/1951)	1.6(31/1951)	0(0/1951)	0.9(18/1951)	0.8(15/1951)
SERT-NK_1	Multi-target NK1SRIs	0(0/45)	0(0/45)	0(0/45)	0(0/45)	0(0/45)	44.4(20/45)	33.3(15/45)
	NK1 receptor antagonists inactive against SERT	0.2(4/1787)	0(0/1787)	0(0/1787)	0(0/1787)	0(0/1787)	0.06(1/1787)	0(0/1787)
	SERT reuptake inhibitors inactive against NK_1 receptors	0.1(2/1910)	0(0/1910)	(20/1910)	1.0(33/1910)	0(0/1910)	0.9(18/1910)	0.6(11/1910)

5.3.2 5-fold cross-validation tests of SVM, k-NN and PNN models

The parameters of the SVM, k-NN and PNN models were determined by 5-fold cross- validation studies of individual target inhibitors and putative non-inhibitors of each target pair. Additionally, each 5-fold cross-validation model was tested by dual target NETSRIs, H3SRIs, 5HT1aSRIs, 5HT1bSRIs, 5HT2cSRIs, MC4SRIs and NK1SRIs and real non-inhibitors of the individual target of each target pair. Non-inhibitors of a target refer to compounds with IC50 or K_i value >20μM. The results of these tests for SVM, k-NN and PNN are shown in **Table 5-3, 5-4 and 5-5** respectively. The 5-fold cross-validation tests were measured by sensitivity, specificity and over all accuracy given as TP/(TP+FN), TN/(TN+FP) and TP+TN/(TP+TN+FP+FN) respectively in terms of the numbers of true positives TP (true inhibitors), true negatives TN (true non-inhibitors), false positives FP (false inhibitors), and false negatives FN (false non-inhibitors). Overall, the sensitivity of SVM, k-NN and PNN is in the range of 78.0%-99.8%, 79%-99.7% and 89%-99.7%, the specificity in the range of 99.4%-99.98%, 99%-99.98%, and 95.1%-99.4%, and overall accuracy in the range of 93.6%-99.6%, 99.0%-99.98%, and 96.5%-99.3% respectively. The dual inhibitor accuracy of SVM, k-NN and PNN are in the range of 15%-83%, 10%-83%, and 17%-58% respectively. The non-inhibitor prediction accuracy of SVM, k-NN and PNN are in the range of 73%-100%, 62%-97% and 72%-89% respectively. Therefore, SVM showed comparable overall performance in these 5-fold cross-validation tests.

Table 5-4 5-fold cross-validation of k-NN models for parameter selection and additional tests of these models for predicting dual-inhibitors and non-inhibitors, SEN sensitivity, SPE specificity, AC overall accuracy, AVE average; SD standard deviation, and SEM standard error of means.

Target Pair	**5-fold C.V. for parameter selection**							**5-fold C.V. tests for dual and non-inhibitors**		
SERT-NET	**C.V. group**	**SERT**			**NET**			**NETSRIs**	**non-SSRIs**	**non-NRIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	85%	99.60%	99%	89%	99.60%	99.30%	40%	83%	62%
	2	84%	99.60%	99%	88%	99.40%	99.20%	38%	89%	84%
	3	88%	99.60%	99%	86%	99.60%	99.30%	37%	89%	81%
	4	89%	99.60%	99%	87%	99.60%	99.30%	36%	80%	83%
	5	87%	99.60%	99%	91%	99.50%	99.30%	39%	85%	82%
	AVE	**87%**	**99.60%**	**99%**	**88%**	**99.50%**	**99%**	**38%**	**85%**	**78%**
	S.D	**0.023**	**0**	**0**	**0.018**	**0.001**	**0.001**	**0.02**	**0.04**	**0.09**
	S.E.M	**0.01**	**0**	**0**	**0.008**	**0**	**0**	**0.01**	**0.02**	**0.04**
SERT-H_3	**C.V. group**	**SERT**			**H3**			**H3SRIs**	**non-SSRIs**	**non-H3Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	91%	99.50%	99.20%	93%	99.50%	99.40%	10%	85%	88%
	2	92%	99.40%	99.20%	92%	99.60%	99.40%	14%	80%	83%
	3	87%	99.30%	99.00%	92%	99.60%	99.40%	13%	77%	83%
	4	90%	99.50%	99.30%	93%	99.60%	99.50%	10%	84%	88%
	5	91%	99.40%	99.10%	91%	99.50%	99.30%	12%	87%	88%
	AVE	**90%**	**99%**	**99%**	**92%**	**99.60%**	**99%**	**12%**	**82%**	**86%**
	S.D	**0.02**	**0.001**	**0.001**	**0.008**	**0.001**	**0.001**	**0.02**	**0.04**	**0.03**
	S.E.M	**0.009**	**0**	**0.001**	**0.003**	**0**	**0**	**0.01**	**0.02**	**0.01**
SERT-$5HT_{1A}$	**C.V. group**	**SERT**			**5HT1a**			**5HT1aSRIs**	**non-SSRIs**	**non-5HT1aIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	89.30%	99.50%	99.20%	85%	99.60%	99.40%	32%	83%	79%

	2	89.30%	99.40%	99.10%	78%	99.50%	99.20%	33%	89%	81%
	3	90.80%	99.40%	99.20%	84%	99.60%	99.30%	34%	80%	82%
	4	93.80%	99.40%	99.30%	85%	99.60%	99.30%	36%	80%	78%
	5	89.60%	99.60%	99.30%	84%	99.60%	99.30%	35%	79%	82%
	AVE	**91%**	**99%**	**99%**	**83%**	**99.60%**	**99%**	**34%**	**82%**	**80%**
	S.D	**0.019**	**0.001**	**0.001**	**0.029**	**0**	**0.001**	**0.01**	**0.04**	**0.02**
	S.E.M	**0.009**	**0**	**0**	**0.013**	**0**	**0**	**0.01**	**0.02**	**0.01**
SERT-5HT$_{1B}$	**C.V.**	**SERT**			**5HT1b**			**5HT1bSRIs**	**non-SSRIs**	**non-5HT1bIs**
	group	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	91%	99.50%	99.20%	84%	99.70%	99.50%	30%	82%	83%
	2	92%	99.30%	99.10%	86%	99.80%	99.60%	30%	86%	84%
	3	91%	99.40%	99.10%	86%	99.70%	99.50%	25%	82%	84%
	4	91%	99.40%	99.10%	82%	99.60%	99.40%	30%	88%	87%
	5	92%	99.30%	99.10%	83%	99.70%	99.40%	28%	88%	84%
	Average	**91%**	**99%**	**99%**	**84%**	**99.70%**	**99%**	**28%**	**85%**	**84%**
	S.D	**0.006**	**0.001**	**0**	**0.019**	**0.001**	**0.001**	**0.02**	**0.03**	**0.01**
	S.E.M	**0.003**	**0**	**0**	**0.008**	**0**	**0**	**0.01**	**0.01**	**0.01**
SERT-5HT$_{2C}$	**C.V.**	**SERT**			**5HT2c**			**5HT2cSRIs**	**non-SSRIs**	**non-5HT2cIs**
	group	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	92.2%	99.4%	99.2%	86%	99%	99.2%	26%	88%	87%
	2	90.1%	99.3%	99.1%	85%	99.6%	99.3%	26%	87%	83%
	3	90.9%	99.3%	99.1%	82%	99%	99.1%	26%	86%	84%
	4	91.9%	99.3%	99.1%	79%	99.5%	99.2%	26%	88%	82%
	5	90.9%	99.4%	99.2%	81%	99.5%	99.2%	26%	87%	82%
	AVE	**83%**	**99.5%**	**99%**	**83%**	**99.5%**	**99.2%**	**26%**	**87%**	**84%**
	S.D	**0.028**	**0.000**	**0.000**	**0.028**	**0.000**	**0.001**	**0.00**	**0.01**	**0.02**
	S.E.M	**0.012**	**0.000**	**0.000**	**0.012**	**0.000**	**0.000**	**0.00**	**0.00**	**0.01**
SERT-MC$_4$	**C.V.**	**SERT**			**MC4**			**MC4SRIs**	**non-SSRIs**	**non-MC4Is**
	group	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	89.8%	99.3%	99.0%	99.7%	99.98%	99.98%	17%	84%	95%
	2	91.3%	99.3%	99.1%	98.3%	99.7%	99.7%	17%	81%	90%
	3	92.8%	99.5%	99.3%	97.7%	99.8%	99.7%	17%	86%	90%

	4	88.7%	99.4%	99.1%	98.0%	99.7%	99.7%	83%	84%	95%
	5	91.5%	99.4%	99.1%	99.1%	99.8%	99.7%	17%	83%	97%
	AVE	**91%**	**99.4%**	**99%**	**98.5%**	**99.8%**	**99.8%**	**30.2%**	**84%**	**93.4%**
	S.D	**0.016**	**0.001**	**0.001**	**0.008**	**0.001**	**0.001**	**0.30**	**0.02**	**0.03**
	S.E.M	**0.007**	**0.000**	**0.000**	**0.004**	**0.0005**	**0.001**	**0.132**	**0.01**	**0.01**
SERT-NK$_1$	**C.V.**	**SERT**			**NK1**			**NK1SRIs**	**non-SSRIs**	**non-NK1Is**
	group	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	91.4%	99.3%	99.1%	95.5%	99.7%	99.6%	20%	88%	81%
	2	92.7%	99.4%	99.2%	95.0%	99.5%	99.4%	24%	89%	82%
	3	91.6%	99.3%	99.1%	95.0%	99.6%	99.4%	22%	88%	85%
	4	90.3%	99.5%	99.2%	95.8%	99.6%	99.5%	20%	87%	85%
	5	89.3%	99.4%	99.1%	95.2%	99.5%	99.4%	20%	87%	85%
	AVE	**91%**	**99.4%**	**99%**	**95%**	**99.6%**	**99%**	**21%**	**88%**	**84%**
	S.D	**0.013**	**0.001**	**0.001**	**0.004**	**0.001**	**0.001**	**0.02**	**0.01**	**0.02**
	S.E.M	**0.006**	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**	**0.01**	**0.00**	**0.01**

Table 5-5 5-fold cross-validation of PNN models for parameter selection and additional tests of these models for predicting dual-inhibitors and non-inhibitors, SEN sensitivity, SPE specificity, AC overall accuracy, AVE average; SD standard deviation, and SEM standard error of means.

Target Pair	5-fold C.V.Performance for parameter selection							5-fold C.V. tests for dual and non-inhibitors		
SERT-NET	C.V. group	SERT			NET			NETSRIs	non-SSRIs	non-NRIs
		SEN	SPE	AC	SEN	SPE	AC	SEN	SPE	SPE
	1	94%	97.50%	97.50%	95%	97.00%	96.90%	58%	86%	80%
	2	94%	97.60%	97.50%	96%	96.60%	96.60%	57%	88%	79%
	3	95%	97.50%	97.50%	94%	97.00%	96.90%	50%	85%	76%
	4	94%	97.40%	97.30%	94%	96.90%	96.80%	55%	84%	72%
	5	93%	97.70%	97.60%	97%	96.90%	96.90%	55%	83%	76%
	AVE	**94%**	**98%**	**98%**	**95%**	**97%**	**97%**	**55%**	**85%**	**77%**
	S.D	**0.005**	**0.001**	**0.001**	**0.013**	**0.002**	**0.002**	**0.03**	**0.02**	**0.03**
	S.E.M	**0.002**	**0**	**0**	**0.006**	**0.001**	**0.001**	**0.01**	**0.01**	**0.01**
SERT-H_3	**C.V. group**	**SERT**			**H3**			**H3SRIs**	**non-SSRIs**	**non-H3Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	96%	96.80%	96.80%	98%	97.50%	97.50%	40%	80%	90%
	2	96%	96.80%	96.80%	97%	97.70%	97.70%	39%	84%	84%
	3	94%	97.00%	96.90%	96%	97.80%	97.80%	41%	83%	84%
	4	96%	96.90%	96.90%	98%	97.70%	97.70%	34%	83%	84%
	5	96%	96.70%	96.60%	95%	97.90%	97.80%	37%	83%	84%
	AVE	**96%**	**97%**	**97%**	**97%**	**98%**	**98%**	**38%**	**83%**	**85%**
	S.D	**0.011**	**0.001**	**0.001**	**0.014**	**0.002**	**0.001**	**0.03**	**0.01**	**0.03**
	S.E.M	**0.005**	**0.001**	**0**	**0.006**	**0.001**	**0.001**	**0.01**	**0.01**	**0.01**
SERT-$5HT_{1A}$	**C.V. group**	**SERT**			**5HT1a**			**5HT1aSRIs**	**non-SSRIs**	**non-5HT1aIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	97.30%	97.00%	97.00%	93%	97.80%	97.70%	46%	85%	72%
	2	94.90%	96.60%	96.60%	91%	97.70%	97.60%	48%	82%	73%

	3	96.40%	96.60%	96.60%	93%	97.50%	97.40%	45%	82%	71%
	4	97.30%	96.70%	96.70%	92%	97.40%	97.30%	49%	84%	72%
	5	97.30%	96.70%	96.70%	92%	97.70%	97.60%	46%	83%	73%
	AVE	**97%**	**97%**	**97%**	**92%**	**98%**	**98%**	**47%**	**83%**	**72%**
	S.D	**0.01**	**0.002**	**0.002**	**0.009**	**0.002**	**0.002**	**0.02**	**0.02**	**0.01**
	S.E.M	**0.005**	**0.001**	**0.001**	**0.004**	**0.001**	**0.001**	**0.01**	**0.01**	**0**
SERT-$5HT_{1B}$	**C.V. group**	**SERT**			**5HT1b**			**5HT1bSRIs**	**non-SSRIs**	**non-5HT1bIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	98%	97.00%	97.10%	91%	98.00%	97.90%	44%	78%	83%
	2	96%	96.90%	96.90%	92%	98.40%	98.30%	35%	77%	82%
	3	96%	96.90%	96.90%	92%	98.40%	98.30%	39%	75%	82%
	4	96%	96.60%	96.50%	91%	98.20%	98.10%	47%	75%	83%
	5	97%	97.00%	97.00%	89%	98.20%	98.10%	46%	75%	81%
	AVE	**97%**	**97%**	**97%**	**91%**	**98%**	**98%**	**42%**	**76%**	**82%**
	S.D	**0.011**	**0.002**	**0.002**	**0.016**	**0.002**	**0.002**	**0.05**	**0.01**	**0.01**
	S.E.M	**0.005**	**0.001**	**0.001**	**0.007**	**0.001**	**0.001**	**0.02**	**0.01**	**0**
SERT-$5HT_{2C}$	**C.V. group**	**SERT**			**5HT2c**			**5HT2cSRIs**	**non-SSRIs**	**non-5HT2cIs**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	96.1%	97.0%	97.0%	91%	97%	97.4%	33%	84%	84%
	2	98.7%	97.0%	97.0%	89%	97.2%	97.0%	33%	83%	85%
	3	95.8%	96.7%	96.7%	93%	97%	97.3%	30%	83%	84%
	4	98.2%	97.0%	97.0%	91%	97.0%	96.9%	30%	82%	87%
	5	96.4%	96.8%	96.8%	93%	97.1%	97.0%	33%	83%	84%
	AVE	**91%**	**97.2%**	**97%**	**91%**	**97.2%**	**97.1%**	**32%**	**83%**	**85%**
	S.D	**0.016**	**0.002**	**0.001**	**0.016**	**0.002**	**0.002**	**0.02**	**0.01**	**0.01**
	S.E.M	**0.007**	**0.001**	**0.001**	**0.007**	**0.001**	**0.001**	**0.01**	**0.00**	**0.00**
SERT-MC_4	**C.V. group**	**SERT**			**MC4**			**MC4SRIs**	**non-SSRIs**	**non-MC4Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	96.2%	96.7%	96.7%	96.5%	99.4%	99.3%	33%	82%	89%

	2	95.6%	97.0%	97.0%	98.5%	99.2%	99.2%	17%	82%	78%
	3	96.4%	97.0%	97.0%	98.5%	99.3%	99.3%	17%	81%	79%
	4	97.2%	96.4%	96.5%	98.8%	99.2%	99.1%	33%	82%	81%
	5	95.1%	97.1%	97.0%	99.7%	99.2%	99.3%	33%	82%	77%
	AVE	**96%**	**96.9%**	**97%**	**98.4%**	**99.3%**	**99.2%**	**27%**	**82%**	**81%**
	S.D	**0.008**	**0.003**	**0.002**	**0.011**	**0.0007**	**0.0006**	**0.09**	**0.00**	**0.05**
	S.E.M	**0.003**	**0.001**	**0.001**	**0.005**	**0.0003**	**0.0002**	**0.039**	**0.00**	**0.02**
SERT-NK$_1$	**C.V. group**	**SERT**			**NK1**			**NK1SRIs**	**non-SSRIs**	**non-NK1Is**
		SEN	**SPE**	**AC**	**SEN**	**SPE**	**AC**	**SEN**	**SPE**	**SPE**
	1	96.1%	96.8%	96.8%	97.2%	98.6%	98.6%	29%	83%	89%
	2	95.5%	97.1%	97.1%	96.9%	98.7%	98.6%	33%	83%	85%
	3	97.1%	96.7%	96.7%	98.6%	98.6%	98.6%	36%	83%	86%
	4	96.3%	96.8%	96.8%	97.8%	98.5%	98.5%	33%	82%	85%
	5	96.1%	96.9%	96.8%	98.0%	98.6%	98.6%	33%	83%	83%
	AVE	**96%**	**96.9%**	**97%**	**98%**	**98.6%**	**99%**	**33%**	**83%**	**86%**
	S.D	**0.006**	**0.002**	**0.001**	**0.007**	**0.001**	**0.001**	**0.02**	**0.00**	**0.02**
	S.E.M	**0.003**	**0.001**	**0.001**	**0.003**	**0.000**	**0.000**	**0.01**	**0.00**	**0.01**

5.3.3 Virtual screening performance of Combinatorial SVM in searching multi-target serotonin inhibitors from large compound libraries

The VS performance of COMBI-SVM in identifying dual inhibitors of the seven target-pairs is summarized in **Table 5-6** together with the similarity level (sequence identity) between the drug-binding domains of each pair. Rost has found that proteins with >40% sequence identity unambiguously distinguish similar and non-similar structures and the signal gets blurred in the twilight zone of 20–35% sequence identity (271). Thus, target-pairs can be classified into high, intermediate, and low similarity classes with their drug-binding domains at sequence identity levels of >40%, 20%-40% and <20% respectively. Based on this criterion, SERT-NET (72.3% sequence identity) is of high similarity, and the other six target-pairs (1.7%~15.1% sequence identity) are of low sequence similarity.

In terms of the numbers of true positives TP (true inhibitors), true negatives TN (true non-inhibitors), false positives FP (false inhibitors), and false negatives FN (false non-inhibitors), the yield and false-hit rate are given by TP/(TP+FN) and FP/(TP+FP) respectively. The dual inhibitor yields are 49.5% for NETSRIs, 25.9% for H3SRIs, 47.7% for 5HT1aSRIs, and 22.8% for 5HT1bSRIs, 22.0% for 5HT2cSRIs, 83.3% for MC4SRIs and 31.1% for NK1SRIs respectively. Therefore, COMBI-SVMs showed reasonably good capability in identifying dual inhibitors of the seven evaluated target pairs without explicit knowledge of dual inhibitors. Target selectivity was tested by using COMBI-SVM to screen the 917-1951 individual-target inhibitors of each target-pair, which misidentified 22.4% and 29.8% of the individual target inhibitors as dual inhibitors for the SERT-NET pair, 5.4% and 8.2% for SERT-H_3, 15.4% and 19.4% for SERT-$5HT_{1A}$, 13.8% and 12.3% for SERT-$5HT_{1B}$, 14.2% and 12.4% for SERT-$5HT_{2C}$, 2.2% and 8.0%

for SERT-MC_4 and 4.2% and 6.3% for SERT-NK_1 respectively. Therefore, COMBI-SVM is reasonably selective in distinguishing multi-target inhibitors from individual-target inhibitors of the same target pair.

The misidentification of a substantial percentage of individual target inhibitors as dual inhibitors could have been caused by the similar reasons discussed in Chapter 4.3.1.

Target selectivity was further tested by using COMBI-SVM to screen the 8110-8688 (**Table 5-1**) inhibitors of the other six targets outside a given target-pair with the results summarised in **Table 5-6**. We found that 2.4%, 3.5%, 7.1%, 0.95%, 4.0%, 0.58%, and 1.16% of the inhibitors of the other six targets were misclassified as NETSRIs, H3SRIs, 5HT1aSRIs, 5HT1bSRIs, 5HT2cSRIs, MC4SRIs and NK1SRIs respectively. These data showed that COMBI-SVM is fairly selective in separating multi-target inhibitors of specific target pair from antidepressant inhibitors of other targets outside the target pair.

Virtual hit rates and false-hit rates of COMBI-SVM in screening compounds that resemble the structural and physicochemical properties of the training datasets were evaluated by using 7-8181 MDDR compounds (**Table 5-1**) similar to a multi-target inhibitor of each target pair. Similarity was defined by Tanimoto similarity coefficient ≥0.9 between a MDDR compound and its closest dual inhibitor (102). As shown in **Table 5-6,** COMBI-SVM identified 81, 3, 256, 249, 66, 1 and 1 virtual-hit(s) from 8181, 1486, 7349, 7475, 1302, 7 and 275 MDDR compounds similar to NETSRI, H3SRI, 5HT1aSRI, 5HT1bSRI, 5HT2cSRI, MC4RI and NK1SRI respectively. Disregarding the target-pairs with only 1 MDDR virtual-hits (which is statistically less meaningful for estimating virtual hit rates), the virtual hit rates in selecting MDDR compounds similar to the dual-inhibitors are in the range of 0.2%~5.1%. As majority of the MDDR compounds similar to the known dual inhibitors are expected to be non-inhibitors for the target pairs, these virtual hit rates can be considered as the upper limit of the false-hit rates.

Significantly lower virtual hit rates and thus false-hit rates were found in screening large libraries of 168,000 MDDR and 17 million PubChem compounds. As shown in **Table 5-6,** the numbers of multi-target virtual hits (virtual hit rate) in screening 168,000 MDDR compounds are 201 (0.12%) for NETSRIs, 112 (0.067%) for H3SRIs, 464 (0.28%) for 5HT1aSRIs, 241 (0.14%) for 5HT1bSRIs, 353 (0.21%) for 5HT2cSRIs, 70 (0.042%) for MC4SRIs and 92 (0.055%) for NK1SRIs respectively. The numbers of multi-target virtual hits (virtual hit rate) in screening 17 million PubChem compounds are 6,305 (0.035%) for NETSRIs, 4,993 (0.028%) for H3SRIs, 9,603 (0.054%) for 5HT1aSRIs, 6,326 (0.011%) for 5HT1bSRIs, 7574 (0.042%) for 5HT2cSRIs, 1252 (0.007%) for MC4SRIs and 1136 (0.006%) for NK1SRIs respectively. Substantial percentages of the MDDR virtual-hits belong to the classes of antidepressant, anxiolytic, antimigraine, and antipsychotic (**Table 5-7**, details in the next section), some of which may be true multi-target serotonin inhibitors. Therefore, the true false-hits rates of the COMBI-SVM are likely smaller than the computed rates, i.e., the false-hit rates of COMBI-SVM are ≤0.2%-4.0%, ≤0.042%-0.28% and ≤0.011%-0.054% in screening MDDR similarity compounds, all MDDR compounds, and PubChem compounds respectively. These rates are similar to the false-hit rates of ≤1.4%-9.4%, ≤0.057%-0.104%, and ≤0.013%-0.036% in COMBI-SVM screening of multi-target kinase inhibitors from MDDR and PUBCHEM compounds (105). These rates are also comparable and sometime better than the false-hit rates of 0.02%-0.37% and 0.05%-0.35% produced by other machine learning methods and molecular docking tools (105).

Table 5-6 The virtual screening performance of combinatorial SVMs for identifying multi-target serotonin inhibitors of the seven target pairs SERT-NET, SERT-H3, SERT-5HT1A, SERT-5HT1B, SERT-5HT2C, SERT-MC4 and SERT-NK1; The target-pairs in this table are arranged with decreasing similarity level between their drug-binding domains. There are only 7 MDDR compounds similar to a dual-inhibitor of SERT-MC_4, the corresponding virtual hit rate was thus un-computed because the small number of compounds may not provide statistically meaningful test of the SVM performance.

Target Pair	**Virtual Screening Performance**							
Target A – Target B (sequence identity between drug-binding domain)	Multi-target inhibitors		Inhibitors of individual target of the target pair inactive against another target of the target pair		Inhibitors of other three targets outside the target pair	MDDR compounds similar to multi-target inhibitors of the target pair	All 168,000 MDDR compounds	17 million PubChem compounds
	Yield	No (%) of identified true hits outside the common training active families of both targets	False hit rate for inhibitors of target A	False hit rate for inhibitors of target B	False hit rate	Virtual hit rate (No of virtual hits)	Virtual hit rate (No of virtual hits)	Virtual hit rate (No of virtual hits)
SERT-NET (72.3%)	49.50%	8 (7.9%)	22.40%	29.80%	2.40%	0.99% (81)	0.12% (201)	0.035% (6305)
SERT-$5HT_{1B}$ (15.1%)	22.8%	2 (3.5%)	13.80%	12.30%	0.95%	2.5% (185)	0.14% (241)	0.011% (6326)
SERT-MC_4 (11.7%)	83.33%	0	2.20%	8.02%	0.58%	-	0.042%(70)	0.007%(1252)
SERT-NK_1 (9.6%)	31.11%	13(28.9%)	4.20%	6.30%	1.16%	0.36%(1)	0.055%(92)	0.006%(1136)
SERT-$5HT_{1A}$ (8%)	47.70%	12 (5.6%)	15.40%	19.40%	7.10%	3.5% (256)	0.28% (464)	0.054% (9603)
SERT-$5HT_{2C}$ (3.2%)	22.0%	5(18.5%)	14.24%	12.40%	4.0%	4.0%(52)	0.21%(353)	0.042%(7574)
SERT-H_3 (1.7%)	25.90%	7 (4.8%)	5.40%	8.20%	3.50%	0.2% (3)	0.067% (112)	0.028% (4993)

Table 5-7 MDDR classes in which higher percentage (≥5%) of COMBI-SVM identified MDDR multi-target virtual hits are distributed in

Target pair(No.of COMBI-SVM virtual hits)	MDDR class that contains higher percentage of these virtual hits	Number (percentage) of COMBI-SVM identified multi-target virtual-hits in class	Percentage of MDDR class members as virtual hits
SERT-NET (201)	Antidepressant	56 (27.9%)	0.91%
	5-HT Reuptake Inhibitor	36 (17.9%)	3.68%
	Dopamine Reuptake Inhibitor	28 (13.9%)	14.29%
	Antipsychotic	25 (12.4%)	0.48%
	Norepinephrine Uptake Inhibitor	20 (10.0%)	6.33%
	Treatment of Cocaine Dependency	20 (10.0%)	25.97%
	Anxiolytic	15 (7.5%)	0.22%
	Calcium Channel Blocker	15 (7.5%)	0.88%
	Antimigraine	14 (7.0%)	0.81%
	Analgesic, Non-Opioid	13 (6.5%)	0.27%
SERT-H_3 (112)	Antidepressant	30 (26.8%)	0.49%
	Antipsychotic	23 (20.5%)	0.44%
	Analgesic, Non-Opioid	13 (11.6%)	0.27%
	5-HT Reuptake Inhibitor	10 (8.9%)	1.02%
	Anxiolytic	10 (8.9%)	0.15%
	Cognition Disorders, Agent for	10 (8.9%)	0.13%
	Antiparkinsonian	9 (8.0%)	0.48%
	Anticonvulsant	8 (7.1%)	0.26%
	Antifungal	7 (6.3%)	0.24%

	Calcium Channel Blocker	7 (6.3%)	0.41%
SERT-5HT$_{1A}$ (464)	Antidepressant	177 (38.1%)	11.91%
	Antimigraine	118 (25.4%)	2.27%
	Antipsychotic	113 (24.3%)	1.67%
	5-HT$_{1D}$ receptor Agonist	100 (21.6%)	10.21%
	Anxiolytic	62 (13.4%)	3.58%
	5-HT Reuptake Inhibitor	49 (10.6%)	8.52%
	5-HT$_{1A}$ receptor Agonist	48 (10.3%)	4.66%
	5 HT2A Antagonist	47 (10.1%)	7.40%
	Dopamine (D$_4$) Antagonist	26 (5.6%)	8.05%
	Analgesic, Non-Opioid	25 (5.4%)	3.63%
SERT-5HT$_{1B}$ (241)	Antidepressant	82 (34.0%)	1.33%
	Antimigraine	76 (31.5%)	4.39%
	5-HT$_{1D}$ receptor Agonist	63 (26.1%)	9.62%
	5-HT reuptake Inhibitor	53(22.0%)	5.41%
	Antipsychotic	47(19.5%)	0.9%
	Anxiolytic	44(18.32%)	0.65%
	5-HT2A receptor Antagonist	25 (10.4%)	3.63%
	5-HT1Areceptor Agonist	21 (8.7%)	2.0%
	Dopamine (D$_4$) Antagonist	15 (6.2%)	2.23%
	5-HT$_{1A}$ receptor Antagonist	13(5.4%)	2.26%

Target pair(No.of COMBI-SVM virtual hits)	**MDDR class that contains higher percentage of these virtual hits**	**Number (percentage) of COMBI-SVM identified multi-target virtual-hits in class**	**Percentage of MDDR class members as virtual hits**
SERT-5HT$_{2C}$ (353)	Antipsychotic	126 (5.7%)	2.42%
	Antidepressant	99 (28.0%)	1.60%
	Anxiolytic	76 (21.5%)	1.12%
	5-HT$_{2A}$ receptor Antagonist	46 (13.0%)	6.68%
	Antimigraine	36 (10.2%)	2.08%
	5-HT$_{1A}$ receptor Agonist	34 (9.6%)	3.30%
	Dopamine (D4) Antagonist	32(9.1%)	4.75%
	5-HT Reuptake Inhibitor	24(6.8%)	2.45%
	Antiparkinsonian	22 (6.2%)	1.16%
	5-HT$_{1D}$ agent Agonist	22 (6.2%)	1.16%
	Antihypertensive	21(5.9%)	0.19%
	Antiallergic/Antiasthmatic	18(5.1%)	0.17%
	Cognition Disorders, Agent for	18(5.1%)	0.24%
SERT-MC$_4$(70)	Antidepressant	15(21.4%)	0.24%
	Anti-allergic/Anti-asthmatic	15(21.4%)	0.14%
	Anxiolytic	13(18.6%)	0.19%
	Neurokinin NK$_2$ Antagonist	9(12.9%)	2.16%
	Neurokinin NK$_3$ Antagonist	8(11.4%)	4.40%
	Antipsychotic	7(10.0%)	0.13%
	Substance P Antagonist	7(10.0%)	0.40%
	Antiviral (AIDS)	5(7.1%)	0.11%

	Analgesic, Non-Opioid	4(5.7%)	0.08%
	Cognition Disorders, Agent for	4(5.7%)	0.05%
	5-HT Reuptake Inhibitor	4(5.7%)	0.41%
	Anti-arthritic	4(5.7%)	0.03%
SERT-NK$_1$(92)	Substance P Antagonist	23(25.0%)	1.31%
	Antidepressant	18(19.6%)	0.29%
	Anxiolytic	15(16.3%)	0.22%
	Antipsychotic	11(12.0%)	0.21%
	Antiallergic/Antiasthmatic	11(12.0%)	0.10%
	5-HT Reuptake Inhibitor	11(12.0%)	1.12%
	Analgesic, Non-Opioid	10(10.9%)	0.20%
	Neurokinin NK$_2$ Antagonist	9(9.8%)	2.16%
	Calcium Channel Blocker	9(9.8%)	0.53%
	5-HT$_{1A}$ receptor Agonist	6(6.5%)	0.58%
	Antihypertensive	6(6.5%)	0.05%
	Antianginal	6(6.5%)	0.18%
	Adrenergic (beta) Blocker	6(6.5%)	2.76%
	Antiarrhythmic	5(5.4%)	0.19%
	Anti-inflammatory	5(5.4%)	0.09%
	Neurokinin Antagonist	5(5.4%)	3.73%
	Cognition Disorders, Agent for	5(5.4%)	0.07%
	5-HT$_{1A}$ receptor Antagonist	5(5.4%)	0.87%
	Antiviral (AIDS)	5(5.4%)	0.11%

5.3.4 Virtual screening performance of Combinatorial SVM in searching multi-target serotonin inhibitors from large compound libraries

COMBI-SVM identified MDDR virtual hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 5-7** gives the MDDR classes in which higher percentage (⩾5%) of COMBI-SVM identified MDDR dual inhibitor virtual hits are distributed. We found that 15-177 (21.4%-38.1%), 10-76 (7.5%-21.5%), and 4-53 (5.7%-22.0%) of the 70-464 dual-inhibitor virtual hits of the seven target-pairs belong to the antidepressant, anxiolytic and 5HT reuptake inhibitor class respectively. It is noted that serotonin reuptake inhibitors have been used as antidepressant and anxiolytic agents (95). Therefore, some of the COMBI-SVM virtual hits are either known SSRIs or have the same therapeutic actions of SSRIs, which were misidentified as dual inhibitors by COMBI-SVM partly because it has 2.2%-22.4% false-hit rates in misclassifying SSRIs as dual inhibitors of the seven target pairs (**Table 5-6**). Moreover, 20 (10.0%) of the 201 SERT-NET dual inhibitor virtual hits belong to the norepinephrine uptake inhibitor class. While some of these virtual hits might be true SERT-NET dual inhibitors, most of these individual target NET inhibitors were falsely selected as dual inhibitors by COMBI-SVM at 6.33% false-hit rate (**Table 5-7**).

We found that 118 (25.5%), 76 (31.5%), 36 (10.2%) and 14 (7.0%) MDDR virtual hits for SERT-5HT$_{1A}$, SERT-5HT$_{1B}$, SERT-5HT$_{2C}$ and SERT-NET belong to the antimigraine class respectively. Serotonin has been implicated in migraine pathophysiology with a low 5-HT state facilitating activation of the trigeminovascular nociceptive pathway (272). Because serotonin is primarily reduced by SERT (240), serotonin reuptake inhibitors may in some circumstances have antimigraine effect in certain patients (273). Some of the MDDR antimigraine virtual hits may be selected by COMBI-SVM partly because they are SERT inhibitors (COMBI-SVMs select individual-target inhibitors as dual-target serotonin reuptake inhibitors at 2.2%-29.8% false-hit rates based on the statistics in **Table 5-6**). Moreover, 25-113 (11.4%-24.3%) MDDR virtual hits of six target pairs (SERT-NET, SERT-H$_{3,}$ SERT-5HT$_{1B}$, SERT-5HT$_{2C}$, SERT-MC$_4$ and SERT-NK$_1$) belong to the antipsychotic class. Some antipsychotic drugs show certain level of activity against serotonin reuptakes and 5-HT receptors (274). It is further noted that serotonin reuptake inhibitors augment and synergize with antipsychotic drugs hence serotonin reuptake inhibitors have been used in combination with antipsychotic drugs in the treatment of some psychiatric disorders (275). Hence, some of the antipsychotic MDDR virtual hits may be selected because they have these activities.

An additional set of 87-100 (21.6%-21.7%), 38-48 (9.5%-10.3%) and 36-47 (9.0%-10.1%) dual inhibitor virtual hits of the SERT-5HT$_{1A}$ and SERT-5HT$_{1B}$

target pairs belong to the 5-HT$_{1D}$ receptor agonist, 5-HT$_{1A}$ receptor agonist, and 5-HT$_{2A}$ receptor antagonist classes respectively. As discussed below, some of these MDDR 5-HT$_{1D}$ receptor agonist, 5-HT$_{1A}$ receptor agonist, and 5-HT$_{2A}$ receptor antagonist virtual hits were falsely selected by COMBI-SVM possibly because they have some level of structural similarity to 5-HT$_{1A}$ receptor antagonists or 5-HT$_{1B}$ receptor antagonists. Analogues of certain scaffolds have been found to bind to both 5-HT$_{1A}$ and 5-HT$_{1D}$ receptors with weak partial agonist activity in cloned receptor and antagonistic activity in *in-vitro* studies (276). Some compounds such as BMY 7378 can act as both 5-HT$_{1A}$ agonist and antagonist depending on the location of 5-HT$_{1A}$. BMY 7378 shows agonist activity at 5-HT$_{1A}$ autoreceptors and act as antagonists or show partial agonist activity at postsynaptic 5-HT$_{1A}$ receptors (277). Both mixed 5-HT$_{1A}$ and 5-HT$_{2A}$ receptor antagonists and 5-HT$_{1A}$ receptor agonists have been derived from the same scaffolds (278). The human 5-HT$_{1B}$ and 5-HT$_{1D}$ receptors are significantly similar in sequence despite being encoded by two distinct genes, and some dual 5HT$_{1B/1D}$ receptor antagonists show substantial degree of structural similarity to dual 5HT$_{1B/1D}$ agonists (279). Some analogs of specific scaffolds are mixed 5-HT$_{1B}$ and 5-HT$_{2A}$ receptor antagonists (280). Moreover, some compounds have been reported to have dual 5-HT$_{1A}$ receptor agonist and serotonin reuptake inhibitory activities (281). It is possible that some of the MDDR 5-HT$_{1A}$ receptor agonist virtual hits were selected by the COMBI-SVM of SERT-5HT$_{1B}$ target pair because they have serotonin reuptake inhibitory activity which may be falsely recognized as multi-target 5HT1bSRIs by COMBI-SVM at 13.8% false-hit rate based on the statistics in **Table 5-6**.

5.3.5 Analysis of MDDR virtual hits of combinatorial SVM

At present, the 3D structure is unavailable for the eight targets considered in this work (serotonin transporter, noradrenaline transporter, H_3 receptor, 5-HT_{1A} receptor, 5-HT_{1B} receptor, 5-HT_{2C} receptor, NK_1 receptor and MC_4 receptor). Only some of their homologous proteins or other members from the same GPCR families, such as H1 receptor, have 3D structure information available (282, 283). While these structures give important insights into functional mechanism and allow the modelling of ligand binding to the eight evaluated targets, the modelled and homologous structures may not provide the most appropriate structural platforms as those of high-resolution crystal structures for fair comparison of the VS performance of COMBI-SVM with molecular docking methods. We therefore only compared the VS performance of COMBI-SVMs with three VS methods, i.e., similarity searching (284), k-NN (285), and PNN (286), by using the common testing datasets composed of 6~216 dual inhibitors of the seven evaluated target pairs, 917-1951 individual target inhibitors of the same target pairs, 8110-8688 inhibitors of the other six target pairs outside a given target pair, and 168,000 MDDR compounds respectively. Similarity searching was conducted against known multi-target inhibitors of each target pair. The training datasets of k-NN and PNN and the methods for estimating the yield and virtual hit rate are the same as those of SVM.

Table 5-8 shows the comparison of the performance of COMBI-SVM with the other three VS methods for identifying multi-target inhibitors of the seven target-pairs from the four common testing datasets. Overall, the dual-inhibitor yields of all VS methods are comparable, mostly in the ranges of 20%~83% for the seven target-pairs with the exception of k-NN for SERT-NK$_1$ (7.7%) and similarity searching for SERT-5HT$_{2c}$ (11.1%). Compared to COMBI-SVM, k-NN produced comparable false-hit rates, and similarity searching and PNN produced slightly higher false-hit rates in misidentifying individual-target inhibitors of the same target-pair and inhibitors of the other six target pairs outside a target pair as dual-inhibitors

The false-hit rates of the similarity searching method may be significantly reduced by adjusting the similarity cut-off values for individual targets, which may however lead to significantly reduced yields. The higher false-hit rates likely arise in part from the difficulty in establishing optimal molecular similarity threshold values that correlate with biological activity, and in separating active and inactive close analogs of reference molecules (287). Data fusion and group fusion approaches may be explored to conduct multiple similarity searches using different sets of molecular representations, similarity measure and parameters followed by the combination of the resulting search outputs to give a single fused output (288), (289). The higher false-hit rates may also arise from the bias linked to molecular complexity and size, i.e., reference molecules of increasing size generate systematically higher Tanimoto coefficient values in database searching (290). This bias may be partly reduced by exploring bit density reduction methods (290), complexity-independent molecular representations (290) and complexity-independent similarity metrics (290).

In screening the MDDR compounds, COMBI-SVM produced slightly to substantially lower virtual hit rates (0.042%~0.28%) than those of similarity searching (2.81%-8.2%), k-NN (0.15%-0.83%) and PNN (0.93%-3.4%) in identifying the MDDR compounds as dual inhibitor virtual hits of the evaluated target pairs. The numbers of MDDR compounds in the antidepressant and 5-HT reuptake inhibitor classes are 6182 and 979 respectively. It is expected that no more than half of the MDDR antidepressant compounds are SSRIs. Therefore, the total number of labelled and unlabelled SSRIs in MDDR can be crudely estimated as ~1000-3000, most likely significantly less than 3000. Assuming that the ratio of the dual-target serotonin reuptake inhibitors against SSRIs in MDDR is roughly similar to those of known dual-target serotonin reuptake inhibitors against SSRIs which are 9.0% (101 vs. 1125) for NETSRIs, 8.2% (147 vs. 1804) for H3SRIs, 12.9% (216 vs. 1679) for 5HT1aSRIs, 3.0% (57 vs. 1894) for 5HT1bSRIs, 1.4% (27 vs. 1924) for 5HT2cSRIs, 0.3% (6 vs. 1951) for MC4SRIs and 2.4% (45 vs. 1910) for NK1SRIs. Then the numbers of dual-target serotonin reuptake inhibitors in MDDR can be crudely estimated as ~3-380 (1000×0.3% - 3000×12.9%), most likely significantly less than 380. Therefore the numbers of COMBI-SVM identified MDDR dual inhibitor virtual hits of the evaluated target pairs (70-464) are consistent to the crudely estimated numbers of dual inhibitors in MDDR than the identified numbers from the other three methods (971-12,698).

Table 5-8 Comparison of the performance of combinatorial SVMs with other virtual screening methods for identifying multi-target inhibitors of the four target pairs

Virtual Screening Performance Measure	**Method**	**Virtual Screening Performance for Target pair**						
		SERT-NET	SERT-H_3	SERT-$5HT_{1A}$	SERT-$5HT_{1B}$	SERT-$5HT_{2C}$	SERT-MC_4	SERT-NK_1
Yield of Multi-Target Inhibitors of Target pair	SVM	49.50%	25.90%	47.70%	22.8%	22%	83.33%	31.11%
	Similarity Searching	53.50%	20.40%	63.90%	40.40%	11.11%	33.3%	48.89%
	k-NN	59.40%	10.90%	34.30%	31.60%	18.52%	16.7%	6.67%
	PNN	57.40%	38.10%	45.40%	45.60%	33.33%	66.7%	20.00%
False-Hit Rate for "Individual-Target" Inhibitors of the Same Target pair	SVM	22.4%-29.8%	5.4%-8.2%	15.4%-19.4%	13.8%-12.3%	14.24%-12.4%	2.2%-8.02%	4.2%-6.3%
	Similarity Searching	46.5%-42.4%	35.6%-21.8%	28.6%-46.9%	28.6%-65.4%	19.4%-13.6%	8.5%-26.3%	17.7%-16.0%
	k-NN	19.8%-25.1%	9%-8.5%	16.6%-24.3%	14.1%-32.6%	15.0%-16.3%	3.1%-11.5%	4.5%-4.9%
	PNN	38.4%-52.3%	22.2%-25.5%	34.3%-38.9%	30.3%-4.7%	34.8%-31.8%	6.6%-27.9%	11.8%-9.8%
False-Hit Rate for Inhibitors of the Other Six Targets Outside the Target pair	SVM	2.40%	3.50%	7.10%	0.95%	4.0%	0.58%	1.16%
	Similarity Searching	15.90%	20.50%	24.10%	11.60%	10.6%	5.0%	8.4%
	k-NN	3.00%	3.50%	9.30%	5.20%	4.2%	1.2%	2.9%
	PNN	16.90%	13.50%	24.20%	10.80%	14.2%	0.37%	8.8%
Virtual Hit Rate for 168,000 MDDR Compounds	SVM	0.12%	0.067%	0.28%	0.14%	0.21%	0.042%	0.055%
	Similarity Searching	6.80%	8.20%	7.60%	7.60%	3.81%	3.26%	3.54%
	k-NN	0.58%	0.41%	0.83%	0.75%	0.52%	0.15%	0.81%
	PNN	3.14%	2.35%	3.40%	2.83%	3.90%	0.93%	2.24%

5.4 Conclusion

In-silico methods have been increasingly explored for facilitating multi-target drug discovery, and shown promising potential in identifying selective multi-target agents. In Chapter 4, I explored and discussed their application in discovering multi-target kinase inhibitors as the preliminary tests of the performances of combinatorial machine learning methods and promising results were obtained. Therefore, this chapter was the application of those methods and it further suggested that combinatorial SVM VS tools developed from individual target inhibitors are capable of identifying dual target serotonin reuptake inhibitors at comparably good yields and low false-hit rates, and in some cases substantially lower false-hit rates than some of the other VS tools in screening large chemical libraries. COMBI-SVMs, in combination with other methods, may be useful for facilitating the search of novel multi-target antidepressants by screening larger chemical libraries. With more knowledge of newly discovered selective multi-target agents from the current and future drug discovery efforts (291, 292), COMBI-SVMs and other in-silico methods will have opportunities for enhanced performances. Further improvement of the algorithms and parameters of VS methods (103, 293-298) also enhance their capability and application range in facilitating multi-target drug discovery. Moreover, the introduction of more comprehensive elements of distinguished structural and physicochemical features of selective multi-target agents and multi-target activity and binding site profiles enable the development of more effective and relevant tools for the identification of selective multi-target agents as well as active compounds against an individual target.

Chapter 6 Concluding Remarks

Multi-target agents have become a new and promising trend in the treatment for many diseases due to their improvement in efficacy and the averting of side effects. The focus of the thesis work was to assist the discovery of multi-target agents. This study was divided into two big compartments. The first part consisted of the construction and updates of the two chemoinformatics databases Kinetics Database of Biomolecular Interactions (KDBI) and Therapeutic Target Database (TTD) (Chapter 3). The second one discussed the application of virtual screening methods in discovery of two different systems, kinase inhibitors which perform as a major drug class and antidepressants, which are very important drug class especially in the modern societies where major depression has been empowered by the stresses (Chapter 4 and Chapter 5). This last chapter summarizes the major contributions and findings of this study (section 6.1). Section 6.2 discusses the limitations and presents suggestions for future studies.

6.1 Major Findings and Merits

6.1.1 Merits of the updates of KDBI and TTD in facilitating multi-target drug discovery

The kinetic information of biomolecular interactions plays the key factor in the quantitative investigation of the components of cellular networks and their interactions. They can promote the studies of cellular functions and interactions on a system level. Pathway studies are found especially interesting in understanding the mechanisms behind complex diseases which usually involve interactions within or between diseases related pathways. On the other hand, a better understating of those mechanisms offers the theoretic foundation and guidance for the discovery of chemical agents that can improve the drug efficacy by acting on multiple targets. Therefore, the update of (KDBI) of pathway interaction kinetic information can greatly enhance the usefulness of KDBI. It is also found that together with this improvement, other factors such as the manual annotation, presentation and speed of database opening, cross-referencing to other databases, and the inclusion of critical information that could significantly increase the speed of their research of other researchers can greatly improve the quality of databases. Another merit of the updates of KDBI is data integration of the simulating models by Systems Biology Markup Language (SBML). Systems biology is characterized by synergistic integration of theory, computational modeling, and experiment (299). Nowadays, there is a proliferation of research institutions that produce sources of huge amounts of biological data derived from experimentation with biological systems and construct numerous stimulating models based on those data. Therefore, it is in great demand for a common format for describing models in the exchange models between different simulation and analysis tools. SMBL hence is developed as an exchange format used by different present-day software tools to communicate the essential aspects of a computational model (300). The integration of SBML into the simulation models data of KDBI hence can bring great convenience for users with different software in system biology simulation and studies.

Therapeutic target database (TTD) was a pioneer for providing pharmaceutical information on therapeutic target and it has become a very functional tool to facilitate drug target studies. After its update in 2010 (4), major improvements and updates have been made to TTD. The information coverage has significantly increased from 1,894 targets and 5,028 drugs to 2,025 targets and 17,816 drugs plus 3,681multi-target agents. Table 6-1 summarizes the statistics of the current TTD version. The new features added are the highlights of the update of TTD this time. These new features include (1) target validation information, quantitative structure activity relationship (QSAR) models for active compounds; (2) multi-target agents data with structure and potency information; (3) drug combinations; (4) nature-derived drugs together with drug species origin data. Therefore, informative and comprehensive information has been integrated to this version of TTD. It has become a very reliable, informative, useful, multifunctional and convenient source of drug target information.

Table 6-1 The data statistics of the updated Target Therapeutic Database

Target coverage (2025)	Successful		364
	Clinical trial		286
	Discontinued		44
	Research		1,331
Drug coverage (17,816)	Approved		1,540
	Clinical trial		1,423
	Experimental		14,853
	Multi-target agents	Small molecules	14,170
		Antisense drugs	652
Other coverage	Protein biochemical class		61
	Drug therapeutic class		140
New features	Quantitative structure-activity relationship (QSAR) models		
	Target validation	Drug potency against target	
		Drug action against disease model	
		Effect of target knockout, knockdown or genetic variations	
	Drug combinations	Pharmacodynamically synergistic	72
		Pharmacodynamically additive	14
		Pharmacodynamically antagonist	4
		Pharmacokinetically potentiative	19
		Pharmacokinetically	7
	Nature-derived drugs	Approved	939
		Clinical trial	369
		Preclinical	119

6.1.2 Findings of combinatorial machine learning methods for virtual screening in the multi-target kinase inhibitors and antidepressant agents

Machine learning (ML) methods have been broadly applied as virtual screening tools due to their capability of high-CPU speed and the ability to cover highly diverse spectrum of compounds. However, while presenting equally good hit selection performance in screening extremely-large and large libraries, the currently developed machine learning tools have the tendency for lower hit-rate and, in some cases, lower enrichment factor than the best performing structure based virtual screening tools.

To improve the performance of one of the most popular ML method support vector machine (SVM), diverse inactive compounds apart from the known inactive compounds and active compounds of other biological target classes were used as negative data in training sets in this work. It was achieved by generating putative inactive compounds by the in-house programs. An advantage of this approach is that it is independent on the knowledge of known inactive compounds and active compounds of other biological target classes. This enables more extended coverage of the "inactive" chemical space compared to when only the limited knowledge of inactive compounds and compounds of other biological classes are used. In the virtual screening for active compounds in large libraries such as PubChem and MDDR, the hit-rates of the methods used in this work are comparable and the enrichment factors are substantially better than the best results of other VS tools. And the usage of putative negatives contributes to it. This method greatly increased the performance of VS without losing much positive accuracy. This showed that a fulfilled presentation in the chemical space can provide improvement of machine learning methods in virtual screening, although some noises could be introduced with the generation of putative inactive compounds.

In this work, combinatorial support vector machines (COMBI-SVMs) were tested as VS tools for searching dual-inhibitors of 7 combinations of 6 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck) and 7 combinations of 8 antidepressant target (serotonin transporter, noradrenaline transporter, H_3 receptor, 5-HT_{1A} receptor, 5-HT_{1B} receptor, 5-HT_{2C} receptor, melancortin 4 receptor and neurokinin 1 receptor). COMBI-SVMs Models were fairly selective in misidentifying as dual-inhibitors of the non-dual inhibitors of the same kinase-pairs and produced low false-hit rates in misidentifying as dual-inhibitors of PubChem and MDDR databases. The performance of COMBI-SVM was compared with DOCK, k-NN and PNN methods. COMBI-SVM VS tools showed good capability in identifying dual-inhibitors of several anticancer target kinase-pairs at comparable and in many cases substantially lower false-hit rates. In the studies of multi-target antidepressants, COMBI-SVMs showed moderate to relatively good target selectivity in misclassifying as dual of the individual target inhibitors of the same target pair and of the other 6 targets outside the target pair. COMBI-SVMs showed low dual inhibitor false hit rates in screening 17 million PubChem compounds, 168,000 MDDR compounds, and 7-8,181 MDDR compounds similar to the dual inhibitors. Compared with similarity searching, k-NN and PNN methods, COMBI-SVM produced comparable dual inhibitor yields, similar target selectivity, and lower false hit rate in screening 168,000 MDDR compounds.

Comparing the performances of COMBI-SVMs in virtual screening multi-target agents in anticancer kinase target combinations and in multi-target antidepressant agents, I found that the sequence similarity could affect the selectivity of COMBI-SVMs against the same target pair. **Table 6-2** shows the target selectivity as the false hit rate of misidentifying the other target pair as dual inhibitor in a target pair and dual inhibitor yield of each pair in the 4 kinase pairs and 7 antidepressant pairs. It is found that generally, there seems to be a tendency that the higher the sequence identity of the target pair is, the lower the target selectivity tends to be but the higher the dual yield tends to be. The dual inhibitor yield of target pair SERT-MC$_4$ is excluded because very few dual inhibitors (only 6) could be found for testing and the result might not be so valid statistically.

Table 6-2 Target pair (sequence identity) and the false hit rate for inhibitor pairs and their dual inhibitor yields

Target pair (sequence identity)	False hit rate for inhibitors of target A	False hit rate for inhibitors of target B	Average false hit rate for the target pair	Dual inhibitor yields
SERT-NET (72.3%)	22.40%	29.80%	26.1%	49.50%
Src-Lck (67.6%)	15.80%	18.70%	17.3%	48.20%
EGFR-Src(37.4%)	12.90%	11.10%	12.0%	26.80%
EGFR-FGFR (33.2%)	10.10%	8.70%	9.4%	40.90%
VEGFR-Lck (32.7%)	6.60%	29.20%	17.9%	52.60%
SERT-5HT$_{1B}$ (15.1%)	13.80%	12.30%	13.1%	22.80%
SERT-MC$_4$ (11.7%)	2.20%	8.02%	5.1%	-
SERT-NK$_1$ (9.6%)	4.20%	6.30%	5.3%	31.11%
SERT-5HT$_{1A}$ (8%)	15.40%	19.40%	17.4%	47.70%
SERT-5HT$_{2C}$ (3.2%)	14.24%	12.40%	13.3%	22.00%
SERT-H$_3$ (1.7%)	5.40%	8.20%	6.8%	25.90%

6.2 Limitations and Suggestions for the Future Studies

The work for updates of KDBI and TTD has few limitations due to the data availability and the methods used in the development of databases. The information abundance of a database is limited to the current availability of accessible chemoinformatics data. On the other hand, new findings and results in chemoinformatics fields have been proliferating. Therefore, it is suggested to stay in tone with the latest findings in systems biology, therapeutic target research and drug discovery and constantly update the data collection in the two databases. Hence continuous efforts are required to maintain the quality and quantity of useful and comprehensive databases such as KDBI and TTD. Moreover for KDBI, its server is currently running on IIS 5.0 web server which has limitation in the processing of requests (maximum 10 requests at a time). Therefore, future improvement in the requests can be done by upgrading the system. In KDBI, the SBML file for pathway simulation model is created based on Java API of SBML version 2.4. The system biology related software which process SBML file may not support lower version of SBML after their upgrading. Then the SBML file downloaded from KDBI will not open in that particular software. In these situations, users are advised to edit these SBML file using some SBML editor.

In my study, I applied the generation of putative negatives for the machine learning methods application. This approach requires a classification of the chemical space which has always been a difficult task in chemoinformatics. The classification of the chemical space needs a clustering method, a distance matrix selection and descriptors. K-means clustering method was used in this work. It is not the best clustering method but is suitable and computable for large chemical spaces. In future studies, more advanced clustering algorithm can be developed for improving the accuracy of chemical space clustering. Additionally, the selection of correlation coefficients and other chemical descriptors such as fingerprint can also help the improvement. Another possible drawback associated with the putative negatives generation approach is the possible inclusion of some undiscovered active compounds in the “inactive” class. This may hinder the identification of novel active compounds by machine learning methods. However, such an adverse effect is expected to be relatively small for many biological target classes.

As for the virtual screening (VS) for multi-target agents, the support vector machine (SVM) is a robust but not a perfect machine learning method. The SVM models developed using the putative negative dataset have been proven to be able to improve the false hit rates. However, there are still some false hits that cannot be excluded easily. These false hits are selected as positive agents by the SVM models mostly due to the structural framework similarities with the actual active compounds. This could be caused by the molecular descriptors used in the SVM models in that they are insufficient to adequately differentiate the compounds with similar structural frameworks. In order to solve this problem, it is necessary to test different combinations of descriptors and apply optimal sets of descriptors by using more refined feature selection algorithms and parameters in future work.

Besides, the integration of new descriptors may help appropriate representations of compounds. Therefore, it is encouraging to employ new descriptors in the model constructions.

The increase of positive compounds number in the model construction means a better representation of the positive agents in the chemical space. Hence, the capability of the combinatorial SVMs in identifying multi-kinase inhibitors and the antidepressant multi-target agents can be further enhanced by more data availability in the VS tool development processes. With the development of selective multi-target agents discovery from the current and future drug discovery efforts, it is possible to introduce more comprehensive elements of distinguished structural and physicochemical features of selective multi-target agents into the training of combinatorial VS tools for more effective identification of selective multi-target agents.

There is no conclusive answer to which VS approach is the best. Both ligand based and structural based methods have their own advantages and drawbacks. Therefore, the choice of one or another depends on the specific case faced by the medicinal chemist. In terms of performance, ligand based methods have the advantage of better enrichment factors and higher speed serving and they are more efficient in removing non active compounds; structure based methods provide a more direct view of the interactions between the ligand and molecular target and it has an advantage for the detecting of novel structures. Nowadays a synergistic, rational, synthetic combination of different approaches has become a trend. The combined VS approaches aims to firstly include less costly approaches, usually ligand based VS, at the first stage and apply the most demanding methods, such as docking, for the last stage when the original large compound library has been reduced to a manageable size after the previous stage.

BIBLIOGRAPHY

1. M.J. Millan. Multi-target strategies for the improved treatment of depressive states: Conceptual foundations and neuronal substrates, drug discovery and therapeutic application. Pharmacol Ther. 110:135-370 (2006).
2. T.D.J.P. B. Waszkowycz, R. A. Sykes, J. Li. Large-scale virtual screening for discovering leads in the postgenomic era. IBM SYSTEMS JOURNAL. 40:360-376 (2001).
3. P. Kumar, B.C. Han, Z. Shi, J. Jia, Y.P. Wang, Y.T. Zhang, L. Liang, Q.F. Liu, Z.L. Ji, and Y.Z. Chen. Update of KDBI: Kinetic Data of Bio-molecular Interaction database. Nucleic Acids Res. 37:D636-641 (2009).
4. F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, and Y. Chen. Update of TTD: Therapeutic Target Database. Nucleic Acids Res. 38:D787-791 (2010).
5. J. Drews. Drug discovery: a historical perspective. Science. 287:1960-1964 (2000).
6. J. Bajorath. Integration of virtual and high-throughput screening. Nat Rev Drug Discov. 1:882-894 (2002).
7. R.S. Bohacek, C. McMartin, and W.C. Guida. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev. 16:3-50 (1996).
8. T.F.a.J.-L. Reymond. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. J Chem Inf Model. (published on Web 01/30/2007): (2007).
9. M. Rareyand M. Stahl. Similarity searching in large combinatorial chemistry spaces. J Comput Aided Mol Des. 15:497-520 (2001).
10. R.V. Guido, G. Oliva, and A.D. Andricopulo. Virtual screening and its integration with modern drug design technologies. Curr Med Chem. 15:37-46 (2008).
11. N. Brooijmansand I.D. Kuntz. Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct. 32:335-373 (2003).
12. I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins. 47:409-443 (2002).
13. R. Wang, Y. Lu, and S. Wang. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem. 46:2287-2303 (2003).
14. N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C.R. Corbeil. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol. 153 Suppl 1:S7-26 (2008).

15. G.L. Warren, C.W. Andrews, A.M. Capelli, B. Clarke, J. LaLonde, M.H. Lambert, M. Lindvall, N. Nevins, S.F. Semus, S. Senger, G. Tedesco, I.D. Wall, J.M. Woolven, C.E. Peishoff, and M.S. Head. A critical assessment of docking programs and scoring functions. J Med Chem. 49:5912-5931 (2006).

16. T. Schulz-Gaschand M. Stahl. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model. 9:47-57 (2003).

17. R. Kimand J. Skolnick. Assessment of programs for ligand binding affinity prediction. J Comput Chem. 29:1316-1331 (2008).

18. J. Kirchmair, P. Markt, S. Distinto, G. Wolber, and T. Langer. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes? J Comput Aided Mol Des. 22:213-228 (2008).

19. R.P. Sheridan, G.B. McGaughey, and W.D. Cornell. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. J Comput Aided Mol Des. 22:257-265 (2008).

20. A.N. Jain. Bias, reporting, and sharing: computational evaluations of docking methods. J Comput Aided Mol Des. 22:201-212 (2008).

21. P.C. Hawkins, A.G. Skillman, and A. Nicholls. Comparison of shape-matching and docking as virtual screening tools. J Med Chem. 50:74-82 (2007).

22. G. Wolber, T. Seidel, F. Bendix, and T. Langer. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today. 13:23-29 (2008).

23. H. Sun. Pharmacophore-based virtual screening. Curr Med Chem. 15:1018-1024 (2008).

24. E. Carosati, R. Budriesi, P. Ioan, M.P. Ugenti, M. Frosini, F. Fusi, G. Corda, B. Cosimelli, D. Spinelli, A. Chiarini, and G. Cruciani. Discovery of novel and cardioselective diltiazem-like calcium channel blockers via virtual screening. J Med Chem. 51:5552-5565 (2008).

25. K. Moffat, V.J. Gillet, M. Whittle, G. Bravi, and A.R. Leach. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. J Chem Inf Model. 48:719-729 (2008).

26. G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.F. Truchon, and W.D. Cornell. Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model. 47:1504-1519 (2007).

27. F.L. Stahuraand J. Bajorath. New methodologies for ligand-based virtual screening. Curr Pharm Des. 11:1189-1202 (2005).

28. B.K. Shoichet. Virtual screening of chemical libraries. Nature. 432:862-865 (2004).

29. S. Yamane, S. Ishida, Y. Hanamoto, K. Kumagai, R. Masuda, K. Tanaka, N. Shiobara, N. Yamane, T. Mori, T. Juji, N. Fukui, T. Itoh, T. Ochi, and R. Suzuki. Proinflammatory role of amphiregulin, an epidermal growth

factor family member whose expression is augmented in rheumatoid arthritis patients. J Inflamm (Lond). 5:5 (2008).

30. J.M. Maris, J. Courtright, P.J. Houghton, C.L. Morton, E.A. Kolb, R. Lock, M. Tajbakhsh, C.P. Reynolds, S.T. Keir, J. Wu, and M.A. Smith. Initial testing (stage 1) of sunitinib by the pediatric preclinical testing program. Pediatr Blood Cancer. 51:42-48 (2008).

31. R. Gozalbes, L. Simon, N. Froloff, E. Sartori, C. Monteils, and R. Baudelle. Development and experimental validation of a docking strategy for the generation of kinase-targeted libraries. J Med Chem. 51:3124-3132 (2008).

32. X.Q. Deng, H.Y. Wang, Y.L. Zhao, M.L. Xiang, P.D. Jiang, Z.X. Cao, Y.Z. Zheng, S.D. Luo, L.T. Yu, Y.Q. Wei, and S.Y. Yang. Pharmacophore modelling and virtual screening for identification of new Aurora-A kinase inhibitors. Chem Biol Drug Des. 71:533-539 (2008).

33. F. Deanda, E.L. Stewart, M.J. Reno, and D.H. Drewry. Kinase-Targeted Library Design through the Application of the PharmPrint Methodology. J Chem Inf Model. 48:2395-2403 (2008).

34. P. Willett, J.M. Barnard, and G.M. Downs. Chemical Similarity Searching. J Chem Inf Comput Sci. 38:983-996 (1998).

35. R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers and Chemistry. 26:5-14 (2001).

36. D.T. Manallackand D.J. Livingstone. Neural networks in drug discovery: have they lived up to their promise? European Journal of Medicinal Chemistry. 34:195-208 (1999).

37. M.W.B. Trotterand S.B. Holden. Support vector machines for ADME property classification. QSAR & Combinatorial Science. 22:533-548 (2003).

38. T. Lengauer, C. Lemmen, M. Rarey, and M. Zimmermann. Novel technologies for virtual screening. Drug Discov Today. 9:27-34 (2004).

39. T.I. Opreaand H. Matter. Integrating virtual screening in lead discovery. Curr Opin Chem Biol. 8:349-358 (2004).

40. A. Bocker, G. Schneider, and A. Teckentrup. NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. J Chem Inf Model. 46:2220-2229 (2006).

41. D. Schuster, E.M. Maurer, C. Laggner, L.G. Nashev, T. Wilckens, T. Langer, and A. Odermatt. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. J Med Chem. 49:3454-3466 (2006).

42. T. Steindl, C. Laggner, and T. Langer. Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening. J Chem Inf Model. 45:716-724 (2005).

43. T. Schroeter, A. Schwaighofer, S. Mika, A.T. Laak, D. Suelzle, U. Ganzer, N. Heinrich, and K.R. Muller. Machine Learning Models for Lipophilicity and Their Domain of Applicability. Mol Pharm. 4:524-538 (2007).

44. H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.R. Li, L.Y. Han, H.H. Lin, and Y.Z. Chen. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. J Pharm Sci. (Published Online): (2007).

45. T. Foxand J.M. Kriegl. Machine learning techniques for in silico modeling of drug metabolism. Curr Top Med Chem. 6:1579-1591 (2006).

46. W. Duch, K. Swaminathan, and J. Meller. Artificial intelligence approaches for rational drug design and discovery. Curr Pharm Des. 13:1497-1508 (2007).

47. B. Chen, R.F. Harrison, G. Papadatos, P. Willett, D.J. Wood, X.Q. Lewell, P. Greenidge, and N. Stiefl. Evaluation of machine-learning methods for ligand-based virtual screening. J Comput Aided Mol Des (2007).

48. B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365-370 (2003).

49. H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, and C. Zardecki. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr. 58:899-907 (2002).

50. D.O. Stichtenothand J.C. Frolich. The second generation of COX-2 inhibitors: what advantages do the newest offer? Drugs. 63:33-45 (2003).

51. L.A. Linkinsand J.I. Weitz. Pharmacology and clinical potential of direct thrombin inhibitors. Current Pharmaceutical Design. 11:3877-3884 (2005).

52. S. Ribeiroand R. Horuk. The clinical potential of chemokine receptor antagonists. Pharmacology & Therapeutics. 107:44-58 (2005).

53. A. Spaltenstein, W.M. Kazmierski, J.F. Miller, and V. Samano. Discovery of next generation inhibitors of HIV protease. Current topics in medicinal chemistry. 5:1589-1607 (2005).

54. D. Fabbro, S. Ruetz, E. Buchdunger, S.W. Cowan-Jacob, G. Fendrich, J. Liebetanz, J. Mestan, T. O'Reilly, P. Traxler, B. Chaudhuri, H. Fretz, J. Zimmermann, T. Meyer, G. Caravatti, P. Furet, and P.W. Manley. Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. Pharmacology & Therapeutics. 93:79-98 (2002).

55. R. Kumar, V.P. Singh, and K.M. Baker. Kinase inhibitors for cardiovascular disease. Journal of Molecular and Cellular Cardiology. doi:10.1016/j.yjmcc.2006.09.005: (2006).

56. D.P. Rotella. Phosphodiesterase 5 inhibitors: current status and potential applications. Nature reviews Drug discovery. 1:674-682 (2002).

57. P. Pacherand V. Kecskemeti. Trends in the development of new antidepressants. Is there a light at the end of the tunnel? Current Medicinal Chemistry. 11:925-943 (2004).

58. L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, and G. Schneider. Extraction and visualization of potential pharmacophore points

using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. J Med Chem. 48:6997-7004 (2005).

59. R.N. Jorissenand M.K. Gilson. Virtual screening of molecular databases using a support vector machine. J Chem Inf Model. 45:549-561 (2005).

60. J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model. 46:462-470 (2006).

61. B. Chen, R.F. Harrison, K. Pasupa, P. Willett, D.J. Wilton, D.J. Wood, and X.Q. Lewell. Virtual screening using binary kernel discrimination: effect of noisy training data and the optimization of performance. Journal of Chemical Information and Modeling. 46:478-486 (2006).

62. M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, and J.W. Davies. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. J Chem Inf Model. 46:193-200 (2006).

63. Z. Lepp, T. Kinoshita, and H. Chuman. Screening for new antidepressant leads of multiple activities by support vector machines. J Chem Inf Model. 46:158-167 (2006).

64. G. Harper, J. Bradshaw, J.C. Gittins, D.V. Green, and A.R. Leach. Prediction of biological activity for high-throughput screening using binary kernel discrimination. J Chem Inf Comput Sci. 41:1295-1300 (2001).

65. K. Yamazaki, N. Kusunose, K. Fujita, H. Sato, S. Asano, A. Dan, and M. Kanaoka. Identification of phosphodiesterase-1 and 5 dual inhibitors by a ligand-based virtual screening optimized for lead evolution. Bioorganic & Medicinal Chemistry Letters. 16:1371-1379 (2006).

66. D. Vidal, M. Thormann, and M. Pons. A novel search engine for virtual screening of very large databases. J Chem Inf Model. 46:836-843 (2006).

67. J.C. Mozziconacci, E. Arnoult, P. Bernard, Q.T. Do, C. Marot, and L. Morin-Allory. Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. J Med Chem. 48:1055-1068 (2005).

68. E. Vangrevelinghe, K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro, and P. Furet. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J Med Chem. 46:2656-2662 (2003).

69. I.J. Enyedy, Y. Ling, K. Nacro, Y. Tomita, X. Wu, Y. Cao, R. Guo, B. Li, X. Zhu, Y. Huang, Y.Q. Long, P.P. Roller, D. Yang, and S. Wang. Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. J Med Chem. 44:4313-4324 (2001).

70. T.N. Doman, S.L. McGovern, B.J. Witherbee, T.P. Kasten, R. Kurumbail, W.C. Stallings, D.T. Connolly, and B.K. Shoichet. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem. 45:2213-2221 (2002).

71. M.D. Cummings, R.L. DesJarlais, A.C. Gibbs, V. Mohan, and E.P. Jaeger. Comparison of automated docking programs as virtual screening tools. J Med Chem. 48:962-976 (2005).

72. J.L. Wang, D. Liu, Z.J. Zhang, S. Shan, X. Han, S.M. Srinivasula, C.M. Croce, E.S. Alnemri, and Z. Huang. Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. Proc Natl Acad Sci U S A. 97:7124-7129 (2000).

73. A. Eversand T. Klabunde. 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).

74. N. Stiefland A. Zaliani. A knowledge-based weighting approach to ligand-based virtual screening. J Chem Inf Model. 46:587-596 (2006).

75. D.M. Lorberand B.K. Shoichet. Hierarchical docking of databases of multiple ligand conformations. Curr Top Med Chem. 5:739-749 (2005).

76. B. Pirard, J. Brendel, and S. Peukert. The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. J Chem Inf Model. 45:477-485 (2005).

77. M. Rella, C.A. Rushworth, J.L. Guy, A.J. Turner, T. Langer, and R.M. Jackson. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. J Chem Inf Model. 46:708-716 (2006).

78. K.S. Smalley, N.K. Haass, P.A. Brafford, M. Lioni, K.T. Flaherty, and M. Herlyn. Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. Mol Cancer Ther. 5:1136-1144 (2006).

79. Y. Pilpel, P. Sudarsanam, and G.M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 29:153-159 (2001).

80. R. Muller. Crosstalk of oncogenic and prostanoid signaling pathways. J Cancer Res Clin Oncol. 130:429-444 (2004).

81. N.V. Sergina, M. Rausch, D. Wang, J. Blair, B. Hann, K.M. Shokat, and M.M. Moasser. Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. Nature. 445:437-441 (2007).

82. Christopher M., Overall, and O. Kleifeld. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. Nature Reviews Cancer. 6:227-239 (2006).

83. T. Force, D.S. Krause, and R.A. Van Etten. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. Nat Rev Cancer. 7:332-344 (2007).

84. R. Morphy, C. Kay, and Z. Rankovic. From magic bullets to designed multiple ligands. Drug Discov Today. 9:641-651 (2004).

85. C.T. Keith, A.A. Borisy, and B.R. Stockwell. Multicomponent therapeutics for networked systems. Nat Rev Drug Discov. 4:71-78 (2005).

86. R. Morphyand Z. Rankovic. The physicochemical challenges of designing multiple ligands. J Med Chem. 49:4961-4970 (2006).

87. R. Morphy. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. J Med Chem. 49:2969-2978 (2006).

88. J. Jia, F. Zhu, X. Ma, Z. Cao, Y. Li, and Y.Z. Chen. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov. 8:111-128 (2009).

89. D. Vina, E. Uriarte, F. Orallo, and H. Gonzalez-Diaz. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharm. 6:825-835 (2009).

90. F.J. Prado-Prado, E. Uriarte, F. Borges, and H. Gonzalez-Diaz. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. Eur J Med Chem. 44:4516-4521 (2009).

91. H. Gonzalez-Diazand F.J. Prado-Prado. Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: multispecies activity models for antifungals. J Comput Chem. 29:656-667 (2008).

92. H. Gonzalez-Diaz, F.J. Prado-Prado, L. Santana, and E. Uriarte. Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species. Bioorg Med Chem. 14:5973-5980 (2006).

93. V.G. Petrelli A. Multitarget drugs: the present and the future of cancer therapy. Expert Opin Pharmacother. 10:589-600 (2009).

94. K. Garber. The second wave in kinase cancer drugs. Nat Biotechnol. 24:127-130 (2006).

95. S.K. Kulkarniand A. Dhir. Current investigational drugs for major depression. Expert Opin Investig Drugs. 18:767-788 (2009).

96. J. Overington. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. J Comput Aided Mol Des. 23:195-198 (2009).

97. T. Scior, J.L. Medina-Franco, Q.T. Do, K. Martinez-Mayorga, J.A. Yunes Rojas, and P. Bernard. How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem. 16:4297-4313 (2009).

98. R.G. Susnowand S.L. Dixon. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. J Chem Inf Comput Sci. 43:1308-1315 (2003).

99. J.J. Perez. Managing molecular diversity. Chemical Society Reviews, Vol. 34, Royal Society of Chemistry, 2005, pp. 143-152.

100. W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang, and R. Perkins. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environ Health Perspect. 112:1249-1254 (2004).

101. I. Ijjaali, F. Petitet, E. Dubus, O. Barberan, and A. Michel. Assessing potency of c-Jun N-terminal kinase 3 (JNK3) inhibitors using 2D molecular descriptors and binary QSAR methodology. Bioorg Med Chem. 15:4256-4264 (2007).

102. X.H. Ma, R. Wang, S.Y. Yang, Z.R. Li, Y. Xue, Y.C. Wei, B.C. Low, and Y.Z. Chen. Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. J Chem Inf Model. 48:1227-1237 (2008).

103. L.Y. Han, X.H. Ma, H.H. Lin, J. Jia, F. Zhu, Y. Xue, Z.R. Li, Z.W. Cao, Z.L. Ji, and Y.Z. Chen. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. J Mol Graph Model. 26:1276-1286 (2008).

104. H. Briemand J. Gunther. Classifying "kinase inhibitor-likeness" by using machine-learning methods. Chembiochem. 6:558-566 (2005).

105. X.H. Ma, R. Wang, C.Y. Tan, Y.Y. Jiang, T. Lu, H.B. Rao, X.Y. Li, M.L. Go, B.C. Low, and Y.Z. Chen. Virtual Screening of Selective Multitarget Kinase Inhibitors by Combinatorial Support Vector Machines. Mol Pharm. 7:1545-1560 (2010).

106. C.W. Yap. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 32:1466-1474 (2011).

107. V.C. A Mauri, M Pavan, R Todeschini. Dragon software: An easy approach to molecular descriptor calculations. Match Communications In Mathematical And In Computer Chemistry. 56:237-248 (2006).

108. K.G. Hall LH, Haney DN. *Molconn-Z*, eduSoft LC: Ashland VA, 2002.

109. Z.R. Li, L.Y. Han, Y. Xue, C.W. Yap, H. Li, L. Jiang, and Y.Z. Chen. MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. Biotechnology and Bioengineering. 97:389-396 (2007).

110. C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci. 43:493-500 (2003).

111. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E.L. Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr Pharm Des. 12:2111-2120 (2006).

112. J.K. Wegner. JOELib/JOELib2, Department of Computer Science,University of Tübingen: Germany, 2005.

113. Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, and Y.Z. Chen. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. J Chem Inf Comput Sci. 44:1630-1638 (2004).

114. M.C. Hemmer, V. Steinhauer, and J. Gasteiger. Deriving the 3D structure of organic molecules from their infrared spectra. Vibrational Spectroscopy. 19:151-164 (1999).

115. G. Rückerand C. Rücker. Counts of all walks as atomic and molecular descriptors. Journal of Chemical Information and Computer Sciences. 33:683-695 (1993).

116. J.H. Schuur, P. Setzer, and J. Gasteiger. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. Journal of Chemical Information and Computer Sciences. 36:334-344 (1996).

117. R.S. Pearlmanand K.M. Smith. Metric validation and the receptor-relevant subspace concept. Journal of Chemical Information and Computer Sciences. 39:28-35 (1999).

118. G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, and A. Zaliani. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. Journal of Computer-Aided Molecular Design. 11:79-92 (1997).

119. J. Galvez, R. Garcia, M.T. Salabert, and R. Soler. Charge indexes. New topological descriptors. Journal of Chemical Information and Computer Sciences. 34:520-525 (1994).

120. V. Consonni, R. Todeschini, and M. Pavan. Structure/Response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. Journal of Chemical Information and Computer Sciences. 42:682-692 (2002).

121. M. Randic. Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. Tetrahedron. 31:1477-1481 (1975).

122. M. Randic. Molecular profiles. Novel geometry-dependent molecular descriptors. New Journal of Chemistry. 19:781-791 (1995).

123. L.B. Kierand L.H. Hall. Molecular structure description: The electrotopological state, Academic Press, San Diego, 1999.

124. J.A. Platts, D. Butina, M.H. Abraham, and A. Hersey. Estimation of molecular free energy relation descriptors using a group contribution approach. Journal of Chemical Information and Computer Sciences. 39:835-845 (1999).

125. T. Liu, Y. Lin, X. Wen, R.N. Jorissen, and M.K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 35:D198-201 (2007).

126. J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. J Chem Inf Comput Sci. 34:1000-1008 (1994).

127. D.J. Livingstone. Data analysis for chemists: Applications to QSAR and chemical product design, Oxford University Press, Oxford, 1995.

128. L. Eriksson, E. Johansson, N. Kettaneh-Wold, and K.M. Wade. Multi- and megavariate data analysis - Principles and applications, Umetrics, AB, Umea, Sweden, 2001.

129. P. Larranaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armananzas, G. Santafe, A. Perez, and V. Robles. Machine learning in bioinformatics. Brief Bioinform. 7:86-112 (2006).

130. S.B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249-268 (2007).

131. D. Fisher. Unsupervised Learning. Machine Learning. 45:5-7 (2001).

132. V.N. Vapnik. The Nature of Statistical Learning Theory, Springer-Verlag New York Inc, New York, 1995.

133. C.J.C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery. 2:127-167 (1998).

134. N. Pochet, F. De Smet, J.A. Suykens, and B.L. De Moor. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics. 20:3185-3195 (2004).

135. F. Liand Y. Yang. Analysis of recursive gene selection approaches from microarray data. Bioinformatics. 21:3741-3747 (2005).

136. L.Y.H. J. Cui, H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, and Y.Z. Chen. Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties. Mol Immunol. 44:866-877 (2007).

137. C.W. Yapand Y.Z. Chen. Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. J Pharm Sci. 94:153-168 (2005).

138. C.W. Yapand Y.Z. Chen. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. J Chem Inf Model. 45:982-992 (2005).

139. I.I. Grover, I.I. Singh, and I.I. Bakshi. Quantitative structure-property relationships in pharmaceutical research - Part 2. Pharm Sci Technol Today. 3:50-57 (2000).

140. M.W.B. Trotter, B.F. Buxton, and S.B. Holden. Support vector machines in combinatorial chemistry. Meas Control. 34:235-239 (2001).

141. R. Czerminski, A. Yasri, and D. Hartsough. Use of support vector machine in pattern classification: Application to QSAR studies. Quantitative Structure-Activity Relationships. 20:227-240 (2001).

142. R.A. Johnsonand D.W. Wichern. Applied multivariate statistical analysis, Prentice Hall, Englewood Cliffs, NJ, 1982.

143. E. Fixand J.L. Hodges. Discriminatory analysis: Non-parametric discrimination: Consistency properties, USAF School of Aviation Medicine, Texas, 1951.

144. S. Fujishimaand Y. Takahashi. Classification of dopamine antagonists using TFS-based artificial neural network. J Chem Inf Comput Sci. 44:1006-1009 (2004).

145. J. Bostrom, A. Hogner, and S. Schmitt. Do structurally similar ligands bind in a similar fashion? J Med Chem. 49:6716-6725 (2006).

146. N. Huang, B.K. Shoichet, and J.J. Irwin. Benchmarking sets for molecular docking. J Med Chem. 49:6789-6801 (2006).

147. C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, and Y.Z. Chen. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 31:3692-3697 (2003).

148. L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, and Y.Z. Chen. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res. 32:6437-6444 (2004).

149. H.H. Lin, L.Y. Han, C.Z. Cai, Z.L. Ji, and Y.Z. Chen. Prediction of transporter family from protein sequence by support vector machine approach. Proteins. 62:218-231 (2006).

150. C.J.Z. L.Y. Han, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, and Y.Z. Chen. Support vector machine approach for predicting druggable proteins: Recent progress in its exploarion and investigation of its usefulness. Drug Discovery Today. (accepted): (2007).

151. T.I. Opreaand J. Gottfries. Chemography: the art of navigating in chemical space. J Comb Chem. 3:157-166 (2001).

152. Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, and Y.Z. Chen. Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci. 44:1497-1505 (2004).

153. M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, and H. Waldmann. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci U S A. 102:17272-17277 (2005).

154. M.M.D. H.P. Rang, J.M. Ritter. Pharmacology, Churchill Livingstone, 2001.

155. P.D. Mosierand P.C. Jurs. QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci. 42:1460-1470 (2002).

156. D.M. Hawkins. The problem of overfitting. J Chem Inf Comput Sci. 44:1-12 (2004).

157. S. Woldand L. Eriksson. Statistical validation of QSAR results. In H. Van de Waterbeemd (ed.), Chemometric methods in molecular design, Wiley-VCH, Weinheim; New York, 1995, pp. 309-318.

158. A. Golbraikhand A. Tropsha. Beware of q2! J Mol Graph Model. 20:269-276 (2002).

159. B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 405:442-451 (1975).

160. X.H. Ma, Z. Shi, C. Tan, Y. Jiang, M.L. Go, B.C. Low, and Y.Z. Chen. In-silico approaches to multi-target drug discovery : computer aided multi-target drug design, multi-target virtual screening. Pharm Res. 27:739-749 (2010).

161. H. Kitano. Innovation - A robustness-based approach to systems-oriented drug design. Nature Reviews Drug Discovery. 6:202-210 (2007).

162. P. Legrain, J. Wojcik, and J.M. Gauthier. Protein-protein interaction maps: a lead towards cellular functions. Trends in Genetics. 17:346-352 (2001).

163. J. Downward. The ins and outs of signalling. Nature. 411:759-762 (2001).

164. J.W. Lengeler. Metabolic networks: a signal-oriented approach to cellular models. Biol Chem. 381:911-920 (2000).

165. A. Beyer, S. Bandyopadhyay, and T. Ideker. Integrating physical and genetic maps: from genomes to interaction networks. Nature Reviews Genetics. 8:699-710 (2007).

166. B.L. Drees, B. Sundin, E. Brazeau, J.P. Caviston, G.C. Chen, W. Guo, K.G. Kozminski, M.W. Lau, J.J. Moskow, A. Tong, L.R. Schenkman, A. McKenzie, 3rd, P. Brennwald, M. Longtine, E. Bi, C. Chan, P. Novick, C. Boone, J.R. Pringle, T.N. Davis, S. Fields, and D.G. Drubin. A protein interaction map for cell polarity development. J Cell Biol. 154:549-571 (2001).

167. A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 415:141-147 (2002).

168. J. Qian, J. Lin, N.M. Luscombe, H. Yu, and M. Gerstein. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics. 19:1917-1926 (2003).

169. S.L. Lo, C.Z. Cai, Y.Z. Chen, and M.C.M. Chung. Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics. 5:876-884 (2005).

170. E. Phizicky, P.I. Bastiaens, H. Zhu, M. Snyder, and S. Fields. Protein analysis on a proteomic scale. Nature. 422:208-215 (2003).

171. M. Pellegrini, E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96:4285-4288 (1999).

172. T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23:324-328 (1998).

173. H.W. Mewes, D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30:31-34 (2002).

174. L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research. 32:D449-D451 (2004).

175. C. Alfarano, C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S.

Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore, B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie, R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao, Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F.F. Ouellette, and C.W.V. Hogue. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Research. 33:D418-D424 (2005).

176. P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahren, S. Tsoka, N. Darzentas, V. Kunin, and N. Lopez-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Research. 33:6083-6089 (2005).

177. A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni. MINT: a Molecular INTeraction database. Febs Letters. 513:135-140 (2002).

178. N. Le Novere, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J.L. Snoep, and M. Hucka. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Research. 34:D689-D691 (2006).

179. C. von Mering, L.J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. STRING 7 - recent developments in the integration and prediction of protein interactions. Nucleic Acids Research. 35:D358-D362 (2007).

180. S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, and H. Hermjakob. IntAct - open source resource for molecular interaction data. Nucleic Acids Research. 35:D561-D565 (2007).

181. S. Okuda, T. Yamada, M. Hamajima, M. Itoh, T. Katayama, P. Bork, S. Goto, and M. Kanehisa. KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Research. 36:W423-W426 (2008).

182. B.J. Breitkreutz, C. Stark, T. Reguly, L. Boucher, A. Breitkreutz, M. Livstone, R. Oughtred, D.H. Lackner, J. Bahler, V. Wood, K. Dolinski, and M. Tyers. The BioGRID interaction database: 2008 update. Nucleic Acids Research. 36:D637-D640 (2008).

183. R. Linding, L.J. Jensen, A. Pasculescu, M. Olhovsky, K. Colwill, P. Bork, M.B. Yaffe, and T. Pawson. NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Research. 36:D695-D699 (2008).

184. M. Kuhn, C. von Mering, M. Campillos, L.J. Jensen, and P. Bork. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Research. 36:D684-D688 (2008).

185. B. Raghavachari, A. Tasneem, T.M. Przytycka, and R. Jothi. DOMINE: a database of protein domain interactions. Nucleic Acids Research. 36:D656-D661 (2008).

186. H.C. Mak, M. Daly, B. Gruebel, and T. Ideker. CellCircuits: a database of protein network models. Nucleic Acids Research. 35:D538-D545 (2007).

187. G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33:D428-432 (2005).

188. S. Goto, Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 30:402-404 (2002).

189. M. Fussenegger, J.E. Bailey, and J. Varner. A mathematical model of caspase function in apoptosis. Nat Biotechnol. 18:768-774 (2000).

190. J.M. Haugh, A. Wells, and D.A. Lauffenburger. Mathematical modeling of epidermal growth factor receptor signaling through the phospholipase C pathway: mechanistic insights and predictions for molecular interventions. Biotechnol Bioeng. 70:225-238 (2000).

191. H. Sahm, L. Eggeling, and A.A. de Graaf. Pathway analysis and metabolic engineering in Corynebacterium glutamicum. Biol Chem. 381:899-910 (2000).

192. B. van den Broek, M.C. Noom, and G.J. Wuite. DNA-tension dependence of restriction enzyme activity reveals mechanochemical properties of the reaction pathway. Nucleic Acids Res. 33:2676-2684 (2005).

193. B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, and G. Muller. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. Nature Biotechnology. 20:370-375 (2002).

194. S. Sasagawa, Y. Ozaki, K. Fujita, and S. Kuroda. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. Nature Cell Biology. 7:365-U331 (2005).

195. M.R. Birtwistle, M. Hatakeyama, N. Yumoto, B.A. Ogunnaike, J.B. Hoek, and B.N. Kholodenko. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. Molecular Systems Biology. 3: (2007).

196. C.Y. Ung, H. Li, X.H. Ma, J. Jia, B.W. Li, B.C. Low, and Y.Z. Chen. Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk. Febs Letters. 582:2283-2290 (2008).

197. B.C.V. Suresh, S.M.E. Babar, E.J. Song, E. Oh, and Y.S. Yoo. Kinetic analysis of the MAPK and PI3K/Akt signaling pathways. Molecules and Cells. 25:397-406 (2008).

198. G. Altan-Bonnetand R.N. Germain. Modeling T cell antigen discrimination based on feedback control of digital ERK responses. Plos Biology. 3:1925-1938 (2005).

199. B.J. Bornstein, S.M. Keating, A. Jouraku, and M. Hucka. LibSBML: an API library for SBML. Bioinformatics. 24:880-881 (2008).

200. A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano. CellDesigner 3.5: A versatile modeling tool for biochemical networks. Proceedings of the Ieee. 96:1254-1265 (2008).

201. S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI- A COmplex PAthway SImulator. Bioinformatics. 22:3067-3074 (2006).

202. E.G. Cerami, G.D. Bader, B.E. Gross, and C. Sander. cPath: open source software for collecting, storing, and querying biological pathways. BMC Bioinformatics. 7: (2006).

203. A. Ludemann, D. Weicht, J. Selbig, and J. Kopka. PaVESy: Pathway visualization and editing system. Bioinformatics. 20:2841-2844 (2004).

204. R. Nicolas, M. Donizelli, and N. Le Novere. SBMLeditor: effective creation of models in the Systems Biology Markup Language (SBML). Bmc Bioinformatics. 8: (2007).

205. T. Franch, M. Petersen, E.G.H. Wagner, J.P. Jacobsen, and K. Gerdes. Antisense RNA regulation in prokaryotes: Rapid RNA/RNA interaction facilitated by a general U-turn loop structure. Journal of Molecular Biology. 294:1115-1125 (1999).

206. N.L. Korneeva, B.J. Lamphear, F.L.C. Hennigan, W.C. Merrick, and R.E. Rhoads. Characterization of the two eIF4A-binding sites on human eIF4G-1. Journal of Biological Chemistry. 276:2872-2879 (2001).

207. M. Hoshino, Y. Kawata, and Y. Goto. Interaction of GroEL with conformational states of horse cytochrome c. Journal of Molecular Biology. 262:575-587 (1996).

208. M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J.C. Doyle, H. Kitano, A.P. Arkin, B.J. Bornstein, D. Bray, A. Cornish-Bowden, A.A. Cuellar, S. Dronov, E.D. Gilles, M. Ginkel, V. Gor, Goryanin, II, W.J. Hedley, T.C. Hodgman, J.H. Hofmeyr, P.J. Hunter, N.S. Juty, J.L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L.M. Loew, D. Lucio, P. Mendes, E. Minch, E.D. Mjolsness, Y. Nakayama, M.R. Nelson, P.F. Nielsen, T. Sakurada, J.C. Schaff, B.E. Shapiro, T.S. Shimizu, H.D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and S. Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 19:524-531 (2003).

209. R. Alves, F. Antunes, and A. Salvador. Tools for kinetic modeling of biochemical networks. Nature Biotechnology. 24:667-672 (2006).

210. A. Deckard, F.T. Bergmann, and H.M. Sauro. Supporting the SBML layout extension. Bioinformatics. 22:2966-2967 (2006).

211. H. Schmidt, G. Drews, J. Vera, and O. Wolkenhauer. SBML export interface for the systems biology toolbox for MATLAB. Bioinformatics. 23:1297-1298 (2007).

212. Z. Ziand E. Klipp. SBML-PET: A systems biology markup language-based parameter estimation tool. Bioinformatics. 22:2704-2705 (2006).

213. C. Smith. Drug target validation: Hitting the target. Nature. 422:341, 343, 345 passim (2003).

214. D.A. Winkler. The role of quantitative structure--activity relationships (QSAR) in biomolecular discovery. Brief Bioinform. 3:73-86 (2002).

215. G.R. Zimmermann, J. Lehar, and C.T. Keith. Multi-target therapeutics: when the whole is greater than the sum of the parts. Drug Discov Today. 12:34-42 (2007).

216. F. Zhu, C. Qin, L. Tao, X. Liu, Z. Shi, X. Ma, J. Jia, Y. Tan, C. Cui, J. Lin, C. Tan, Y. Jiang, and Y. Chen. Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. Proc Natl Acad Sci U S A. 108:12943-12948 (2011).

217. A. Aroraand E.M. Scholar. Role of tyrosine kinase inhibitors in cancer therapy. J Pharmacol Exp Ther. 315:971-979 (2005).

218. G. Vlahovicand J. Crawford. Activation of tyrosine kinases in cancer. Oncologist. 8:531-538 (2003).

219. M.K. Pauland A.K. Mukhopadhyay. Tyrosine kinase - Role and significance in Cancer. Int J Med Sci. 1:101-115 (2004).

220. A. Petrelliand G. Valabrega. Multitarget drugs: the present and the future of cancer therapy. Expert Opin Pharmacother. 10:589-600 (2009).

221. X. Zhangand A. Fernandez. In silico drug profiling of the human kinome based on a molecular marker for cross reactivity. Mol Pharm. 5:728-738 (2008).

222. M.E. Noble, J.A. Endicott, and L.N. Johnson. Protein kinase inhibitors: insights into drug design from structure. Science. 303:1800-1805 (2004).

223. A. Vema, S.K. Panigrahi, G. Rambabu, B. Gopalakrishnan, J.A. Sarma, and G.R. Desiraju. Design of EGFR kinase inhibitors: a ligand-based approach and its confirmation with structure-based studies. Bioorg Med Chem. 11:4643-4653 (2003).

224. H. Yu, Z. Wang, L. Zhang, J. Zhang, and Q. Huang. Pharmacophore modeling and in silico screening for new KDR kinase inhibitors. Bioorg Med Chem Lett. 17:2126-2133 (2007).

225. K. Matsuno, M. Ichimura, T. Nakajima, K. Tahara, S. Fujiwara, H. Kase, J. Ushiki, N.A. Giese, A. Pandey, R.M. Scarborough, N.A. Lokker, J.C. Yu, J. Irie, E. Tsukuda, S. Ide, S. Oda, and Y. Nomoto. Potent and selective inhibitors of platelet-derived growth factor receptor phosphorylation. 1. Synthesis, structure-activity relationship, and biological effects of a new class of quinazoline derivatives. J Med Chem. 45:3057-3066 (2002).

226. A.M. Thompson, C.J. Connolly, J.M. Hamby, S. Boushelle, B.G. Hartl, A.M. Amar, A.J. Kraker, D.L. Driscoll, R.W. Steinkampf, S.J. Patmore, P.W. Vincent, B.J. Roberts, W.L. Elliott, W. Klohs, W.R. Leopold, H.D. Showalter, and W.A. Denny. 3-(3,5-Dimethoxyphenyl)-1,6-naphthyridine-2,7-diamines and related 2-urea derivatives are potent and selective inhibitors of the FGF receptor-1 tyrosine kinase. J Med Chem. 43:4200-4211 (2000).

227. D. Dalgarno, T. Stehle, S. Narula, P. Schelling, M.R. van Schravendijk, S. Adams, L. Andrade, J. Keats, M. Ram, L. Jin, T. Grossman, I. MacNeil, C. Metcalf, 3rd, W. Shakespeare, Y. Wang, T. Keenan, R. Sundaramoorthi, R. Bohacek, M. Weigele, and T. Sawyer. Structural basis of Src tyrosine

kinase inhibition with a new class of potent and selective trisubstituted purine-based compounds. Chem Biol Drug Des. 67:46-57 (2006).

228. L. Abbott, P. Betschmann, A. Burchat, D.J. Calderwood, H. Davis, P. Hrnciar, G.C. Hirst, B. Li, M. Morytko, K. Mullen, and B. Yang. Discovery of thienopyridines as Src-family selective Lck inhibitors. Bioorg Med Chem Lett. 17:1167-1171 (2007).

229. H.D. Showalter, A.D. Sercel, B.M. Leja, C.D. Wolfangel, L.A. Ambroso, W.L. Elliott, D.W. Fry, A.J. Kraker, C.T. Howard, G.H. Lu, C.W. Moore, J.M. Nelson, B.J. Roberts, P.W. Vincent, W.A. Denny, and A.M. Thompson. Tyrosine kinase inhibitors. 6. Structure-activity relationships among N- and 3-substituted 2,2'-diselenobis(1H-indoles) for inhibition of protein tyrosine kinases and comparative in vitro and in vivo studies against selected sulfur congeners. J Med Chem. 40:413-426 (1997).

230. T. Asano, T. Yoshikawa, T. Usui, H. Yamamoto, Y. Yamamoto, Y. Uehara, and H. Nakamura. Benzamides and benzamidines as specific inhibitors of epidermal growth factor receptor and v-Src protein tyrosine kinases. Bioorg Med Chem. 12:3529-3542 (2004).

231. J. Caballero, M. Fernandez, M. Saavedra, and F.D. Gonzalez-Nilo. 2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-d]pyrimidine derivatives. Bioorg Med Chem. 16:810-821 (2008).

232. X.H. Ma, J. Jia, F. Zhu, Y. Xue, Z.R. Li, and Y.Z. Chen. Comparative Analysis of Machine Learning Methods in Ligand Based Virtual Screening of Large Compound Libraries. Comb Chem High Throughput Screen:(accepted) (2009).

233. J.F. Carvalho, M. Blank, and Y. Shoenfeld. Vascular endothelial growth factor (VEGF) in autoimmune diseases. J Clin Immunol. 27:246-256 (2007).

234. S. Daouti, B. Latario, S. Nagulapalli, F. Buxton, S. Uziel-Fusi, G.W. Chirn, D. Bodian, C. Song, M. Labow, M. Lotz, J. Quintavalla, and C. Kumar. Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis. Osteoarthritis Cartilage. 13:508-518 (2005).

235. M.A. Meynand T.E. Smithgall. Small molecule inhibitors of Lck: the search for specificity within a kinase family. Mini Rev Med Chem. 8:628-637 (2008).

236. T. Raj, P. Kanellakis, G. Pomilio, G. Jennings, A. Bobik, and A. Agrotis. Inhibition of fibroblast growth factor receptor signaling attenuates atherosclerosis in apolipoprotein E-deficient mice. Arterioscler Thromb Vasc Biol. 26:1845-1851 (2006).

237. M.J. Millan. Dual- and triple-acting agents for treating core and co-morbid symptoms of major depression: novel concepts, new drugs. Neurotherapeutics. 6:53-77 (2009).

238. X.H. Ma, C.J. Zheng, L.Y. Han, B. Xie, J. Jia, Z.W. Cao, Y.X. Li, and Y.Z. Chen. Synergistic therapeutic actions of herbal ingredients and their

mechanisms from molecular interaction and network perspectives. Drug Discov Today. 14:579-588 (2009).

239. L.D. Jayanthiand S. Ramamoorthy. Regulation of monoamine transporters: influence of psychostimulants and therapeutic antidepressants. Aaps J. 7:E728-738 (2005).

240. L.C. Daws. Unfaithful neurotransmitter transporters: focus on serotonin uptake and implications for antidepressant efficacy. Pharmacol Ther. 121:89-99 (2009).

241. M.D.A. Gavin A. Whitlock, Alan D. Brown, Paul V. Fish, Alan Stobie, Florian Wakenhut. Design of Monoamine Reuptake Inhibitors: SSRIs, SNRIs and NRIs. Transporters as Targets for Drugs Topics in Medicinal Chemistry. 4:42 (2009).

242. C. Davidsonand J.A. Stamford. Evidence that 5-hydroxytryptamine release in rat dorsal raphe nucleus is controlled by 5-HT1A, 5-HT1B and 5-HT1D autoreceptors. Br J Pharmacol. 114:1107-1109 (1995).

243. L. Romero, I. Hervas, and F. Artigas. The 5-HT1A antagonist WAY-100635 selectively potentiates the presynaptic effects of serotonergic antidepressants in rat brain. Neurosci Lett. 219:123-126 (1996).

244. F. Artigas, L. Romero, C. de Montigny, and P. Blier. Acceleration of the effect of selected antidepressant drugs in major depression by 5-HT1A antagonists. Trends Neurosci. 19:378-383 (1996).

245. E. Schlicker, R. Betz, and M. Gothert. Histamine H3 receptor-mediated inhibition of serotonin release in the rat brain cortex. Naunyn Schmiedebergs Arch Pharmacol. 337:588-590 (1988).

246. S. Threlfell, S.J. Cragg, I. Kallo, G.F. Turi, C.W. Coen, and S.A. Greenfield. Histamine H3 receptors inhibit serotonin release in substantia nigra pars reticulata. J Neurosci. 24:8704-8710 (2004).

247. K.S. Ly, M.A. Letavic, J.M. Keith, J.M. Miller, E.M. Stocking, A.J. Barbier, P. Bonaventure, B. Lord, X. Jiang, J.D. Boggs, L. Dvorak, K.L. Miller, D. Nepomuceno, S.J. Wilson, and N.I. Carruthers. Synthesis and biological activity of piperazine and diazepane amides that are histamine H3 antagonists and serotonin reuptake inhibitors. Bioorg Med Chem Lett. 18:39-43 (2008).

248. G.A. Whitlock, J. Blagg, and P.V. Fish. 1-(2-Phenoxyphenyl)methanamines: SAR for dual serotonin/noradrenaline reuptake inhibition, metabolic stability and hERG affinity. Bioorg Med Chem Lett. 18:596-599 (2008).

249. M.J. Fray, P.V. Fish, G.A. Allan, G. Bish, N. Clarke, R. Eccles, A.C. Harrison, J.L. Le Net, S.C. Phillips, N. Regan, C. Sobry, A. Stobie, F. Wakenhut, D. Westbrook, S.L. Westbrook, and G.A. Whitlock. Second generation N-(1,2-diphenylethyl)piperazines as dual serotonin and noradrenaline reuptake inhibitors: improving metabolic stability and reducing ion channel activity. Bioorg Med Chem Lett. 20:3788-3792 (2010).

250. K. Takeuchi, T.J. Kohn, N.A. Honigschmidt, V.P. Rocco, P.G. Spinazze, S.K. Hemrick-Luecke, L.K. Thompson, D.C. Evans, K. Rasmussen, D.

Koger, D. Lodge, L.J. Martin, J. Shaw, P.G. Threlkeld, and D.T. Wong. Advances toward new antidepressants beyond SSRIs: 1-aryloxy-3-piperidinylpropan-2-ols with dual 5-HT1A receptor antagonism/SSRI activities. Part 5. Bioorg Med Chem Lett. 16:2347-2351 (2006).

251. Z. Shen, P. Siva Ramamoorthy, N.T. Hatzenbuhler, D.A. Evrard, W. Childers, B.L. Harrison, M. Chlenov, G. Hornby, D.L. Smith, K.M. Sullivan, L.E. Schechter, and T.H. Andree. Synthesis and structure-activity relationship of novel lactam-fused chroman derivatives having dual affinity at the 5-HT(1A) receptor and the serotonin transporter. Bioorg Med Chem Lett. 20:222-227 (2010).

252. L. Matzen, C. van Amsterdam, W. Rautenberg, H.E. Greiner, J. Harting, C.A. Seyfried, and H. Bottcher. 5-HT reuptake inhibitors with 5-HT(1B/1D) antagonistic activity: a new approach toward efficient antidepressants. J Med Chem. 43:1149-1157 (2000).

253. M.J. Millan, M. Brocco, A. Gobert, and A. Dekeyne. Anxiolytic properties of agomelatine, an antidepressant with melatoninergic and serotonergic properties: role of 5-HT2C receptor blockade. Psychopharmacology (Berl). 177:448-458 (2005).

254. S. Chaki, Y. Oshida, S. Ogawa, T. Funakoshi, T. Shimazaki, T. Okubo, A. Nakazato, and S. Okuyama. MCL0042: a nonpeptidic MC4 receptor antagonist and serotonin reuptake inhibitor with anxiolytic- and antidepressant-like activity. Pharmacol Biochem Behav. 82:621-626 (2005).

255. T. Ryckmans, L. Balancon, O. Berton, C. Genicot, Y. Lamberty, B. Lallemand, P. Pasau, N. Pirlot, L. Quere, and P. Talaga. First dual NK(1) antagonists-serotonin reuptake inhibitors: synthesis and SAR of a new class of potential antidepressants. Bioorg Med Chem Lett. 12:261-264 (2002).

256. A. Rupp, K.A. Kovar, G. Beuerle, C. Ruf, and G. Folkers. A new pharmophoric model for 5-HT reuptake-inhibitors: differentiation of amphetamine analogues. Pharm Acta Helv. 68:235-244 (1994).

257. R. Bureau, C. Daveu, J.C. Lancelot, and S. Rault. Molecular design based on 3D-pharmacophore. Application to 5-HT subtypes receptors. J Chem Inf Comput Sci. 42:429-436 (2002).

258. C.Y. Kim, P.E. Mahaney, O. McConnell, Y. Zhang, E. Manas, D.M. Ho, D.C. Deecher, and E.J. Trybulski. Discovery of a new series of monoamine reuptake inhibitors, the 1-amino-3-(1H-indol-1-yl)-3-phenylpropan-2-ols. Bioorg Med Chem Lett. 19:5029-5032 (2009).

259. D.J. O'Neill, A. Adedoyin, P.D. Alfinito, J.A. Bray, S. Cosmi, D.C. Deecher, A. Fensome, J. Harrison, L. Leventhal, C. Mann, C.C. McComas, N.R. Sullivan, T.B. Spangler, A.J. Uveges, E.J. Trybulski, G.T. Whiteside, and P. Zhang. Discovery of novel selective norepinephrine reuptake inhibitors: 4-[3-aryl-2,2-dioxido-2,1,3-benzothiadiazol-1(3H)-yl]-1-(methylamino)butan -2-ols (WYE-103231). J Med Chem. 53:4511-4521 (2010).

260. A.J. Bojarski. Pharmacophore models for metabotropic 5-HT receptor ligands. Curr Top Med Chem. 6:2005-2026 (2006).

261. K.C. Weber, L.B. Salum, K.M. Honorio, A.D. Andricopulo, and A.B. da Silva. Pharmacophore-based 3D QSAR studies on a series of high affinity 5-HT1A receptor ligands. Eur J Med Chem. 45:1508-1514 (2010).

262. S. Lorenzi, M. Mor, F. Bordi, S. Rivara, M. Rivara, G. Morini, S. Bertoni, V. Ballabeni, E. Barocelli, and P.V. Plazzi. Validation of a histamine H3 receptor model through structure-activity relationships for classical H3 antagonists. Bioorg Med Chem. 13:5647-5657 (2005).

263. B. Schlegel, C. Laggner, R. Meier, T. Langer, D. Schnell, R. Seifert, H. Stark, H.D. Holtje, and W. Sippl. Generation of a homology model of the human histamine H(3) receptor for ligand docking and pharmacophore-based screening. J Comput Aided Mol Des. 21:437-453 (2007).

264. N. Dessalew. QSAR study on dual SET and NET reuptake inhibitors: an insight into the structural requirement for antidepressant activity. J Enzyme Inhib Med Chem. 24:262-271 (2009).

265. F. Micheli, P. Cavanni, D. Andreotti, R. Arban, R. Benedetti, B. Bertani, M. Bettati, L. Bettelini, G. Bonanomi, S. Braggio, R. Carletti, A. Checchia, M. Corsi, E. Fazzolari, S. Fontana, C. Marchioro, E. Merlo-Pich, M. Negri, B. Oliosi, E. Ratti, K.D. Read, M. Roscic, I. Sartori, S. Spada, G. Tedesco, L. Tarsi, S. Terreni, F. Visentini, A. Zocchi, L. Zonzini, and R. Di Fabio. 6-(3,4-dichlorophenyl)-1-[(methyloxy)methyl]-3-azabicyclo[4.1.0]heptane: a new potent and selective triple reuptake inhibitor. J Med Chem. 53:4989-5001 (2010).

266. A. Bender. Databases: Compound bioactivities go public. Nature Chemical Biology. 309: (2010).

267. T.I. Opreaand J. Gottfries. Chemography: the art of navigating in chemical space. J Comb Chem. 3:157-166 (2001).

268. T.F.a.J.-L. Reymond. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. J Chem Inf Model. 47:342-353 (2007).

269. M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, and H. Waldmann. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci USA. 102:17272-17277 (2005).

270. M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, and J.W. Davies. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. J Chem Inf Model. 46:193-200 (2006).

271. B. Rost. Twilight zone of protein sequence alignments. Protein Eng. 12:85-94 (1999).

272. E. Hamel. Serotonin and migraine: biology and clinical implications. Cephalalgia. 27:1293-1300 (2007).

273. T.A. Smitherman, A.B. Walters, M. Maizels, and D.B. Penzien. The Use of Antidepressants for Headache Prophylaxis. CNS Neurosci Ther (2010).

274. S.M. Stahland D.K. Shayegan. The psychopharmacology of ziprasidone: receptor-binding properties and real-world psychiatric practice. J Clin Psychiatry. 64 Suppl 19:6-12 (2003).

275. Y. Chertkow, O. Weinreb, M.B. Youdim, and H. Silver. Molecular mechanisms underlying synergistic effects of SSRI-antipsychotic augmentation in treatment of negative symptoms in schizophrenia. J Neural Transm. 116:1529-1541 (2009).

276. T.D. Heightman, L.M. Gaster, S.L. Pardoe, J.P. Pilleux, M.S. Hadley, D.N. Middlemiss, G.W. Price, C. Roberts, C.M. Scott, J.M. Watson, L.J. Gordon, V.A. Holland, J. Powles, G.J. Riley, T.O. Stean, B.K. Trail, N. Upton, N.E. Austin, A.D. Ayrton, T. Coleman, and L. Cutler. 8-Piperazinyl-2,3-dihydropyrrolo[3,2-g]isoquinolines: potent, selective, orally bioavailable 5-HT1 receptor ligands. Bioorg Med Chem Lett. 15:4370-4374 (2005).

277. M. Bykuand R.L. Gannon. Effects of the 5HT1A agonist/antagonist BMY 7378 on light-induced phase advances in hamster circadian activity rhythms during aging. J Biol Rhythms. 15:300-305 (2000).

278. P. Zajdel, G. Subra, A.J. Bojarski, B. Duszynska, E. Tatarczynska, A. Nikiforuk, E. Chojnacka-Wojcik, M. Pawlowski, and J. Martinez. Novel class of arylpiperazines containing N-acylated amino acids: their synthesis, 5-HT1A, 5-HT2A receptor affinity, and in vivo pharmacological evaluation. Bioorg Med Chem. 15:2907-2919 (2007).

279. A. Slassi. Recent advances in 5-HT1B/1D receptor antagonists and agonists and their potential therapeutic applications. Curr Top Med Chem. 2:559-574 (2002).

280. G. McCort, C. Hoornaert, M. Aletru, C. Denys, O. Duclos, C. Cadilhac, E. Guilpain, G. Dellac, P. Janiak, A.M. Galzin, M. Delahaye, F. Guilbert, and S. O'Connor. Synthesis and SAR of 3- and 4-substituted quinolin-2-ones: discovery of mixed 5-HT(1B)/5-HT(2A) receptor antagonists. Bioorg Med Chem. 9:2129-2137 (2001).

281. T. Heinrich, H. Bottcher, K. Schiemann, G. Holzemann, M. Schwarz, G.D. Bartoszyk, C. van Amsterdam, H.E. Greiner, and C.A. Seyfried. Dual 5-HT1A agonists and 5-HT re-uptake inhibitors by combination of indole-butyl-amine and chromenonyl-piperazine structural elements in a single molecular entity. Bioorg Med Chem. 12:4843-4852 (2004).

282. R.E. Hubbard. Structure-based drug discovery and protein targets in the CNS. Neuropharmacology. 60:7-23 (2011).

283. T. Shimamura, M. Shiroishi, S. Weyand, H. Tsujimoto, G. Winter, V. Katritch, R. Abagyan, V. Cherezov, W. Liu, G.W. Han, T. Kobayashi, R.C. Stevens, and S. Iwata. Structure of the human histamine H1 receptor complex with doxepin. Nature. 475:65-70 (2011).

284. P. Willett. Chemical Similarity Searching. J Chem Inf Comput Sci. 38:983-996 (1998).

285. J. Liand P. Gramatica. Classification and Virtual Screening of Androgen Receptor Antagonists. J Chem Inf Model.

286. S. Derksen, O. Rau, P. Schneider, M. Schubert-Zsilavecz, and G. Schneider. Virtual screening for PPAR modulators using a probabilistic neural network. ChemMedChem. 1:1346-1350 (2006).

287. Y. Wangand J. Bajorath. Advanced fingerprint methods for similarity searching: balancing molecular complexity effects. Comb Chem High Throughput Screen. 13:220-228 (2010).

288. J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model. 46:462-470 (2006).

289. P. Willett. Similarity searching using 2D structural fingerprints. Methods Mol Biol. 672:133-158 (2011).

290. D.R. Flower. On the properties of bit string-based measures of chemical similarity. J Chem Inf Comput Sci. 38:8 (1998).

291. M. Krugand A. Hilgeroth. Recent advances in the development of multi-kinase inhibitors. Mini Rev Med Chem. 8:1312-1327 (2008).

292. A.L. Gill, M. Verdonk, R.G. Boyle, and R. Taylor. A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development. Curr Top Med Chem. 7:1408-1422 (2007).

293. A. Bender, J.L. Jenkins, M. Glick, Z. Deng, J.H. Nettles, and J.W. Davies. "Bayes affinity fingerprints" improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? J Chem Inf Model. 46:2445-2456 (2006).

294. A. Givehchi, A. Bender, and R.C. Glen. Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. J Chem Inf Model. 46:1078-1083 (2006).

295. S. Renner, S. Derksen, S. Radestock, and F. Morchen. Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. J Chem Inf Model. 48:319-332 (2008).

296. D. Erhan, J. L'Heureux P, S.Y. Yue, and Y. Bengio. Collaborative filtering on a family of biological targets. J Chem Inf Model. 46:626-635 (2006).

297. H. Dragos, M. Gilles, and V. Alexandre. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model. 49:1762-1776 (2009).

298. X.H. Liu, X.H. Ma, C.Y. Tan, Y.Y. Jiang, M.L. Go, B.C. Low, and Y.Z. Chen. Virtual screening of Abl inhibitors from large compound libraries by support vector machines. J Chem Inf Model. 49:2101-2110 (2009).

299. H. Kitano. Systems biology: a brief overview. Science. 295:1662-1664 (2002).

300. A. Finney, Hucka, M., Bornstein, B.J., Keating, S.M., Shapiro, B.E., Matthews, J., Kovitz, B.L., Schilstra, M.J., Funahashi, A., Doyle, J.C., Kitano, H. Software Infrastructure for Effective Communication and

Reuse of Computational Models. Systems Modeling in Cell Biology: From Concepts to Nuts and Bolts:369-378 (2006).