HIGH PERFORMANCE COMPUTATIONAL VIRTUAL SCREENING TOOLS: DEVELOPMENT AND APPLICATION TO THE DISCOVERY OF KINASE INHIBITORS

MA XIAOHUA

(M.Sc, Sichuan Univ.; B.Sc, Sichuan Univ.)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHARMACY

NATIONAL UNIVERSITY OF SINGAPORE

2010

Acknowledgements

First and foremost, I wish to express my heartfelt appreciation to my supervisor, Prof. Chen Yu Zong, who provides me with excellent guidance, invaluable advices and suggestions throughout my Ph.D study. I have tremendously benefited from his profound knowledge, expertise in scientific research, as well as his enormous support, which will inspire and motivate me to go further in my future professional career.

I would also like to thank Prof. Low Boon Chuan. As my co-supervisor, he gave me many valuable comments on my research and kind suggestions for my career. His insights and knowledge always give me new idea during our discussion. Great thanks also go to Prof. YAP Chun Wei for his great supports and encouragements.

I also wish to thank present and previous BIDD group members. In particulars, I would like to thank Dr. Li Hu, Dr. Han Lianyi, Dr. Lin Honghuang, Dr. Zhang Hailei, Dr. Wang Rong, Ms Jia jia, Mr Zhu Feng, Mr Liu Xianghui, Ms Liu Xin, Ms Shi Zhe, Ms Sit Win Yee, Mr Han Bucong, Mr Zhang Jingxian, Ms Wei Xiaona, Ms Huang Lu, Mr Guo Yangfan, Mr Tao Lin, Mr Zhang Cheng, Ms Qin Chu, etc. I am really thankful for their valuable suggestions and support in my project, as well as enjoy the close friendship among us.

Last, but not the least, I am profoundly grateful to my parents, my husband and my son for their encouragement and accompany.

Ma Xiaohua

Aug 2010

Table of Contents

Summary

Virtual screening (VS) can provide valuable contributions in hit and lead compound discovery. Numerous software tools have been developed for this purpose. However, the insufficient coverage of compound diversity, high false positive, high false negative prediction and lower speed of screening compound libraries are also required to address in the development of virtual screening methods. In this work, training-sets of diverse inactive compounds are used to improve the performance of Support vector machine (SVM) virtual screening tools. In retrospective database screening of active compounds of single mechanism (HIV protease inhibitors, DHFR inhibitors, dopamine antagonists) and multiple mechanisms (CNS active agents) from large libraries of 2.986 million compounds, the yields, hit-rates, and enrichment factors of our SVM models are compared to those of structure-based VS and other ligand-based VS tools in screening libraries of $\geqslant$1 million compounds. The hit-rates are comparable and the enrichment factors are substantially better than the best results of other VS tools. SVM appears to be potentially useful for facilitating lead discovery in VS of large compound libraries.

Virtual screening performance of SVM depends on the diversity of training active and inactive compounds. We also evaluated the performance of SVM trained by sparsely distributed actives in six MDDR biological target classes composed of high number of known actives of high, intermediate, and low structural diversity. The results show SVM has substantial capability in identifying novel active compounds from sparse active datasets at low false-hit rates.

c-Src and VEGFR-2 are two important kinases that play various roles in tumour progression, invasion, metastasis, angiogenesis and survival. The successes of their inhibitors and the encountered problems have led to further efforts for discovering new inhibitors for c-Src and VEGFR-2. We applied our developed SVM based virtual screening tools for searching c-Src and VEGFR-2 inhibitors from large compound libraries. SVM models showed around 60% accuracy for independent testing sets and >99.9% accuracy for non-inhibitors (very low false hit-rate) that is favorable for selecting potential leads to further study in wet-lab experiment.

Multi-target agents have been increasingly explored for enhancing therapeutic efficacies and improving safety and resistance profiles by selectively modulating the elements of these counter-target and toxicity activities. In the final part of my thesis, combinatorial support vector machines (C-SVMs), virtual screening tools for searching multi-target agents are developed based on our previous high performance SVM based virtual screening tools. C-SVMs models were tested for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). Moreover, C-SVMs were compared to other VS methods DOCK Blaster, kNN and PNN against the same sets of kinase inhibitors and 1.02M Zinc clean-leads dataset. C-SVMs produced comparable dual-inhibitor yields, slightly better false-hit rates for kinase inhibitors, and significantly lower false-hit rates for the Zinc clean-leads dataset.

List of Tables

List of Figures

List of Acronyms

CDK1	Cyclin dependent kinase 1
CDK2	Cyclin dependent kinase 2
CNS	Central nervous system
C-SVMs	Combinatorial support vector machines
DHFR	Dihydrofolate reductase
EGFR	Epidermal growth factor receptor
FGFR	Fibroblast Growth Factor Receptor
FN	False negative
FP	False positive
GSK3	Glycogen synthase kinase 3
HTS	High throughput screening
kNN	k-nearest neighbors
LBVS	Ligand-based Virtual Screening
Lck	Lymphocyte-specific protein tyrosine kinase
MCC	Matthews correlation coefficient
MDDR	MDL Drug Data Report
ML	Machine Learning
N	Negative
NCEs	Novel chemical entities
P	Positive
PDGFR	Platelet-derived growth factor receptor
PNN	Probabilistic neural network
QSAR	Quantitative structure activity relationship
SAR	Structure-activity relationship

SBVS	Structure-based Virtual Screening
SVM	Support vector machine
TN	True negative
TP	True positive
VEGFR	Vascular endothelial growth factor receptor
VS	Virtual Screening

List of Publications

1. Virtual Screening of Selective Multi-Target Kinase Inhibitors by Combinatorial Support Vector Machines. **X.H. Ma**, R.Wang, C.Y. Tan Y.Y. Jiang, L. Tao, H.B. Rao, X.Y. Li, M.L. Go, B.C. Low and Y. Z. Chen. ***Mol Pharmaceutics.*** 7(5):1545-1560 **(2010)**.
2. Consensus model for identification of novel PI3K inhibitors in large chemical library. C.Y. Liew, **H.X. Ma**, C.W. Yap. ***J Comput Aided Mol Des.*** 24(2):131-141 **(2010)**.
3. 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.*** 27(5):2101-2110 **(2010)**.
4. Identifying Novel Type ZBGs and Non-hydroxamate HDAC Inhibitors through a SVM Based Virtual Screening Approach. X.H. Liu, H.Y. Song, J.X. Zhang, B.C. Han, X.N. Wei, **X.H. Ma**, W.K. Chui, Y.Z. Chen. ***Mol Inf.*** 29(5): 407-420 **(2010)**
5. Virtual Screening Prediction of New Potential Organocatalysts for Direct Aldol Reactions. X.H. Liu, H.Y. Song, **H.X. Ma**, M.J. Lear and Y.Z. Chen. ***J Mol Catal A: Chem***. 319:114-118 **(2010)**
6. Update of TTD: Therapeutic Target Database. F. Zhu, B.C. Han, P. Kumar, X.H. Liu, **X.H. Ma**, X.N. Wei, L. Huang, Y.F. Guo, L.Y. Han, C.J. Zheng, Y.Z. Chen. ***Nucleic Acids Res.*** 38 (Database issue):D787-791 **(2010)**
7. Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines. X.H. Liu, **X.H. Ma**, C.Y. Tan, Y.Y. Jiang,

M.L. Go, B.C. Low and Y.Z. Chen. ***J Chem Info Model.*** 49(9):2101-2110 **(2009)**

8. SVM model for virtual screening of Lck inhibitors. C.Y. Liew, **X.H. Ma**, X.H. Liu, C.W. Yap. ***J Chem Inf Model.*** 49(4):877-885 **(2009)**
9. Identification of Small Molecule Aggregators from Large Compound Libraries by Support Vector Machines. H.B. Rao, Z.R. Li, X.Y. Li, **X.H. Ma**, C.Y. Ung, H. Li, X.H. Liu and Y.Z. Chen. ***J Comput Chem.*** 31(4):752-763 **(2010)**
10. Synergistic therapeutic actions of herbal ingredients and their mechanisms from molecular interaction and network perspectives **X. H. Ma**, C.J. Zheng, L.Y. Han, B. Xie, J. Jia, Z.W. Cao, Y.X. Li and Y. Z. Chen. ***Drug Discov Today.*** 14:579-588 **(2009)**
11. Pathway sensitivity analysis for detecting pro-proliferation activities of oncogenes and tumor suppressors of EGFR-ERK pathway at altered protein levels H. Li, C. Y. Ung, **X. H. Ma**, X. H. Liu, B. W. Li, B. C. Low and Y. Z. Chen. ***Cancer.*** 15(18):4246-4263 **(2009)**
12. Simulation of Crosstalk between Small GTPase RhoA and EGFR-ERK Signaling Pathway via MEKK1. H. Li, C. Y. Ung, **X. H. Ma**, B. W. Li, B. C. Low, Z. W. Cao and Y. Z. Chen. ***Bioinformatics.***25(3):358-364 **(2009)**
13. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. **X.H. Ma**, J. Jia, F. Zhu, Y. Xue, Z.R. Li and Y.Z. Chen. ***Comb. Chem. High Throughput Screen*** 12(4):344-357 **(2009)**
14. Mechanisms of drug combinations from interaction and network perspectives J. Jia, F. Zhu, **X.H. Ma**, Z.W. Cao, Y.X. Li and Y.Z. Chen.

Nat. Rev. Drug Discov., 8(2):111-128 **(2009)**

15. Simulation of the Regulation of EGFR Endocytosis and EGFR-ERK Signaling by Endophilin-Mediated RhoA-EGFR Crosstalk. C.Y. Ung, H. Li, **X.H. Ma**, J. Jia, B.W. Li, B.C. Low and Y.Z. Chen. ***FEBS Lett***. 582:2283-2290 **(2008)**
16. Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds. **X.H. Ma**, R. Wang, S.Y. Yang, Z.R. Li, Y. Xue, Y.Q. Wei, B.C. Low and Y. Z. Chen. ***J Chem Inf Model***. 48(6):1227-1237 **(2008)**
17. 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. L.Y. Han, **X.H. Ma**, H.H. Lin, J. Jia, F. Zhu, Y. Xue, Z.R. Li, Z.W. Cao, Z.L. Ji, Y.Z. Chen. ***J. Mol. Graph. Mod.*** 26(8):1276-1286 **(2008)**
18. Prediction of Antibiotic Resistance Proteins from Sequence Derived Properties Irrespective of Sequence Similarity. H.L. Zhang, H.H. Lin, L. Tao, **X.H. Ma**, J.L. Dai, J.Jia, Z.W. Cao. ***Int J Antimicrob Agents.*** 32(3):221-226 **(2008)**
19. Advances in Machine Learning Prediction of Toxicological Properties and Adverse Drug Reactions of Pharmaceutical Agents. **X.H. Ma**, R. Wang, Y. Xue, Z.R. Li, S.Y. Yang, Y.Q. Wei and Y.Z. Chen. ***Curr Drug Saf.*** 3(2):100-114 **(2008)**
20. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. L.Y.

Han, C.J. Zheng, B. Xie, J. Jia, **X.H. Ma**, F. Zhu, H.H. Lin, X. Chen, and Y.Z. Chen. ***Drug Discov. Today*** 12(7-8): 304-313 **(2007)**

Chapter 1 Introduction

The discovery of novel chemical entities (NCEs) in the pharmaceutical industry is becoming increasingly difficult, costly and time-consuming. Many approaches have been suggested to increase the cost-effectiveness of discovery programmes, one of them being the use of virtual screening methods to complement the more traditional chemical and biological approaches. Presently, a variety of computational virtual screening tools are being developed and refined to effectively employ fast screening methods to yield potent lead hits such as docking, quantitative structure activity relationship (QSAR) and machine learning methods etc. However, virtual screening also faces several fundamental challenges. It can be regarded as less accurate, since speed and the possibility to capture most (but not necessarily all) potentially positives are its key attributes. The insufficient coverage of compound diversity, high false positive, high false negative prediction and lower speed of screening compound libraries are also required to address in the development of virtual screening method. This work on "high performance computational virtual screening tools: development and application to the discovery of kinase inhibitors" is one of such kind of strategies to improve the screening speed and the prediction accuracy and decrease the false hit rate.

The following sections will describe an overview of virtual screening in drug discovery (Section 1.1), machine learning methods in virtual screening (Section 1.2) and discuss the important role of kinase inhibitors in cancer treatment (Section 1.3) and in-silico approaches to multi-target drug

discovery (Section 1.4). In addition, the objectives and outline of this project (Section 1.5) will be introduced.

1.1 Virtual screening in drug discovery

In current drug discovery, lead compounds of high quality and structural diversity are keys to the successful development of the drug candidates. During the last 10 to 15 years, High throughput screening (HTS) of proprietary compound collections at pharmaceutical companies has represented the most important source of leads in the industry. However, the use of HTS is very expensive and companies need to purchase the synthesized compounds to be screened (if available at all). Moreover, these physically existing compounds (in-house libraries) represent only a tiny fraction of the drug-like chemical space. In more recent years, virtual screening (VS) has complemented the experimental identification of bioactive compounds. Virtual screening offers many possibilities for new structures beyond those found in in-house libraries. The term 'virtual screening' was first used in 1997, and relates to the search for compounds with a defined biological activity using computational models[1]. During the last decade, a huge number of different virtual screening methods have been reported and used to search for novel bioactive compounds for many targets. Like HTS, VS searches large libraries of potentially bioactive molecules for hits. Unlike HTS, there is no need for physically existing compounds, which is a key advantage of VS. Another advantage of VS comes from the exploration of the chemical space outside the in-house compound pool. 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[2,3], which is many orders of

magnitude larger than this, even if only drug-like compounds are considered[4]. Of the order of 10 million compounds are commercially available, which are an additional source of potential leads that can be exploited with the VS approach. Another source of accessible compounds is virtual combinatorial libraries. The chemical space accessible through virtual combinatorial libraries is at least 1 million-fold larger than that available from in-house pools and external vendor compounds, respectively, and adds a new dimension to the VS search space (**Figure 1-1**).

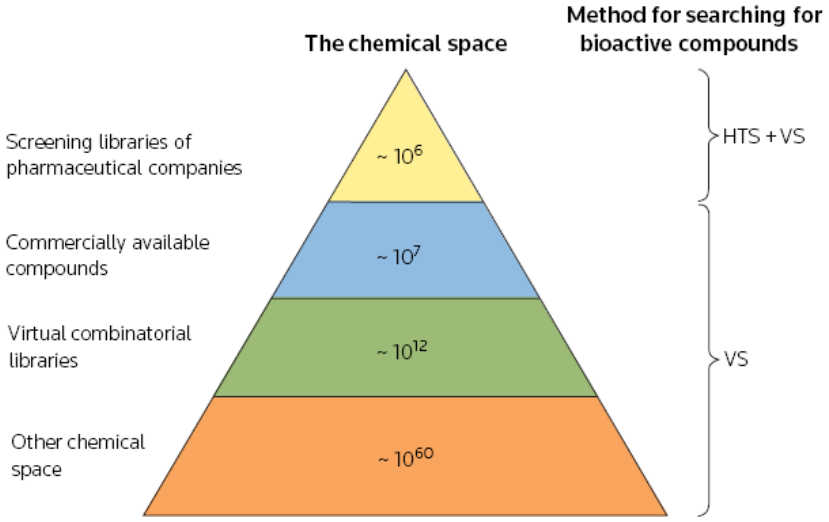


Figure 1-1 Typical numbers of compounds available in the chemical space.

Virtual screening methods are often divided into structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) depending on what is already known about a target and its ligands[5]. Structure-based virtual screening involves docking of candidate ligands into a protein target followed by applying a scoring function to estimate the likelihood that the ligand will bind to the protein with high affinity[6,7]. LBVS methods include pharmacophore methods[8] and chemical similarity analysis methods[9]. **Figure**

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

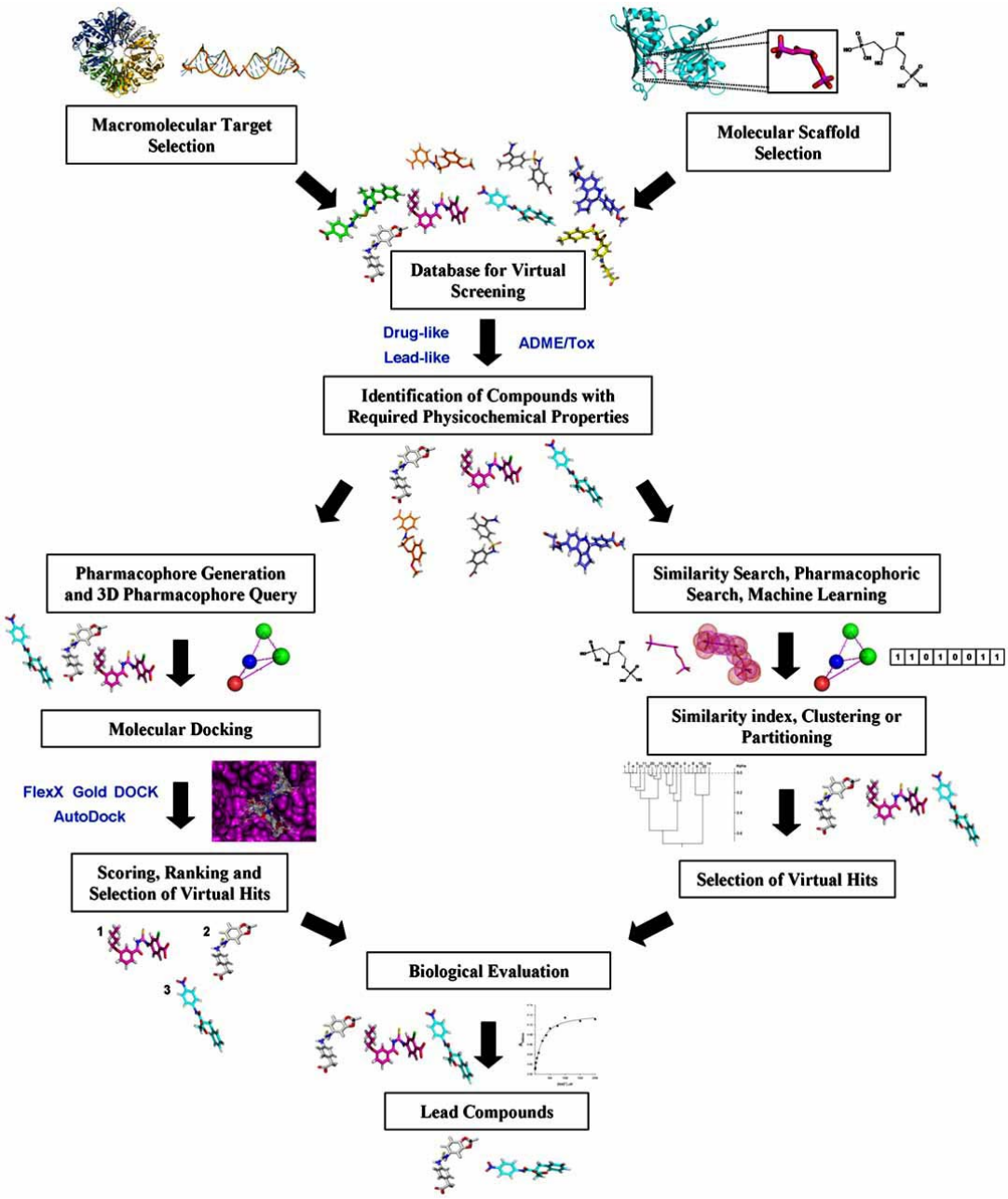


Figure 1-2 General procedure used in SBVS and LBVS (adopted from Rafael V.C. et al[10]).

1.1.1 Structure-based virtual screening

When 3D protein target structure information, derived either from experimental data (X-ray or NMR spectroscopy) or from homology modeling, is available, the most frequently used VS method is docking. Binding modes for each ligand can be predicted in silico, together with numerical assessment (score) of the interaction energy between the ligand and the protein. 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. The major challenges in scoring functions are how to account for the solvent effect and how to accurately account for entropic effect. Now desolvation and entropy contributions of both ligand and protein are included only in an approximate way. To date, more than 60 docking programs and 30 scoring functions have been reported. Both docking programs and scoring functions have been evaluated and reviewed extensively[11,12]. Most researchers agree that there is currently no 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 assess in advance, and depends on the nature of the target[11-13]. Despite all optimization efforts, the currently available scoring functions do not provide reliable estimates of free binding energies, and are not able to rank-order compounds according to affinity[12,14]. The published comparisons of docking programs have been critically reviewed[15-17].

1.1.2 Ligand-based virtual screening

Ligand-based VS begins with the use of one or more active compounds as templates, and no details about the target are needed. In general, LBVS methods depend on the application of computational descriptors of molecular structure, properties, or pharmacophore features and analyze relationships between active and database or test compounds in however defined chemical descriptor spaces. It is computationally efficient and can rapidly search very large databases. As a result, it is often used to sequentially filter large compound sets before more complex tools are applied. Myriads of different methods have been reported, and there are literally thousands of different descriptors, which are derived from the 2D or 3D distribution of atomic properties in compounds, or from the presence or absence of specific structural elements. Many methods exist for the comparison of the similarity of compounds based on these descriptors. In ligand-based VS, shape comparison is frequently used[18], and pharmacophore searches are also a long-established technique[8,19]. Other methods use molecular fields to define the similarity of structures[20,21]. If 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 distinguish active from inactive compounds based on their specific structural features. For a comprehensive overview of ligand-based VS the reader is referred to a number of reviews[22,23].

1.2 Machine learning in virtual screening

Machine learning methods have been explored as an alternative virtual screening method. It utilize nonlinear supervised learning methods to develop statistical models that map physicochemical properties (molecular descriptors) with their activity classes, so they are more capable of predicting a more diverse spectrum of compounds and more complex structure-activity relationships than structure-based virtual screening methods and other ligand-based virtual screening methods such as QSAR, pharmacophore, and clustering methods[24-31]. This capability arises because machine learning methods are capable of generating complex nonlinear mappings from molecular descriptors to activity classes without restriction on structural frameworks, and without requiring prior knowledge of relevant molecular descriptors and functional form of structure-activity relationships[32-36]. Moreover, machine learning methods can overcome several problems that have impeded progress in the application of structure-based virtual screening and other ligand-based virtual screening tools[33,37]. These problems include the vastness and sparse nature of searched chemical space, limited availability of target structures (only 15% of known proteins have known 3D structures); limited diversity biased by training molecules, complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects.

The reported performance of machine learning methods in screening pharmacodynamically active compounds from libraries of >25,000 compounds is summarized in **Table 1-1**. The screening tasks of these reported studies[38-45]

are primarily focused on the prediction of compounds that inhibit, antagonize, block, agonize, or activate specific therapeutic target protein. Machine learning methods have been found to show generally good performances. In the majority of the reported studies, the yields, hit rates, and enrichment factors of machine learning methods are in the range of 50%~94%, 10%~98%, and 30~108 respectively.

For tentative comparison of the performance of machine learning methods with other virtual screening methods, the reported performances of structure-based VS methods and two classes of ligand-based VS methods, pharmacophore and clustering, are summarized in Table 2, Table 3 and Table 4 respectively. The yields, hit rates, and enrichment factors of the majority of the reported studies by other methods shown in **Table 1-2**, **Table 1-3** and **Table 1-4** are 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 VS studies by using structure-based, pharmacophore and clustering methods. However, we can see from the **Table 1-2**, **Table 1-3**, **Table 1-4**, 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 In screening libraries of ~98,000 compounds, the reported hit-rates of some machine learning VS tools are comparable to those

of structure-based VS tools, 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.

Two approaches have been explored to improve hit-rates and enrichment factors. One is the selection of top-ranked hits, which has been extensively used in ligand-based[46-51] and structure-based [52-57] VS tools. The other is the elimination of unlikely hits at the pre-screening stage by using such filters as Lipinski's rule of five[58] for drug-like compounds, identification of specific chemical groups or interaction patterns[52,53,59,60], and pharmacophore recognition[54]. These two methods are effective to improve hit-rates and enrichment factors but they are just supplemental methods combined with virtual screening methods. Higher performance virtual screening methods are required. The performance of machine learning VS tools in screening large libraries can be further improved by using training sets of more diverse spectrum of compounds to develop more optimally performing machine learning VS tools. These tools have been generated by using two-tier supervised classification machine learning methods[36,46-49,61-63], which require training sets of diverse spectrum of active and inactive compounds.

Machine learning methods have shown promising capability in virtual screening of compounds of diverse ranges of structures for identifying

compounds of a wide variety of pharmacodynamic and other properties. In virtual screening of large libraries, these methods have been found to be capable of achieving comparable performance to other structure-based and ligand-based VS methods. By using training sets of more diverse spectrum of inactive compounds, the hit-rates and enrichment factors of machine learning VS tools can be substantially improved to the level comparable to and in some cases higher than those of the best performing structure-based and ligand-based VS tools.

Table 1-1 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 [49]	Molecular fingerprints	94/200K	2,500	0.1%	18	81%	0.7%	795
	25,300	25	SVM+ BKD [46]	DRAGON descriptors	125/5035	506	2%	20	80%	3.9%	39.5
COX inhibitors	102,514	536	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	76	14.3%	1.4%	2.7
	98,435	536	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	240	44.7%	24.4%	44.5
Thrombin inhibitors	2.5M	46	SVM [49]	Molecular fingerprints	188/200K	11,250	0.45%	25	55%	0.2%	108.7
	102,514	703	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	367	52,3%	7.1%	10.3
	98,435	703	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	381	54.2%	38.9%	54.0
Protease inhibitors	171,726	118	SVM [47]	Extended connectivity fingerprints	228/4200	1717	1%	26	22%	1.5%	21.8
			LMNB [47]					19	16%	1%	14.5
Chemokine receptor antagonists	171,560	128	SVM [47]	Extended connectivity fingerprints	258/4199	1716	1%	70	55%	4.1%	54.9
			LMNB [48,62]					68	53%	3.9%	52.3

5HT3 antagonists	102,514	652	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	236	36.3%	4.6%	7.2
	98,435	852	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	529	62.1%	54.0%	62.1
5HT1A antagonists	102,514	727	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	224	30.9%	4.3%	6.1
	98,435	727	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	319	43.9%	32.6%	44.0
5HT reuptake inhibitors	102,514	259	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	65	25%	1.2%	4.7
	98,435	259	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	137	52.9%	14.0%	53.8
D2 antagonists	102,514	295	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	90	30.6%	1.7%	5.9
	98,435	295	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	137	46.4%	14.0%	53.8
Rennin inhibitors	102,514	1030	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	972	94.4%	18.9%	18.9
	98,435	1030	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	710	68.9%	72.4%	69.0
Angiotesin II AT1 antagonists	102,514	843	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	776	92.1%	15.1%	18.4

	98,435	843	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	384	45.6%	39.2%	45.6
Substance P antagonists	102,514	1146	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	378	33%	7.3%	6.5
	98,435	1146	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	509	44.4%	51.9%	44.4
HIV protease inhibitors	102,514	650	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	377	58%	7.3%	11.5
	98,435	650	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	355	54.6%	36.2%	54.9
Protein kinase C inhibitors	102,514	353	BKD [48,51]	Extended connectivity fingerprints	100/400	5125	5%	81	23.1%	1.5%	4.4
	98,435	353	CKD [36]	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 [36]	Pipeline pilot	100/4000	984	1%	206	58.3%	21.0%	58.3
MAO inhibitors	101,437	1166	BKD [61]	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 [36]	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 [36]	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 [36]	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	98,435	782	CKD [36]	Pipeline pilot	100/4000	984	1%	436	55.8%	44.3%	56.1

reductase inhibitors				ECFP4	100/4000	984	1%	665	85.0%	67.6%	85.5
Reverse transcriptase inhibitors	98,435	419	CKD [36]	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 [36]	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 [36]	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 [46]	DRAGON descriptors	125/5035	506	2%	18	72%	3.5%	35.4
FXa inhibitors	25,300	25	SVM+ BKD [46]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	N/A
PDE5 inhibitors	50,000	19	RO5+ DS [64]	Pharmacophore and macroscopic descriptors	130/10K	1821	3.6%	11	57.8%	0.6%	15.8
	25,300	25	SVM+ BKD [46]	DRAGON descriptors	125/5035	506	2%	21	84%	4.1%	41.5
Alpha1A AR antagonists	25,300	25	SVM+ BKD [46]	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 1-2 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 [65]	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 [52]	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 [59]	<400K	N/A	35	0.0087%	4	N/A	11.4%	N/A
Thyroid hormone receptor antagonists	250K	>14	ICM VLS module (Molsoft) [66] + 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 [67]	165,581	Top-500 + Top-500	889	0.38%	127	N/A	14.3%	N/A
	141K	10	GOLD + elements and chemical group screen [53]	<141K	Top-2%	<2820	<2.5%	8	80%	<0.28%	39.4
BCL-2 inhibitors	206,876	>1	DOCK3.5 + non-peptidic	<206,876	Top-500	35	0.017%	1	N/A	2.9%	N/A

			screen [68]								
HIV-1 protease inhibitors	141K	5	GLIDE + elements and chemical group screen [53]	<141K	Top-5%	<7050	<5%	1	20%	<0.014%	4.6
HDM2 inhibitors	141K	14	DOCK + elements and chemical group screen [53]	<141K	Top-5%	<7050	<5%	4	28.6%	<0.056%	5.7
UPA inhibitors	141K	10	GOLD + elements and chemical group screen [53]	<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 [54]	22,950	Top-300	300	0.21%	38	N/A	N/A	N/A
Thrombin inhibitors	141K	10	GLIDE + elements and chemical group screen [53]	<141K	Top-2%	<2820	<2.5%	3	30%	<0.11%	15.5
	133.8K	760	FlexX + Similarity [69]	<133.8K	Top-1%	1338	1%	231	29.3%	17.3%	30.5
DHFR inhibitors	135K	165	DOCK3.5.54 applied to holo form [55]	135K	Top-1% of 50k docked	1350	1%	47	25%	3.4%	27.8
			DOCK3.5.54 applied to appo form [55]	135K	Top-1% of 100k docked	1000	1%	16	9.7%	1.6%	13.1
Neutral endopeptidase inhibitors	135K	356	DOCK3.5.54 [55]	135K	Top-1% of 125.5K docked	1255	0.74%	3	0.8%	0.24%	~1
Thrombin inhibitors	135K	788	DOCK3.5.54 [55]	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 [55]	135K	Top-1% of 54K docked	540	0.4%	49	26.5%	9.1%	66.4
Phospholipase C inhibitors	135K	25	DOCK3.5.54 [55]	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 [55]	135K	Top-5% of database	4500	3.3%	10	2.8%	0.22%	~1
			DOCK3.5.54 applied to appo form [55]	135K	Top-5% of database	4500	3.3%	5	1.4%	0.11%	<1
	133.8K	59	FlexX + Similarity [69]	<133.8K	Top-1%	1338	1%	13	22%	0.97%	22.0
Acetylcholinesterase inhibitors	135K	637	DOCK3.5.54 applied to holo form [55]	135K	Top-1% of 77K docked	770	0.57%	49	7.7%	6.4%	13.6
			DOCK3.5.54 applied to appo form [55]	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 [69]	<133.8K	Top-1%	1338	1%	35	3.4%	2.6%	3.4

Table 1-3 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 [70]	1M	26%	39	70.1%	0.0039%	2.8
	3.8M	55	Structure-based pharmacophore [71]	91K	2.4%	6	10.9%	0.0066%	4.6
11β-hydroxysteroid dehydrogenase 1 inhibitors	1.77M	144	Pharmacophore [30]	20.3K	1.15%	17	11.8%	0.084%	10.3
Rhinovirus 3C protease inhibitors	380K	30	Pharmacophore [31]	6,917	1.82%	23	76.7%	0.33%	41.8

Table 1-4 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 [29]	5590	1.6%	246	50.2%	4.4%	31.2
			NIPALSTREE [29]	8174	2.4%	188	38.4%	2.3%	16.2
			Hierachical k-means + NIPALSTREE disjunction [29]	12240	3.6%	306	62.4%	2.5%	17.6
			Hierachical k-means + NIPALSTREE conjunction [29]	1662	0.48%	128	26.1%	7.7%	54
COX inhibitors	344.5K	1556	Hierachical k-means [29]	15322	4.4%	761	48.9%	5.0%	11
			NIPALSTREE [29]	22321	6.5%	625	40.2%	2.8%	6.16
			Hierachical k-means + NIPALSTREE disjunction [29]	33793	9.8%	980	63.0%	2.9%	6.42
			Hierachical k-means + NIPALSTREE conjunction [29]	3980	1.2%	406	26.1%	10.2%	22.6
Adrenoceptor ligand	344.5K	542	Hierachical k-means [29]	21285	6.2%	298	55.0%	1.4%	8.99
			NIPALSTREE [29]	28125	8.2%	270	49.8%	0.96%	6.14
			Hierachical k-means + NIPALSTREE disjunction [29]	42365	12.3%	394	72.7%	0.93%	5.93
			Hierachical k-means + NIPALSTREE conjunction [29]	6692	1.9%	174	32.1%	2.6%	16..3

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

1.3 Protein kinase inhibitors in cancer treatment

There are some 518 protein kinases that share a catalytic domain highly conserved in sequence and structure in the human genome. The kinase family is one of the largest target families and its key function in signal transduction for all organisms makes it a very attractive target class for therapeutic interventions in many disease states such as cancer, diabetes, inflammation, and arthritis. Protein kinases play important roles in regulating most cellular functions such as proliferation/cell cycle, cell metabolism, survival/apoptosis, DNA damage repair, cell motility, response to the microenvironment, so they are often themselves oncogenes. Kinases such as c-Src, c-Abl, mitogen activated protein (MAP) kinase, phosphotidylinositol-3-kinase (PI3K) AKT, and the epidermal growth factor (EGF) receptor are commonly activated in cancer cells, and are known to contribute to tumorigenesis[72,73]. Small molecule kinase inhibitors have been designed to inhibit the enzyme's adenosine triphosphate (ATP) binding site for cancer treatment[74]. There are currently over 70 reported small molecule kinase inhibitors at various stages of clinical trials in oncology (www.clinicaltrials.gov) which emphasises the potential importance in targeting protein kinases for treating human malignancies. The advent of kinase targeted therapy for the treatment of human cancer offers a potential therapy to improve both patient survival and quality of life during treatment[75].

Kinase inhibitors designed to bind the catalytic ATP-binding site can have broad specificity because of kinases' high conserved sequence and structure. Imatinib (Gleevec, Novartis) is a highly successful cancer drug due to its

activity as an inhibitor of the Abelson cytoplasmic tyrosine kinase (Abl), which is constitutively active in a majority of patients with chronic myelogenous leukemia (FDA-approved in May 2001). Imatinib also inhibit c-Kit and the platelet-derived growth factor (PDGF) receptor tyrosine kinases. So it can be used to treat gastrointestinal stromal and other types of tumors associated with activation of these signaling molecules. Cancer cells use multiple pathways to promote their own survival and proliferation, combination therapies (of multiple targeted therapeutics, or of targeted drugs plus chemotherapy) are likely to be required to completely eradicate a tumor and prevent resistance or relapse. Due to kinases' broad specificity, it is possible to design multi-target kinase inhibitors for achieving enhanced therapeutic efficacies through controlling multiple pathways in cancer network. However, just because of this broad specificity, many kinase inhibitors have "off-target" effect in modulating signaling pathway. It is also necessary to design more specific kinase inhibitors for cancer treatment.

1.4 *In-Silico* approaches to multi-target drug discovery

Therapeutic agents directed at an individual target frequently show reduced efficacies, undesired safety profiles and drug resistances due to network robustness[76], redundancy[77], crosstalk[78], compensatory and neutralizing actions[79], anti-target and counter-target activities[80], and on-target and off-target toxicities[81]. Multi-target agents directed at selected multiple targets have been increasingly explored[76,82] for achieving enhanced therapeutic efficacies, improved safety profiles, and reduced resistance activities by simultaneously modulating the activity of a primary therapeutic target and the counteractive elements and resistance activities[83] while limiting un-wanted cross-reactivities

via optimization of target selectivity[84]. Examples of clinically successful multi-target drugs are anticancer kinase inhibitors sunitinib against PDGFR and VEGFR, dasatinib against Abl and Src, and lapatinib against EGFR and HER2[85,86]. These multi-target anticancer agents inhibit a primary therapeutic target that promote tumor growth in specific cancer patient group and block the alternative signalling or escape mechanism[79,87,88].

In-silico methods have been widely explored for facilitating lead discovery against individual targets[37,89,90]. In particular, molecular docking[91], pharmacophore[92], structure-activity relationship (SAR) and quantitative structure activity relationship (QSAR)[93], machine learning[94], and combination methods[95] have been extensively used for searching and designing active compounds against individual targets. Some of these methods have recently been explored for searching and designing multi-target agents. **Figure 1-3**, **Figure 1-4**, **Figure 1-5**, and **Figure 1-6** outline the strategies 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 are classified into combinatorial approaches and fragment–based approaches. Combinatorial approaches (**Figure 1-3 and Figure 1-4**) straightforwardly conduct parallel search against each individual targets to find virtual hits that simultaneously interact with multiple targets. Combinatorial approaches are practically useful if the retrieval rates against individual targets are sufficiently high and the false-hit rates are sufficiently low. High retrieval rates 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 an individual target.

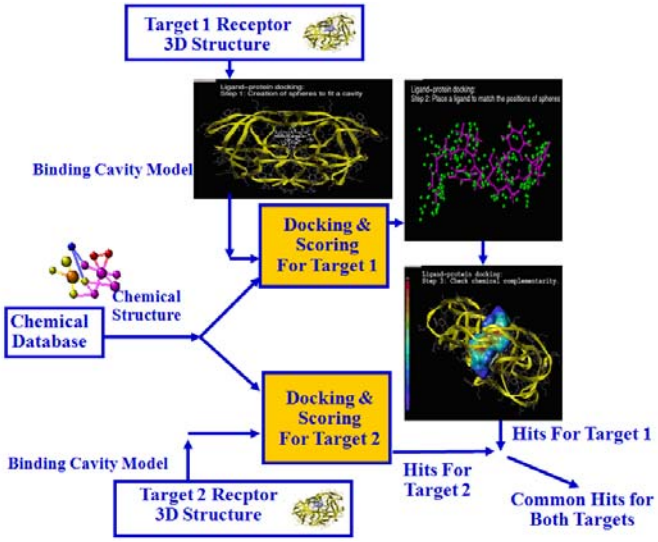


Figure 1-3 Molecular docking strategy for multi-target inhibitor discovery.

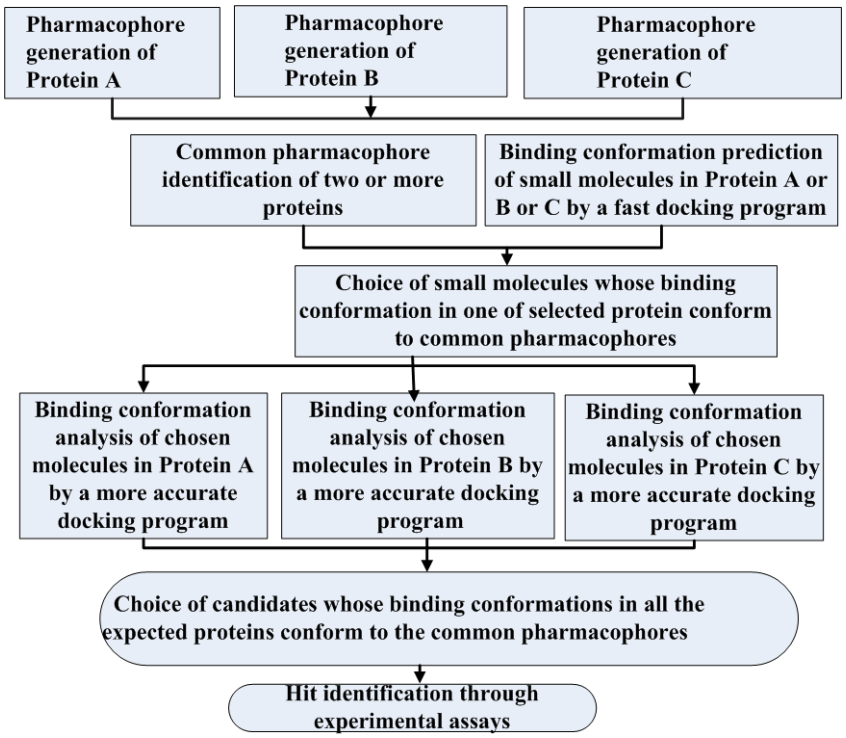


Figure 1-4 Combined pharmacophore and molecular docking strategy of multi-target inhibitor discovery.

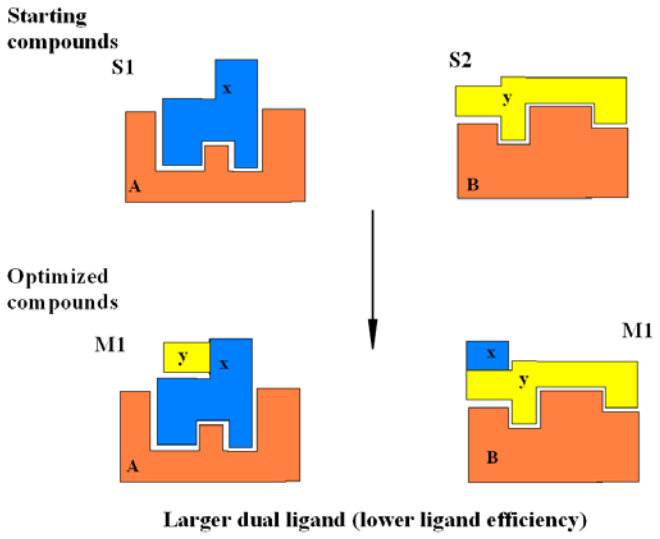


Figure 1-5 Illustration of framework combination approach to multi-target drug discovery.

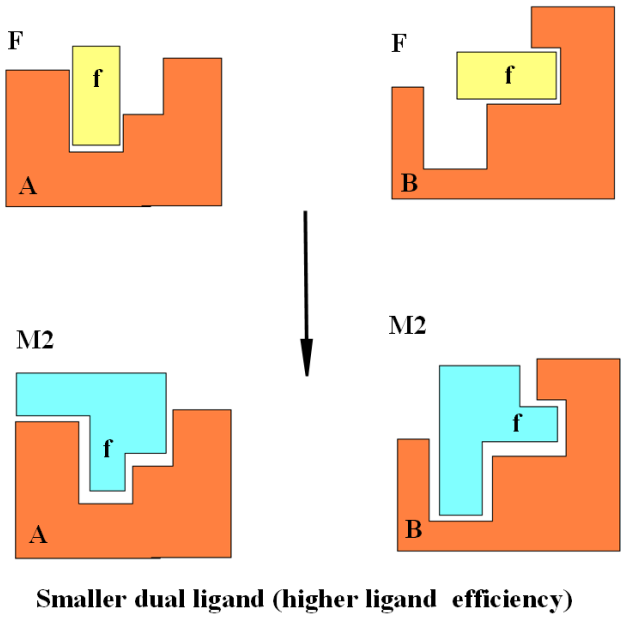


Figure 1-6 Illustration of fragment-based approach to multi-target drug discovery.

Fragment-based approaches (**Figure 1-5 and Figure 1-6**) combine multiple elements of structural frameworks or multiple fragments that bind to each individual target to design compounds that bind to multiple targets, which have been introduced as tools for the design of multi-target agents[96]. In one approach, the structure-activity relationships against individual targets are analyzed to find molecular fragments and essential binding features which are either combined or incorporated into active agents against selected multiple targets[96]. Fragment combination often results in larger and more complex non-drug like molecules. Drug-like features may be retained if the degree of framework overlap is maximized and the size of the selected fragments is minimized. In another 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[96]. Optimizing fragments with weak multiple activities into potent multi-target drug-like agents can be more easily achieved for targets sharing a conserved binding site[97]. As binding sites become more dissimilar, it is increasingly difficult to improve and adequately balance the high binding affinities needed to achieve acceptable *in-vivo* efficacy and safety. One way to reduce this difficulty is to explore synergistic targets, such that multi-target agents with modest activity at one or more of the relevant targets may still produce similar or better *in-vivo* effects compared with higher-affinity target-selective compounds[98].

Moreover, multi-target QSAR models for identification of multi-target agents[99] and active agents against multiple bacterial[100], fungal[101,102] and viral[100] species have been developed by incorporating 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. These multi-target QSAR models achieve high retrieval rates of 72%~85% and moderately low false-hit rates of 15%~28%. Development 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 needs to be in a certain range for accurate computation of multi-target dependent or species-dependent molecular descriptors, which in some cases may also affect one's capability for developing multi-target QSAR models[102].

Multi-target based in-silico methods have been increasingly explored and have shown promising potential as virtual screening tools for identifying selective multi-target agents. The capability of these methods may be further enhanced by incorporating knowledge of newly discovered selective multi-target agents from the current and future drug discovery efforts[85,86], and by the improvement of virtual screening methods[103-109].

1.5 Objectives and outline of this work

Overall, there are four major objectives for this work.

1. To construct high performance virtual screening tools for searching potential inhibitors or antagonists through screening large chemical libraries.
2. To evaluate the robustness of our virtual screening tools. In this work, sparsely distributed active compounds are used to achieve this objective.
3. To search potential c-Src and VEGFR-2 selective kinase inhibitors applying the developed virtual screening tools.
4. To build combinatorial support vector machines (C-SVMs) models applying the developed virtual screening tools to search dual inhibitors of kinase pairs.

In summary, this work is aimed at the development, evaluation and application of high performance virtual screening tools. More specifically, the study seeks to search potential single kinase inhibitors (c-Src and VEGFR-2) and multi-target kinase inhibitors through screening large compound libraries. The present study may shed some light on the capability of machine learning based virtual screening methods in searching potential active agents from large

compound libraries at low false-hit rates, which could help in the lead discovery and optimization.

The complete outline of this thesis is as follows: In Chapter 1, an introduction to virtual screening in drug discovery is described. Machine learning method is compared with other virtual screening method according to the literature review. In addition, the importance of potential kinase inhibitors discovery and in silico approaches of multi-target kinase discovery are presented.

In Chapter 2, methods used in this work are described. In particular, the dataset quality analysis, the molecular descriptors, various statistical learning methods used in this work, and the model evaluation methods are presented in more detail.

Chapter 3 is devoted to the development of high performance virtual screening tools. In particular, putative negative dataset is involved in training dataset to build SVM model to improve the performance of virtual screening. The performance of this virtual screening platform is evaluated using four datasets: HIV-1 protease inhibitors, DHFR inhibitors, Dopamine antagonists and CNS active agents. The results of screening 2.98M PubChem database using this platform show that the hit-rates are comparable and the enrichment factors are substantially better than the best results of other virtual screening (VS) tools.

Chapter 4 is devoted to the evaluation of the virtual screening tools developed in Chapter 3 by using sparsely distributed active compounds. SVM models are

trained by regularly sparse datasets of 100 actives and very sparse datasets of 40 datasets from six MDDR biological target classes. These models' performance of virtual screening PubChem and MDDR database show that the platform developed in Chapter 3 has substantial capability in identifying novel active compounds from sparse active datasets at low false-hit rates.

In Chapter 5, virtual screening models of kinase c-Src and VEGFR-2 inhibitors are built using the method discussed in Chapter 3 to screen large compound libraries. Independent dataset and MDDR screening results show that rational c-Src and VEGFR-2 hits are given by our virtual screening tools.

In Chapter 6, combinatorial support vector machines (C-SVMs) models were provided as virtual screening tools for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). In particular, C-SVMs method was compared to other VS methods DOCK Blaster, kNN and PNN against the same sets of kinase inhibitors and 1.02M Zinc clean-leads dataset.

Finally, in the last chapter, Chapter 7, major findings and contributions of current work for the development and application of high performance virtual screening tools were discussed. Limitations and suggestions for future studies were also rationalized in this chapter.

Chapter 2 Methods

Machine learning based virtual screening for drug leads discovery will normally consist of three main components: (1) pharmaceutical agent datasets and chemical compound libraries (section 2.1), (2) physicochemical and structural descriptions of the compounds in the dataset (section 2.2) and (3) a statistical learning technique used to correlate the first two components (section 2.3). In this chapter, these three components are described and all the methods used in this work for developing virtual screening model are featured. Methods that are used for checking the validity and usefulness of virtual screening models are also described (section 2.4).

2.1 Datasets

2.1.1 Data source

Data accessibility is critical for the success of a drug discovery and development. Huge amounts of small molecules and their related information have been accumulated in scientific literatures and databases. Some important small molecule databases are given in **Table 2-1**.

In this work, datasets are mainly collected from the journals (Bioorganic & Medicinal Chemistry Letters, Bioorganic & Medicinal Chemistry, European Journal of Medicinal Chemistry, European Journal of Organic Chemistry and Journal of Medicinal Chemistry,etc) and databases (BindingDB[110], MDDR, PubChem and ZINC[111], etc).

Table 2-1 Some 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

2.1.2 Data quality analysis

The development of reliable pharmacological properties classification models depends on the availability of high quality pharmacological property descriptor data with low experimental errors[112]. Ideally, these pharmacological properties descriptors should be measured by a single protocol so that different compounds can be reliably compared with each other. However, some pharmacological properties descriptors have been measured only for a limited number of compounds and these data are rarely determined by the same protocol. Thus data selection has been primarily based on comparison of data of compounds commonly studied by different protocols, and incorporation of additional experimental information. For this work, several methods are adopted to ensure that inter-laboratory variations in experimental protocols do not significantly affect the quality of the training sets. The sources for the

pharmacological property descriptor data for each compound were investigated to ensure that there were no wide variations in experimental protocols from those of the majority of the compounds in the training set. Compounds that were investigated in more than one source are used to estimate the quality of each source. It is assumed that the most common range of the pharmacological properties descriptor data for the compounds investigated in more than one source was used to select compounds for the different classes[113].

2.1.3 Determination of structural diversity

Structural diversity of a collection of compounds can be evaluated by using the Diversity Index (DI), which is the average value of the similarity between pairs of compounds in a dataset[114],

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

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 coeeficient[115] were 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 descriptors calculated for the compounds in the dataset.

2.2 Molecular descriptors

2.2.1 Types of molecular descriptors

Molecular descriptors have been extensively used in deriving structure-activity relationships[116,117], quantitative structure activity relationships[118,119], and machine learning prediction models for pharmaceutical agents[120-123]. A descriptor is "the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a compound into a useful number or the result of some standardized experiment". A number of programs e.g. DRAGON[124], Molconn-Z[125], MODEL[126], Chemistry Development Kit (CDK)[127,128], JOELib[129] and Xue descriptor set[130], are available to calculate chemical descriptors. These methods can be used for deriving >3,000 molecular descriptors including constitutional descriptors, topological descriptors, RDF descriptors[131], molecular walk counts[132], 3D-MoRSE descriptors[133], BCUT descriptors[134], WHIM descriptors[135], Galvez topological charge indices and charge descriptors[136], GETAWAY descriptors[137], 2D autocorrelations, functional groups, atom-centred descriptors, aromaticity indices[138], Randic molecular profiles[139], electrotopological state descriptors[140], linear solvation energy relationship descriptors[141], and other empirical and molecular properties. Not all of the available descriptors are needed for representing features of a particular class of compounds. Moreover, without properly selecting the appropriate set of descriptors, the performance of a developed ML VS tool may be affected to some degrees because of 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 used in this work.

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 Intery 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 Estate 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 Estate of all heavy atoms, all C atoms, all hetero atoms, Sum of Estate of H-bond acceptors, Sum of H Estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin, Sum of H Estate of H-bond donors

In our work, descriptors were computed from the 3D structure of the compounds. The 2D structure of each of the compounds was generated by using ChemDraw or downloaded from other database like PubChem, BindingDB[110] database and was subsequently converted into 3D structure by using CORINA[142]. All the generated geometries had been fully optimized without symmetry restrictions. The 3D structure of each compound was manually inspected to ensure that the chirality of each chiral agent was properly generated. All salts and elements, such as sodium or calcium, were removed prior to descriptor calculation.

2.2.2 Scaling of molecular descriptors

Chemical descriptors are normally scaled before they can be employed for machine learning. Scaling of chemical descriptors ensures that each descriptor has an unbiased contribution in creating the prediction models[143]. Scaling can be done by number of ways e.g. auto-scaling, range scaling, Pareto scaling[144], and feature weighting[143]. In this work, range scaling is used to scale the chemical descriptor data. Range scaling is done by dividing the difference between the descriptor value and the minimum value of that descriptor with the in range of that descriptor:

$$d_{ij}^{scaled} = \frac{d_{ij} - d_{j,min}}{d_{j,max} - d_{j,min}} \quad (3)$$

Where d_{ij}^{scaled}, d_{ij} , $d_{j,\max}$ and $d_{j,\min}$ are the scale descriptor value of compound i, absolute descriptor value of compound i , maximum and minimum values of descriptor j respectively. The scaled descriptor value falls in the range of 0 and 1.

2.3 Machine learning classification methods

Machine learning classification methods employ computational and statistical methods to construct mathematical models from training samples which is used to classify independent test sample. The training samples are represented by vectors which can binary, categorical or continuous. Machine learning can be divided into two types: Supervised and Unsupervised. Supervised machine learning, as the name indicates, generally needs feeding which generally involves already labeled or classified training data. 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, as the name indicates, gets unlabeled training data and the learning task involves finding the organization of data. Examples of unsupervised machine learning include Clustering, Adaptive Resonance Theory, and Self Organized Map (SOM). Some of machine learning methods employed in this work are SVM, Probabilistic Neural Network (PNN), k nearest neighbor (KNN). They are explained below in subsequent sub sections. For a comparative study, Tanimoto similarity searching method is also introduced.

2.3.1 Support vector machines method

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**. SVM is based on the structural risk minimization principle of statistical learning theory[145,146], which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for over-fitting[147,148]. 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:

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \qquad \text{Class 1 (active)} \tag{4}$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \qquad \text{Class 2 (inactive)} \tag{5}$$

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**. Base on **w** and *b*, a given vector ***x*** can be classified by $f(\boldsymbol{x}) = sign[(\mathbf{w}\cdot\mathbf{x}) + b]$. A positive or negative *f(**x**)* value indicates that the vector **x** belongs to the active or inactive class respectively.

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures[46-48,63,149-151], SVM maps the input vectors into a higher dimensional feature space by using a kernel function K($\mathbf{x}_i$, $\mathbf{x}_j$). We used RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2 / 2\sigma^2}$ which has been extensively used and

consistently shown better performance than other kernel functions[24,152,153]. Linear SVM can then 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 Langrangian 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*** is an inhibitor or non-inhibitor respectively.

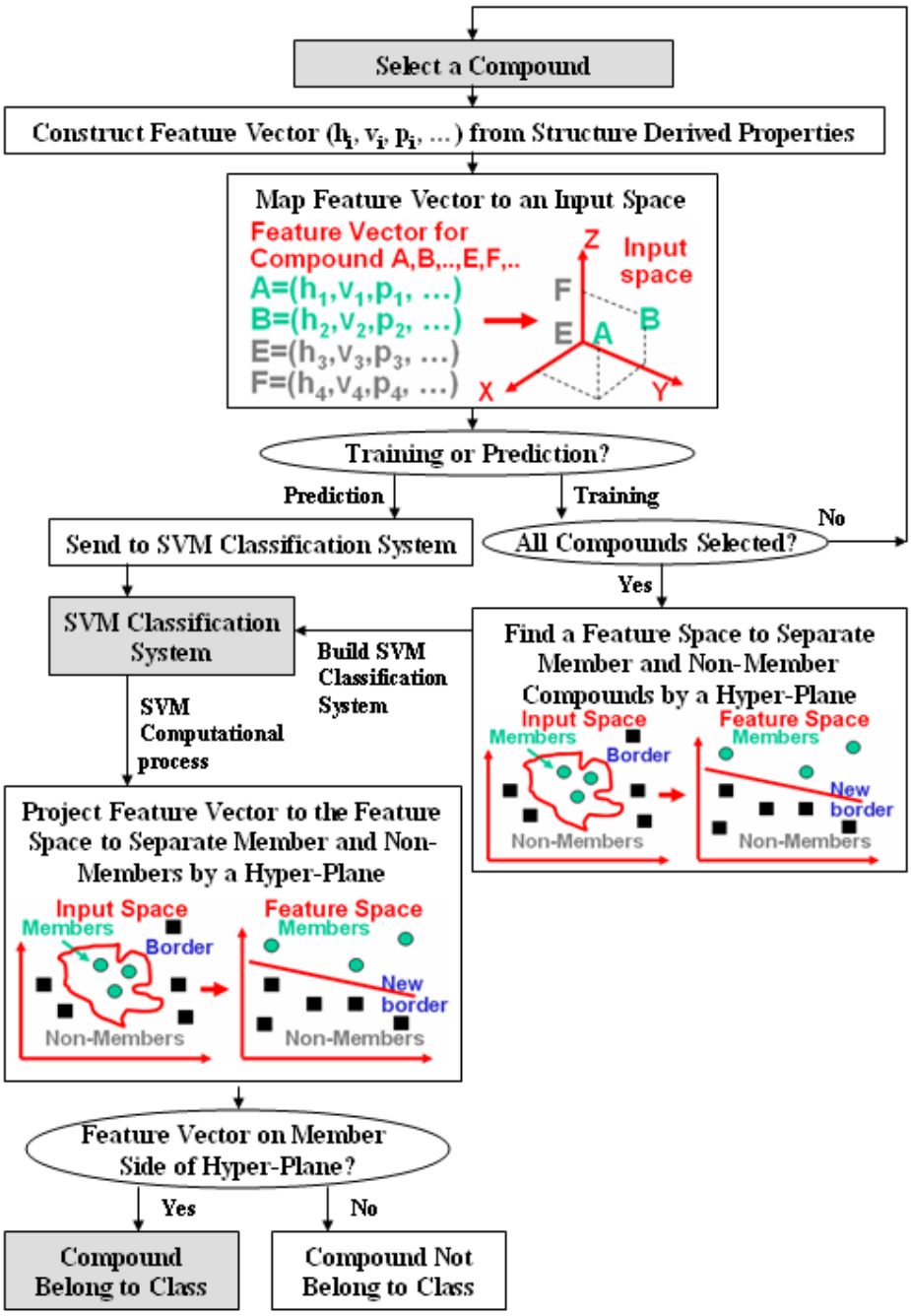


Figure 2-1 Schematic diagram illustrating the process of the training a prediction model and using it for predicting active compounds of a compound class from their structurally-derived properties (molecular descriptors) by using support vector machines. A, B, E, F and (h_j, p_j, vj,…) represents such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

2.3.2 K-nearest neighbor method

k-NN is illustrated in **Figure 2-2**. k-NN measures the Euclidean distance $D = \sqrt{\|\mathbf{x} - \mathbf{x}_i\|^2}$ between a compound ***x*** and each individual inhibitor or non-inhibitor $\boldsymbol{x}_i$ in the training set[154,155]. A total of *k* number of vectors nearest to the vector ***x*** are used to determine the decision function ***f(x)***:

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

where $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ if $a \neq b$, argmax is the maximum of the function, *V* is a finite set of vectors {v1,...,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 neighbours.

2.3.3 Probabilistic neural network method

As illustrated in **Figure 2-3**, PNN is a form of neural network designed for classification through the use of Bayes' optimal decision rule[113]

$$h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$$

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 unknown vector ***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[156],

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

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.

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

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

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

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are 4 layers in a PNN. 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. The single neuron in the output layer then estimates the class of the unknown vector $\boldsymbol{x}$ by comparing all the probability density function from the summation neurons and choosing the class with the highest probability density function.

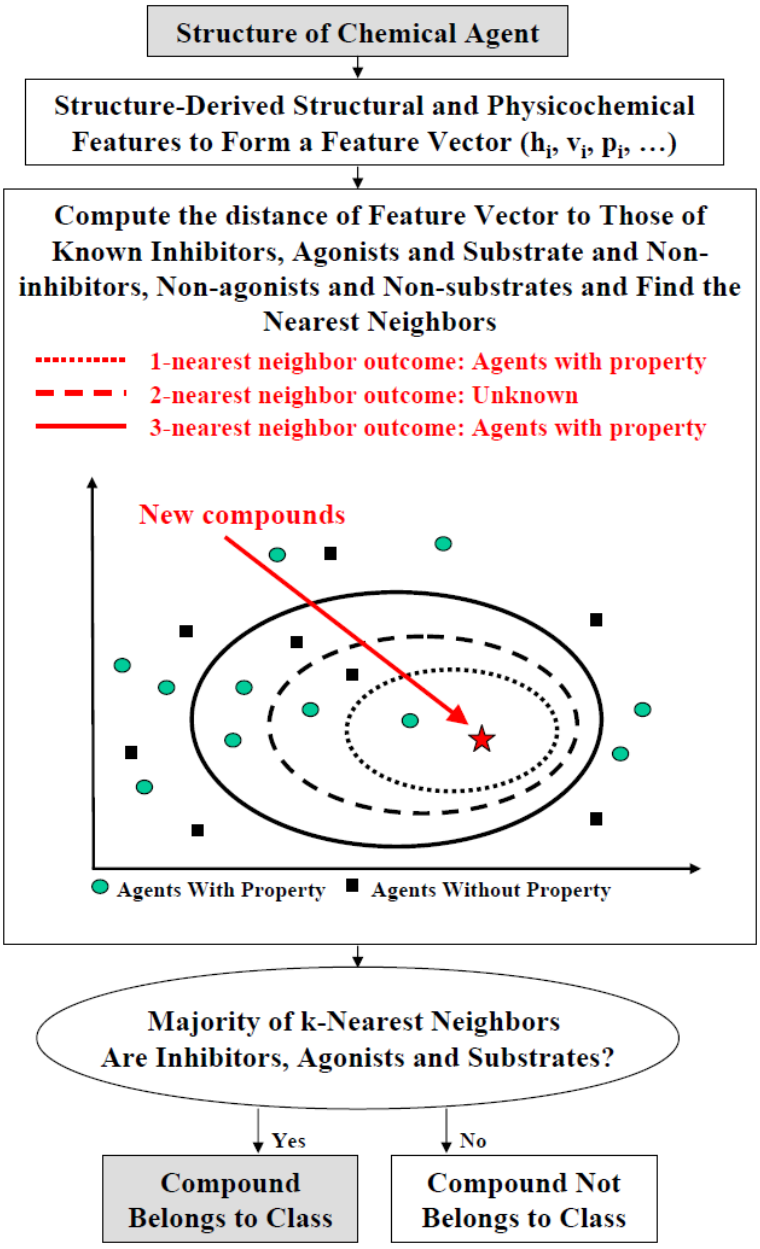


Figure 2-2 Schematic diagram illustrating the process of the prediction of compounds of a particular property from their structure by using a machine learning method – k-nearest neighbors (k-NN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents without the property;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

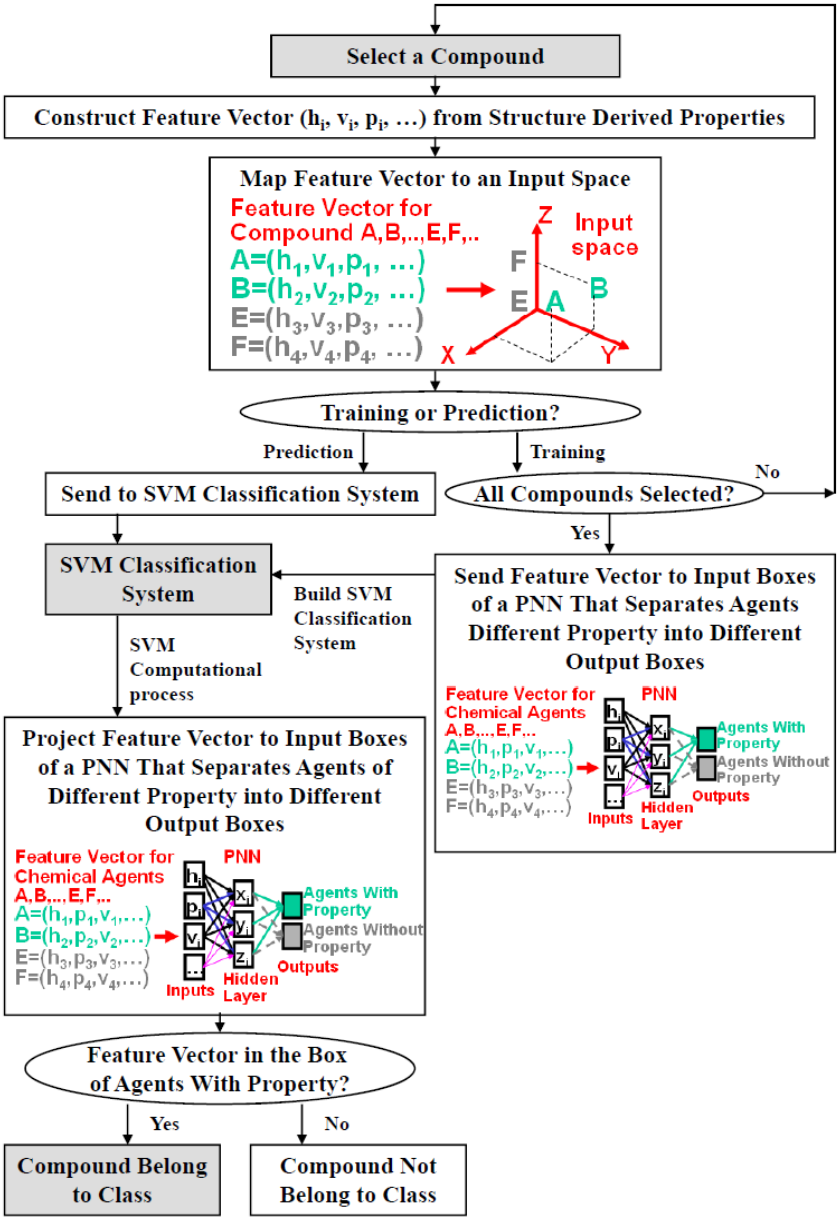


Figure 2-3 Schematic diagram illustrating the process of the prediction of the prediction of compounds of a particular property from their structure by using a machine learning method –probabilistic neural networks (PNN). A, B: feature vectors of agents with the property; E, F: feature vectors of agents 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

Compounds similar to at least one compound in a training dataset can be identified by using the Tanimoto coefficient *sim(i,j)*[115]

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

(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, the similarity search was conducted for MDDR compounds. Therefore, in computing *sim(i,j)*, the molecular descriptor vectors $\mathbf{x}_i$s were scaled with respect to all of the MDDR compounds. The cut-off values for similarity compounds are typically in the range of 0.8 to 0.9[157,158]. A stricter cut-off value of 0.9 was used in this work

2.4 Virtual screening model validation and performance evaluation

2.4.1 Model validation

One of the objectives of modeling is to allow prediction of the pharmacological properties of compounds which have not been clinically and biologically tested. Thus it is important to determine the ability of the derived pharmacological property prediction models to predict the properties of compounds that are not present in the training set. The validation methods used in this work are 5-fold cross validation and independent validation

dataset. In 5-fold cross validation, compounds are randomly divided into five subsets of approximately equal size. Four subsets are used as a training set for developing a model; the remaining one is used as a test set for evaluating the prediction performance of that model. This process is repeated five times such that every subset is used as a testing set once. The average accuracy of the five time models is used for measuring the generalization capability of that method. However, cross validation methods have a tendency of underestimating the prediction capability of a classification model, especially if important molecular features are present in only a minority of the compounds in the training set[159,160]. Thus a model having low cross-validation accurary can still be predictive[159]. This lead to some studies which suggest that an independent validation dateset may provide a more reliable estimation of the prediction capability of a pharmacological property model[161,162]. An independent validation dataset should ideally be obtained independently from the training set and should be representative of the training set so that it can properly assess the prediction capabilities of the pharmacological property model. It is even better if the validation dataset is composed of newly published experimentally validated chemical compounds with a particular pharmacological property.

2.4.2 Performance evaluation methods

The performance of virtual screening model can be evaluated by the quantity 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), 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 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)[163] are used to measure the overall prediction performance:

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

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

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

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

The model performance in screening large libraries can be typically measured[33] by 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 = *hit-rate / (TP+FN)/(TP+FN+TN+FP)*

(18)

2.4.3 Overfitting problem and its prevention

Overfitting is the phenomenon of building a model that agrees well with the observed data but has no predictive ability (it does not agree with unseen or future data). It is a major concern in machine learning classification methods. There are two main types of overfitting: (1) using a model that is more flexible than it needs to be and (2) using a model that includes irrelevant descriptors[160]. A frequently used method for checking whether a prediction system is overfitted is to compare the prediction accuracies determined by using cross validation methods with those determined by using independent validation sets[160]. An over-fitted classification system is expected to have much higher prediction accuracy for the cross validation sets than that for the independent validation sets.

Chapter 3 Development and Evaluation of High Performance Virtual Screening Tools

Support vector machines (SVM) and other machine-learning (ML) methods have been explored as ligand-based virtual screening (VS) tools for facilitating lead discovery. While exhibiting good hit selection performance, in screening large compound libraries, these methods tend to produce lower hit-rate than those of the best performing VS tools, partly because their training-sets contain a limited spectrum of inactive compounds. In this chapter, we tested whether the performance of SVM can be improved by using training-sets of diverse inactive compounds.

3.1 Introduction

Virtual screening (VS) has been extensively explored for facilitating lead discovery [27,37,164,165] and for identifying agents of desirable pharmacokinetic and toxicological properties[26,166]. Machine learning (ML) methods have recently been used for developing ligand-based VS (LBVS) tools[36,46-49,61,62,167] to complement or to be combined with structure-based VS (SBVS) [37,52-55,59,65-67,69,168-170] and other LBVS[27-31] tools aimed at improving the coverage, performance and speed of VS tools.

ML methods have been used as part of the efforts to overcome several problems that have impeded progress in more extensive applications of SBVS and LBVS tools[33,37]. These problems include the vastness and sparse nature of chemical space that needs to be searched, limited availability of target

structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects. LBVS may in some cases limit the diversity of hits due to the bias of training molecules[168]. Therefore, alternative approaches that enhance screening speed and compound diversity without relying on target structural information are highly desired. ML methods have been explored for developing such alternative VS tools[46,47,61] because of their high-CPU speed (100K data points per hour on 3GHz PC)[62] and capability for covering highly diverse spectrum of compounds[171].

The reported performance of various LBVS and SBVS tools in screening compound libraries of >90,000 compounds is summarized in **Table 3-1**. Caution needs to be raised about straightforward comparison of these reported results, which might be misleading because the outcome of VS strongly depends on the datasets used. The dataset-dependence of VS performance can be illustrated by a test shown in a **subsequent section 3.4** of this chapter. Therefore, the listed results should be viewed as providing very crude pictures about the reported VS performances. While exhibiting equally good hit selection performance, in screening extremely-large (≥1 million) and large (100,000~900,000) libraries, the currently developed ML tools tend to show lower hit-rate (ratio of known hits and the predicted hits) and, in some cases, lower enrichment factor (magnitude of hit-rate improvement over random selection) than the best performing SBVS tools. For instance, in screening extremely-large libraries, the reported yield (percentage of known hits predicted), hit-rate and enrichment factor of ML tools are in the range of

55%~81%, 0.2%~0.7% and 110~795 respectively [172], compared to those of 62%~95%, 0.65%~35% and 20~1,200 by SBVS tools [46-51]. While in screening libraries of ~98,000 compounds the reported hit-rates of some ML tools are comparable to those of SBVS tools, their enrichment factors are substantially smaller. A lower hit-rate gives rise to a higher number of false-hits and a lower enrichment factor suggests that there might be bigger room for further optimizing a VS tool. Hence, there is a need for further improving the hit-rate and enrichment factor of ML tools. It is not uncommon for the pharmaceutical industry to screen >1 million compounds per high-throughput screening campaign[172]. The goal of virtual screening is the drastic reduction of compound libraries to a manageable size for synthesis and biological testing. Therefore, improvement of hit-rate and enrichment factor is highly desirable for developing practically useful ML tools for LBVS.

Two approaches have been explored for minimizing false hits. One is the selection of top-ranked hits, which has been extensively used in LBVS[46-51] and SBVS[52-57]. The other is the elimination of potentially unpromising hits in pre-screening stage by using such filters as Lipinski's rule of five[58 65], and recognition of pharmacophore [54] and specific chemical groups or interaction patterns[52,53,59,60]. In addition to the application of these approaches, the performance of ML tools in screening large libraries may be further improved by using training sets of more diverse spectrum of compounds to develop more optimally performing ML models. These models have been generated by using two-tier supervised classification ML methods[36,46-49,61-63], which require training sets of diverse spectrum of active and inactive compounds.

The training inactive compounds in these models have been collected from up to a few hundred known inactive compounds or/and putative inactive compounds from up to a few dozen biological target classes in MDDR database[36,46-49,61-63], which may not always be sufficient to fully represent inactive compounds in the vast chemical space, thereby making it difficult to optimally minimize false hit prediction rate of ML models.

Table 3-1 Comparison of the reported performance of different virtual screening (VS) methods in screening large libraries of compounds.

Type of VS method and size of compound libraries screened	VS method (number of studies) [references]	Compounds screened			Virtual hits selected by VS method		Known hits selected by VS method			
		No of compounds	No of known hits	Percent of known hits	No of compounds selected as virtual hits	Percent of screened compounds selected as virtual hits	No of known hits selected	Yield	Hit rates	Enrichment factor
Structure-based VS, extremely large libraries (≥1M)	Docking + pre-screening filter (2)[52,65]	1M~2M	355~630	~0.03%	1K~60K	0.08%~3%	340~390	62%~ 95%	0.65%~ 35%	20~1200
Structure-based VS, large libraries	Docking + pre-screening filter (11)[53-55,59,66,67,69]	134K~400K	100~1016	0.12%~ 0.76%	375~4.5K	0.28%~3%	5~231	2%~ 30%	0.11%~ 17%	4~66
Ligand-based VS (machine learning), extremely large libraries (≥1M)	Machine learning - SVM (2)[46,48,62]	2.5M	22~46	0.0009%~ 0.0018%	2.5K~11K	0.1%~0.45%	18~25	55%~ 81%	0.2%~ 0.7%	110~795
Ligand-based VS (machine learning), large libraries	Machine learning – SVM (2)[47]	172K	118~128	~0.07%	1.7K	1%	26~70	22%~ 55%	1.5%~ 4.1%	22~55
	Machine learning – SVM (11)[36]	98.4K	259~1146	0.26%~ 1.16%	984	1%	131~710	44%~ 69%	14%~ 72%	44~69
	Machine learning – BKD (12)[47-49,62]	101K~103K	259~1166	0.25%~ 1.2%	5.1K	5%	65~972	14%~ 94%	1.2%~ 18.9%	3~19
	Machine learning – LMNB (1)[48,62]	172K	118	0.069%	1.7K	1%	19	16%	1%	15
	Machine learning – CKD (18)[36]	98.4K	259~1211	0.26%~ 1.23%	984	1%	132~960	34%~ 94%	13%~ 98%	53~94
Ligand-based VS (clustering), large libraries	Hierachical k-means (5)[29]	344.5K	91~1556	0.026% ~0.45%	3750~21285	1.1%~6.2%	27~761	23% ~55%	0.72%~5%	7.97~31.2
	NIPALSTREE (5)[29]	344.5K	91~1556	0.026% ~0.45%	3469~28125	1.0%~8.2%	17~625	18% ~50%	0.49%~ 2.8%	3.51~18.7
	Hierachical k-	344.5K	91~155	0.026%	7317~4316	2.1%~12.3%	30~980	33% ~72%	0.41% ~2.9%	4.86~17.6

	means + NIPALSTREE disjunction (5)[29]		6	~0.45%	5					
	Hierachical k-means + NIPALSTREE conjunction (5)[29]	344.5K	91~1556	0.026% ~0.45%	538~6692	0.16%~1.9%	14~406	6% ~32%	1.1% ~10.2%	7.77~98
Ligand-based VS (structural signatures), extremely large libraries (≥1M)	Pharmacophore (3) [30,70,71]	1.77M~3.8M	55~144	0.0014% ~0.0081%	20K~1M	1.15%~26%	6~39	11% ~70%	0.0039%~ 0.084%	3~10.3
Ligand-based VS (structural signatures), large libraries	Pharmacophore (1) [31]	380K	30	0.0079%	6917	1.82%	23	76.7%	0.33	41.8
Ligand-based VS, extremely large libraries (≥1M) for HIV protease, inhibitors DHFR inhibitors, Dopamine antagonists, CNS active agents	SVM	2.986M	2351	0.076%	8157	0.27%	1833	78.0%	22.5%	296
	SVM	2.986M	225	0.007%	160	0.0054%	118	52.4%	73.8%	10543
	SVM	2.986M	37	0.0012%	299	0.01%	23	62.2%	7.7%	6417
	SVM	2.986M	664	0.022%	9502	0.32%	442	66.6%	4.7%	214

In this work, we examined to what extent hit rate and enrichment factor of ML tools can be improved by using training-sets of more diverse spectrum of inactive compounds. A widely used and better performing ML method, support vector machines (SVM) [36,46,47,49,62,167], was used to develop SVM models for identifying active compounds of single mechanism (HIV-1 protease inhibitors, dihydrofolate reductase (DHFR) inhibitors, dopamine receptor antagonists) and multiple mechanisms (central nervous system (CNS) active agents). HIV-1 protease inhibitors form an important class of anti-HIV agents some of which have been successfully used clinically[41]. DHFR inhibitors are useful for the treatment of microbial infections[173], cancer[174], and parasitic diseases[175]. Dopamine antagonists have been used as antipsychotic agents[176] and for the treatment of cervical dystonia[177], vertigo[178], and gastrointestinal motility disorders[179]. CNS active agents are composed of a diverse spectrum of CNS acting compounds that produce anxiolytic, antipsychotic, antidepressant, analgesic, anticonvulsant, antimigraine, antiischemic, antiparkinsonian, nootropic, neurologic, epileptic, neuroleptic, neurotropic, neuronal injury inhibiting, narcotics antagonizing, and CNS stimuating effects[180]. Because of their diverse therapeutic applications and structural frameworks, these compounds are highly useful for testing the performance of SVM and other ML tools in LBVS of large compound libraries.

Our SVM models were trained by using known active compounds and putative inactive compounds extracted from compound families that contain no known active compound. Compound families can be generated by

clustering distinct compounds of chemical databases into groups of similar structural and physicochemical properties[29]. The developed SVM models were tested in screening libraries of 2.986 million compounds from the PubChem database that are not in the training sets of these SVM models. The yields, hit-rates and enrichment factors derived from these tests were compared with those of SBVS and other LBVS tools applied in the screening of extremely-large libraries to determine to which extent the overall performance of SVM models can be enhanced and whether it is comparable to that of the best performing VS tools reported in the literature. To further evaluate whether our SVM models predict active and inactive compounds rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

3.2 Methods

3.2.1 Collection of active compounds

Table 3-2 gives the statistics of collected active compounds for the four active compound classes and their structural diversity index (DI) (defined in **Methods Chapter section 2.1.3**). The structures of a few selected compounds for each class are shown in **Figure 3-1**. For comparison of structural diversity of the compounds in these and those of the other structurally diverse classes, the statistics and DI values of several such classes are also listed in **Table 3-2**. A total of 5,161 HIV-1 protease inhibitors, with log (IC50) values in the range of -7.85 to -3.30, were selected from the HIV/OI Enzyme Inhibition Database of the National Institute of Allergy and Infectious Diseases of NIH. 76.6% of which are peptide-based inhibitors (66% and 5% are peptidomimetics and

symmetry-based inhibitors respectively) and 23.4% are non-peptide-based inhibitors. The quality of these inhibitors were further validated against literature reports we found from the literature database PUBMED to ensure that they have been described as HIV-1 protease inhibitors with IC50 values in the range of binding potencies considered to be important in various cases.

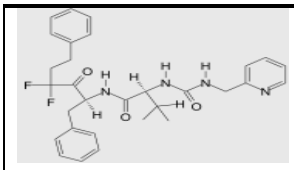
Table 3-2 Diversity index (DI) and number of HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents used for developing support vector machines ligand-based virtual screening tools. For comparison, relevant data of several other compound classes of highly diverse structures are also included. These compound classes are arranged in descending order of structural diversity.

Chemical Class	No. of Active Compounds	DI Value
Blood-brain barrier penetrating agents [181]	276	0.430
FDA approved drugs	1,121	0.495
NCI diversity set	1,804	0.544
P-glycoprotein substrates [130]	116	0.555
CYP 2D6 inhibitors	180	0.575
CNS active agents (this work)	16,182	0.578
CYP 2D6 substrates	198	0.588
Human intestine absorbing agents [182]	131	0.596
Estrogen receptor agonists [167]	243	0.618
HIV protease inhibitors (this work)	5,161	0.626
DHFR inhibitors (this work)	755	0.719
Dopamine antagonists (this work)	1,184	0.741

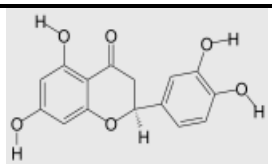
DHFR inhibitors were collected from a publication[183]. We were able to use our software[182] to generate molecular descriptors of 755 of the 756 collected inhibitors. We collected 1,184 distinct dopamine antagonists from three separate sources, which include 1,163 from MDDR database, 126 from PubChem database, and 41 from a publication[184]. CNS active agents were

retrieved from those compounds in MDDR database annotated as anxiolytic, antipsychotic, antidepressant, analgesic (non-opioid and opioid), anticonvulsant, antimigraine, antiischemic (cerebral), antiparkinsonian, stimulant in central, antagonist to narcotics, centrally acting agent, nootropic agent, neurologic agent, epileptic, and neuronal injury inhibitor/neuroleptic/neurotropic. We were able to use our software[182] to derive molecular descriptors for 16,182 of the collected 16,390 non-redundant CNS active compounds. Molecular descriptorsof part of active compounds cannot be calculated because of non-availability of their reasonable 3D structures.

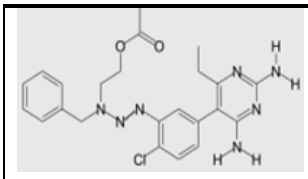
(a) HIV-1 protease inhibitors

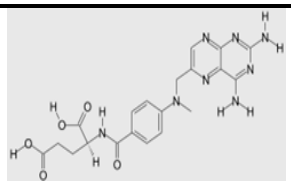

C452983

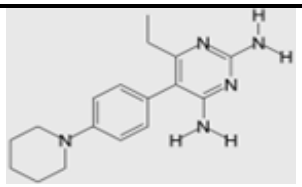

C10459


C373261

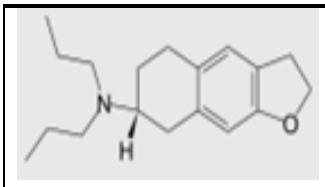
(b) DHFR inhibitors

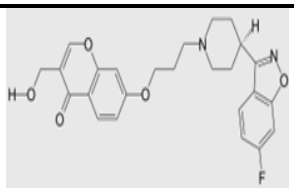

C1987

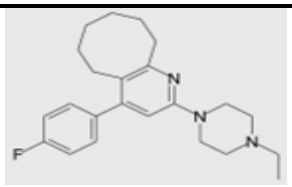

C4112


C341273

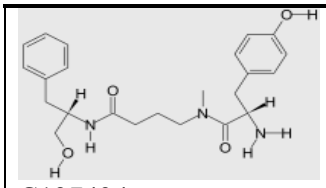
(c) Dopamine antagonists

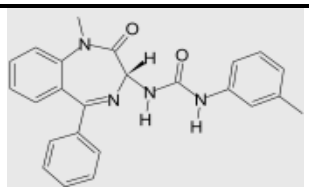

C115007


C3037308


C125564

(e) CNS active agents


C127434


C104929

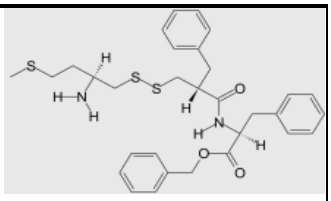

C119600

Figure 3-1 Structures of the selected HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents. The PubChem accession number of these compounds is given.

3.2.2 Generation of putative inactive compounds

Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds[36,46-49,61-63], a new approach extensively used for generating inactive proteins in ML classification of various classes of proteins[185-187] may be applied for generating putative inactive compounds. An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the "inactive" chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. 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. As will be demonstrated, such an adverse effect is expected to be relatively small for many biological target classes.

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[129], 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[188]. Importantly, 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[188]. 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[185-188].

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors[29,189]. As ML methods predict compound activities based on their molecular descriptors, in developing ML tools, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a K-means method [29,189] and molecular descriptors computed from our own software[182], we generated 7,990 cluster families 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[3], and the 2,851 clusters for 171,045 natural products[190]. 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. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered "active" families in PubChem database is expected to be relatively small, 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, putative inactive compounds can be generated 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 signalling molecules in the CNS system[180]. 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 likelihood 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.

3.2.3 Molecular descriptors

A total of 199 descriptors derived by using our software[182] were used in this work. The details of the molecular descriptors are explained in **Chapter 2 Section 2.2**.

3.2.4 Development of support vector machines virtual screening tools

SVM models for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents were developed by a procedure widely used for developing SVM protein classification models of optimal performance[185-187]. In the first step, active and inactive compounds were each divided into separate training, testing and independent evaluation sets. Specifically, active and inactive compounds were each clustered into groups based on their distance in the molecular descriptor space by using a hierarchical clustering method[132] . An upper-limit of the largest separation of 20 was used for each cluster. One representative compound was randomly selected from each group to form a training set that is sufficiently diverse and

broadly distributed in the descriptor space. One or up to 50% of the remaining compounds in each group were randomly selected to form the testing set. The selected compounds from each group were further checked to ensure that they are distinguished from those of other groups. The remaining compounds were used as the independent evaluation set, which are also of reasonable level of diversity. Moreover, an analysis of the compounds in each cluster shows that the majority of the compounds in a cluster are substantially different. Thus, the testing and independent evaluation sets are expected to have certain level of usefulness for performing their task of fine-tuning the parameter of a SVM model and for evaluating its prediction performance. In the second step, SVM models were trained by using the training set and their parameters were optimized by using the testing set. The SVM model with the best overall performance on both the testing and independent evaluation sets was selected as a VS tool.

3.3 Assessment of virtual screening performance

The developed SVM models for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in screening 2.986 million distinct compounds from the PubChem database that are not in the training sets of our developed SVM models. The performance of these SVM models is given in **Table 3-3**, which can be compared with the reported performance of other SBVS and LBVS tools listed in **Table 3-1**. There are 2,351, 225, 37, and 664 known HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in the PubChem database not in the training sets of our SVM models. Our SVM models were able to identify 78.0%, 52.4%, 62.2%, and 66.6% of these known hits, which are comparable

to the range of 62%~95% by the SBVS tools[52,65] and 55%~81% by other LBVS [46,48,62] tools in screening libraries of ≥1 million compounds, and they are also comparable to the percentages in screening libraries of 98,400~344,500 compounds by other SBVS[53-55,59,66,67,69,168-170] and LBVS tools[29,31,36,47-49,62]. These results suggest that our developed SVM models are equally effective in selecting potential hits in VS of large libraries.

In addition to the exhibition of equally effective hit selection performance, our SVM models appear to show relatively lower "false" hit identification rate. Without the use of top-ranked cut-off or additional filter, our SVM models identified a total of 8,157, 160, 299, and 9,502 virtual hits for the four compound classes respectively, which are comparable to and in some cases smaller than those identified by SBVS[52-55,59,65-67,69,168-170] and other LBVS[29,36,46-48,62,150,151] tools even though a substantially larger number of compounds (2.983M vs. 98.4K~2.5M) were screened. As a result, smaller percentages of screened compounds were selected as virtual hits, which are in the range of 0.0054%~0.32% as compared to those of 0.08%~3% by SBVS tools[52-55,59,65-67,69,168-170], 0.1%~5% by other reported ML models[36,47-49,62], 0.16%~82.% by clustering methods[29] , and 1.15%~26% by pharmacophore models[30,31,70,71]. By using Lipinski's rule of five[58] as a filter, the numbers of identified virtual hits are further reduced to 333, 115, 209, and 8,035 for the four compound classes respectively, suggesting that introduction of such filters or combination with other VS methods may enable further reduction of the number of predicted hits.

The hit-rates of our SVM models are 22.5%, 73.8%, 7.7%, and 4.7% for the four classes of compounds respectively, which are comparable to those of 0.65%~35% by SBVS tools and substantially improved against those of 0.2%~0.7% by other reported SVM models in screening extremely large libraries. These hit-rates are also greater than the majority of the hit-rates in screening large libraries of 98,400~344,500 compounds by SBVS and other LBVS tools. The enrichment factors of our SVM models are 296, 10,543, 6,417, and 214 for the four classes of compounds respectively, which are substantially improved against those of 20~1,200 by SBVS tools and 110~795 by other reported SVM models in screening extremely large libraries. Therefore, our method is useful in improving the hit-rate and enrichment factor of SVM while maintaining an equally high hit identification rate as other SBVS and LBVS tools.

To further evaluate whether our SVM models predict active compounds rather than membership of certain compound families, compound family distribution of the predicted active and inactive compounds for the four compound classes were analyzed. As shown in Table 3-3, 24.3%, 71.3%, 87.6%, 85.7% of the predicted HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents belong to the families that contain no known active compound. For those families that contain at least one known active compound, >70% of the compounds (>90% in majority cases) in each of these families were predicted as inactive compounds by our SVM models. These results suggest that our SVM models predict active compounds rather than membership to certain compound families. Some of the predicted active

compounds not in the family of known active compounds may serve as potential "novel" active compounds. Therefore, as in the case shown by an earlier study[191], SVM methods have certain capacity for predicting novel active compounds.

3.4 Comparative analysis of virtual screening performance of our method

The performance of our method can be more appropriately evaluated by using it to develop VS tools and test them based on the same dataset construction and testing procedures as those used in other VS methods. In this work, we specifically developed additional VS prediction models by using the same dataset construction method and same data source of a standard similarity-based method, the data fusion method[48], the performance of both methods were then compared by using the same data source. The data fusion method is based on Taminoto based similarity searching using multiple reference compounds, which have shown good performances for a number of active compound groups by using only a small number of training active compounds[48], and thus is a good reference method for evaluating the performance of our method.

Table 3-3 Performance of support vector machines virtual screening tools developed in this work for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in screening 2.986 million compounds.

Screening task	Compounds screened				Virtual hits selected by SVM					Known hits selected by SVM			
	No of compounds	No of known hits not in training sets of SVM-LBVS tool	Percent of known hits	No of families covered by known hits	No of selected virtual hits	Percent of selected virtual hits not in the families covered by known hits	Percent of screened compounds selected as virtual hits	No of selected virtual hits passed rule-of-five	Percent of selected virtual hits passed rule-of-five and not in the families covered by known hits	No of known hits selected	Yield	Hit rates	Enrichment factor
HIV protease inhibitors	2.986M	2351	0.076%	496	8157	24.3%	0.27%	333	42.6%	1833	78.0%	22.5%	296
DHFR inhibitors	2.986M	225	0.007%	60	160	71.3%	0.0054%	115	64.4%	118	52.4%	73.8%	10543
Dopamine antagonists	2.986M	37	0.0012%	29	299	87.6%	0.01%	209	82.8%	23	62.2%	7.7%	6417
CNS active agents	2.986M	664	0.022%	519	9502	85.7%	0.32%	8035	84.1%	442	66.6%	4.7%	214

We developed three separate HIV protease inhibitor VS tools by using our method and datasets of similar sizes and from the same sources as that used by the reported studies of the data fusion method[48,192]. Our training and testing datasets were generated from 1,054 HIV protease inhibitors extracted from the MDDR database. Based on the training set generation procedure of the data fusion method[192], three sets of 60, 80 and 100 inhibitors were selected from this full set of 1,054 inhibitors as the active compound training sets, from which the inactive compound training sets were generated by using our method. Using the same testing method of the data fusion method, the performance of the three developed SVM VS tools were evaluated by using the remaining 994, 974 and 954 HIV protease inhibitors respectively, which showed that 59.5%, 62.2% and 67.3% of these remaining inhibitors were correctly identified. The performance of these SVM VS tools is similar to and in some cases slightly improved against that of 55.2%~58.0% of the data fusion method that used a similar number of training HIV protease inhibitors[48]. This suggests that, by using the equally small active compounds as training data, our SVM model is capable of performing at the same level and in some cases slightly improved level than that of the data fusion method.

3.5 Discussion

The performance of SVM and other ML methods critically depends on the diversity of compounds in a training dataset and the appropriate description of the compounds. The datasets used in developing ML models described in Table 3-1 and in this work are not expected to be fully representative of all of the active and inactive compounds. Known inactive compounds, particularly those structurally similar to an active compound, may serve to further refine

ML models at higher "structural resolutions" than those achievable by using only the putative inactive compounds generated from this work. Mining of known active compounds and inactive compounds from the literature[103] and other sources [193,194] is a key to developing more optimally performing ML models for VS.

Examination of incorrectly predicted compounds by ML models consistently suggests that the currently-used molecular descriptors are insufficient to adequately represent some of the compounds that contain complex structural or chemical configurations[130,149,181]. Examples of these agents are those with large rigid structures combined with a short flexible hydrophilic tail, compounds that contain multi-rings with various hetero atoms such as nitrogen, oxygen, sulphur, fluorine and chlorine. Due to the limited coverage of the number of bond links in a hetero-atom loop, the currently available topological descriptors are not yet capable of describing the special features of a complex multi-ring structure that contains multiple hetero atoms. It appears that none of the currently-available descriptors are capable of fully representing molecules containing a long flexible chain. Therefore, it might be helpful to explore different combination of descriptors and to select more optimal set of descriptors by using more refined feature selection algorithms and parameters[130,195]. However, indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each other, may introduce noise as well as extend the coverage of some the aspects of these special features. Thus, it may be necessary to introduce new descriptors for more appropriately representing these and other special features.

3.6 Further perspective

By using training sets of more diverse spectrum of inactive compounds, the hit-rates and enrichment factors of SVM models can be substantially improved to the level comparable to and in some cases higher than those of the best performing SBVS and LBVS tools reported in the literature. Because of their high computing speed and capability for covering highly diverse spectrum compounds, SVM and other ML methods can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating lead discovery[65,69,71].

Chapter 4 Evaluation of Virtual Screening by Sparsely Distributed Active Compounds

Virtual screening performance of support vector machines (SVM) depends on the diversity of training active and inactive compounds. While diverse inactive compounds can be routinely generated, the number and diversity of known actives are typically low. In this chapter, we evaluated the performance of SVM trained by sparsely distributed actives in six MDDR biological target classes composed of high number of known actives of high, intermediate, and low structural diversity (muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors).

4.1 Introduction

As part of the efforts in further developing virtual screening (VS) methods for facilitating lead discovery[27,37,164,165], support vector machines (SVM)[167]have recently been explored as ligand-based VS (LBVS) tools to complement or to be used in combination with structure-based VS (SBVS)[37,52-55,65-67,69,168-170] and other LBVS[27] tools. A particular objective for exploring these approaches is to overcome several problems that have impeded progress in more extensive applications of VS[33,37,168]. These problems include the vastness and sparse nature of chemical space to be searched, limited hit diversity due to the bias of training molecules, limited availability of target structures (only 15% of known proteins have experimentally-determined 3D structures), complexity

and flexibility of target structures, and difficulties in computing binding affinity and solvation effects.

SVM is of particular interest because it classifies active compounds based on the differentiating physicochemical profiles between active and inactive compounds rather than structural similarity to active compounds *per se*. Moreover, SVM does not require the knowledge of target structure and activity-related molecular descriptors, and the computation of binding affinity and solvation effects. Its fast speed enables efficient search of vast chemical space. Some of these advantages have been exhibited by good VS performance in screening large compound libraries[48,62,108]. None-the-less, as in the cases of all statistical learning methods, the performance of SVM is significantly influenced by the levels of the training active and inactive compounds in representing the physicochemical profiles of the remaining compounds in the chemical space.

Active compounds (actives) typically occupy small pockets of the chemical space. It may be possible to construct a training active dataset to substantially represent the properties of the remaining actives by using relatively small number of known actives. However, inactive compounds (inactives) generally occupy larger portions of the chemical space. A large number of training inactives is needed to reach sufficient level of diversity for representing the remaining inactives in the chemical space. SVM constructs a hyper-plane in a higher dimensional molecular descriptor space to separate actives from inactives based on whether or not the molecular descriptor vector of a

compound is distributed on the known active side of the hyper-plane. As illustrated in **Figure 4-1**, the position and orientation of the SVM hyper-plane, which extends to far regions of the chemical space, can in many cases be influenced by inactives distributed remotely from the known actives as well as those closely resembling known actives. The level of influence tends to be stronger for sparsely distributed known actives and inactives as there is more room in the local space for altering the position and orientation of the hyper-plane. Therefore, highly diverse inactive datasets are typically needed for constructing SVM VS models[33,108].

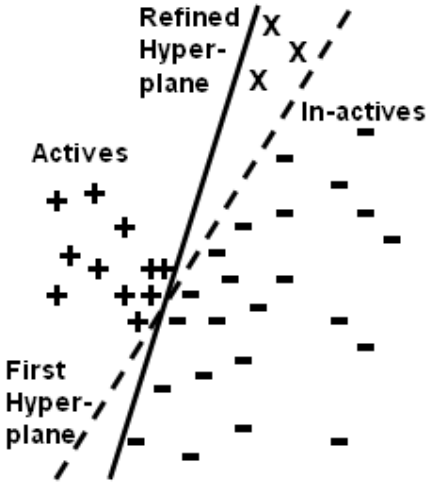


Figure 4-1 Illustration of the influence of the inactive compounds distributed far away from the active compounds on the position and orientation of the hyper-plane of support vector machines that separates active and inactive compounds. +: active compounds, -: inactive compounds used for constructing the first hyper-plane (dashed line), x: additional inactive compounds used for constructing the more-refined hyper-plane (solid line).

Highly diverse inactive training datasets can be routinely generated by large-scale sampling of active compounds of other biological target classes[36,48,62,63] and by using representative compounds from compound families that contain no known actives[108]. In contrast, the diversity and the level of representation

of active training datasets are often constrained by the small number of known actives sparsely distributed in the active regions of chemical space (active regions are defined as regions of chemical space covered by discovered and yet-to-be-discovered actives). There is a need to evaluate the VS performance of SVM trained by sparse active datasets to determine its capability in identifying novel actives from sparsely distributed known actives.

In this work, we examined the VS performance of SVM trained by sparse active datasets generated from available active datasets of sufficiently high number of known actives and varying degrees of structural diversity. The high number of actives in the studied datasets makes it possible to generate sufficiently sparse training active datasets, and varying degrees of diversity enables objective evaluation of the VS performance of SVM on different classes of actives. To facilitate comprehensive analysis and further comparative studies, six of the well-studied MDDR biological target classes[48] of high number of actives (983~1,645) of both high, intermediate, low structural diversity were used for this study. These classes include muscarinic M1 receptor agonists and NMDA receptor antagonists representing high-diversity, thrombin inhibitors and HIV protease inhibitors representing intermediate-diversity, and cephalosporins and rennin inhibitors representing low-diversity classes respectively.

Muscarinic M1 receptor agonists are useful for the treatment of Alzheimer's disease by improving the performance in cognitive tests in Alzheimer's patients[196]. NMDA receptor antagonists have been explored for

neuroprotection[197] and the treatment of postoperative pain[198]. Thrombin inhibitors produce anticoagulant effects and have been used as antithrombotic agents[199]. HIV protease inhibitors form an important class of anti-HIV agents some of which have been successfully used clinically[41]. Cephalosporins are in clinical development as broad-spectrum antibacterial agents[200]. Rennin inhibitors have shown effectiveness in cardiovascular pharmacotherapy[201]. Because of their diverse therapeutic applications and structural frameworks, these compounds are highly useful for testing the performance of SVM as well as other methods[48].

For each biological target class, two training datasets were generated. A regularly sparse active dataset, which contains the same number of actives as those in earlier sparse dataset studies[33,48] was generated by extracting 100 actives (representing 6.1%~10.2% of the known actives) scattered in the known active region of chemical space. A very sparse active dataset was generated by extracting 40 active compounds (representing 2.4%~4.1% of the known actives) scattered in the known active region of chemical space. To generate a dataset of N number of actives from a larger number actives, all actives were clustered into N clusters followed by the extraction of one compound from each of these clusters. Putative inactive datasets were generated by extracting representative compounds from all compound families that contain no known active compound[108]. Compound families can be generated by clustering distinct compounds of chemical databases into groups of similar structural and physicochemical properties[29,189].

The regularly sparse active datasets were used for facilitating crude estimation of the level of performance of our SVM VS tools with respect to those of other VS tools such as the data fusion method[48] and other methods[33] that have been frequently developed by using ~100 active compounds. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. To further evaluate whether the performance of our SVM VS tools are attributed to the SVM classification models or the molecular descriptors used, a study was conducted to compare the performance our SVM VS tools with that of the Tanimoto-based similarity searching method[115] using the same datasets and the same molecular descriptors.

The yields (percent of testing actives identified as active) of our SVM VS tools were estimated by using the remaining 89.7%~97.4% of the known actives. The false-hit rates (percent of inactives identified as active) of our SVM VS tools were estimated by using the remaining 167K MDDR compounds outside the training datasets and by using the 9.997M PubChem compounds that exclude the known actives. To further evaluate whether our SVM VS tools predict active and inactive compounds rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

VS performance may be over-estimated by training datasets that contain higher percentages of inactives significantly different from the known actives,

because the easily distinguishable features may make VS enrichments appearing artificially good[202]. Therefore, VS performance may be more strictly tested by using subsets of inactives that resemble the physicochemical properties of the known actives so that enrichment is not simply a separation of trivial physicochemical features[158]. In this work, the performance of our SVM VS tools was further evaluated by the subsets of MDDR compounds that are similar in physicochemical properties to those of the known actives.

4.2 Methods

4.2.1 Construction of active training and testing datasets

All actives of the six biological target classes are from MDDR, from which we obtained 983 muscarinic M1 receptor agonists, 1,510 NMDA receptor antagonists, 1,252 thrombin inhibitors, 1,054 HIV protease inhibitors, 1,645 cephalosporins, and 1,241 rennin inhibitors. The structure of representative compounds of these six classes is shown in **Figure 4-2**. To generate the popular-sized sparse and highly sparse active training datasets and the corresponding testing datasets, all known actives of each of these classes were clustered into 100 and 40 clusters respectively by using a K-means method[29,189] and molecular descriptors computed from our own software[182]. For each class, the regularly sparse and very sparse active training datasets of 100 and 40 active compounds were generated by extracting one compound from each of the 100 and 40 active clusters respectively. The remaining actives were used as the corresponding active testing set.

4.2.2 Generation of putative inactive training and testing datasets

Methods in **Chapter 3 Section 3.2.2** are used to generate putative inactive training and testing datasets.

The classes of muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors are distributed in 203, 538, 161, 281, 95, and 138 families respectively. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered "active" families in PubChem database is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered "active" families (hundreds) and the families that contain no known active compound (~8,993 based on the current versions of PubChem and MDDR) for these and possibly many other target classes is expected to be <15%. Therefore, putative inactive training datasets can be generated by extracting a few representative compounds of those families that contain no known active compound in the active training set, with a maximum possible "wrong" family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class, and with the expectation that a substantial percentage of active members in the putative "inactive" families can be classified as active despite of their family representatives are placed into the inactive training sets.

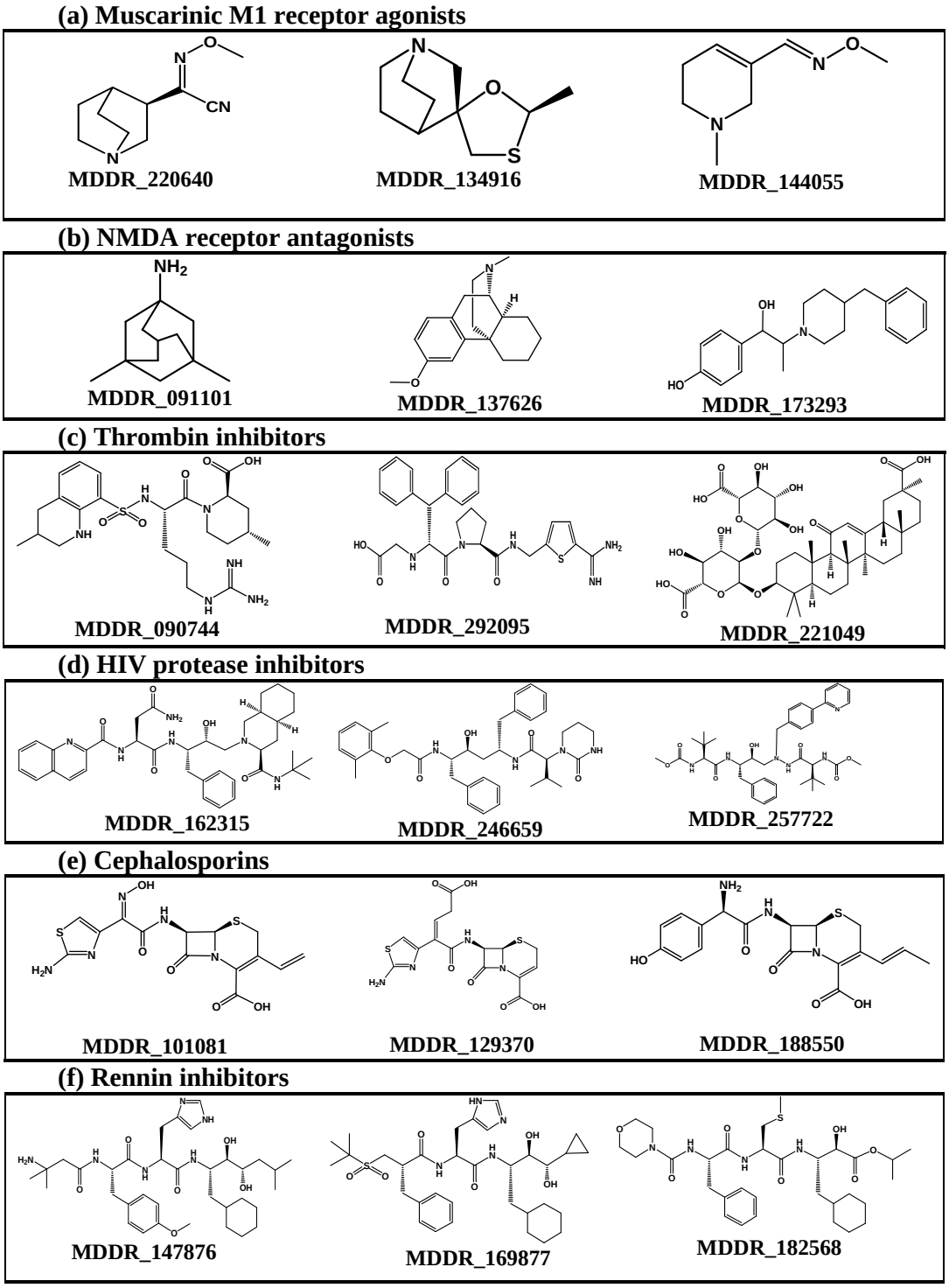


Figure 4-2 Structures of the selected muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors. PubChem accession number of these compounds is given.

There are 8790, 8455, 8832, 8712, 8898, and 8855 compound families that contain no known muscarinic M1 receptor agonist, NMDA receptor antagonist, thrombin inhibitor, HIV protease inhibitor, cephalosporin, and rennin inhibitor respectively. Thus the inactive training dataset corresponding to each sparse or biased active training dataset was generated by random selection of 5~6 representative compounds from each of these "inactive" families and those active families with none of their members in the active training set. The remaining compounds of the "inactive" families in PubChem and MDDR can be used as putative inactive testing sets. It is noted that 9.6%~68.7% of the active containing families are not covered in the active training set, and their representative compounds were deliberately placed into the inactive training set as they are not supposed to be known in our study. As shown in an earlier study[48,49,62] (**Chapter 3**) and in this work, a substantial percentage of the active compounds in these misplaced active containing families were predicted as active by our SVM models. Moreover, a small percentage of the compounds in these putative inactive datasets are expected to be un-reported and un-discovered actives for each of the six biological target classes, their presence in these datasets is not expected to significantly affect the estimated false positive rate of the developed SVM VS tools.

4.2.3 Molecular descriptors

A total of 98 important descriptors were chosen from the chemical descriptors calculated by our program MODEL which were used in this work. The detail about molecular descriptors is explained in **Chapter 2 Section 2.2**.

4.3 Results and discussion

4.3.1 Comparative analysis of virtual screening performance of SVM trained by regularly sparse active datasets

It is of interest to evaluate the performance of SVM trained from regularly sparse active datasets by comparison with literature reported VS performance based on similar dataset construction/testing procedures and the same data sources. As discussed in the introduction section, the comparison of these results should be viewed as providing very crude pictures about the level of performance of SVM. In this work, we specifically compared the performance of SVM VS tools with those a standard similarity-based method, the data fusion method[48,192]. The data fusion method is based on Tanimoto-based similarity searching using multiple reference compounds, which have shown good performances for a number of active compound groups by using only a small number of ~100 training active compounds[48,192], which serves as a good reference method for evaluating the performance of SVM. To further evaluate whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of the Tanimoto-based similarity searching method based on the same training and testing datasets and molecular descriptors.

The statistics of the regularly sparse active datasets, the performance of our method, the reported performance of the data fusion method, and the results of the Tanimoto-based similarity searching method for the six biological classes are given in **Table 4-1**. As shown in **Table 4-1**, the percentage of known actives in these datasets is in the range of 6.1%~10.2%. The percentage of

"active" families (defined as the families that include at least one known active compound) covered by these datasets is in the range of 15.4%~67.2% with five of the sets below 31.5%. Therefore, these datasets are reasonably sparse.

By using the same testing procedure of the data fusion method, the performance of the six developed SVM VS tools were evaluated by using the remaining 883~1,545 actives and ~167K MDDR compounds of other biological target classes. The yields of our SVM VS tools are 26.7%~49.5% for the high, 60.0%~67.3% for the intermediate, and 82.1%~91.9% for the low diversity classes respectively. The reported yields of the data fusion method are 15.7%~46.6% for the high, 44.5%~58.0% for the intermediate, and 90.4%~94.7% for the low diversity classes respectively[48,192]. The false-hit rates (estimated from the percentage of the ~167K MDDR compounds of other biological target classes identified as active) of our SVM VS tools are in the range of 1.0%~2.9%. The false-hit rates of data fusion method can be deduced as 4% based on the reported top 5% hit selection criterion from ~150K compounds of other MDDR biological target classes[48,192]. Compared with those of data fusion method, the yields of our SVM VS tools are slightly improved for the high and intermediate classes, and the false-hit rates of our SVM VS tools are substantially reduced for all three classes. These results suggest that, by using the equally small number of active compounds as training data, SVM is capable of producing equally good or slightly better yields and generalization capability at substantially reduced false-hit rates than those of the data fusion method.

As shown in **Table 4-1**, the yields of the Tanimoto-based similarity searching method are 9.4%~ 24.2% for the high, 19.0%~27.8% for the intermediate, and 38.4%~39.3% for the low diversity classes respectively. The false-hit rates are in the range of 3.3%~4.4%. Compared to these results, the yields of SVM are significantly improved and the false-hit rates of SVM are substantially reduced. This suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used.

Table 4-1 Dataset statistics and the virtual screening performance of support vector machines developed by using regularly sparse datasets of 100 active compounds for screening MDDR database. The results are compared with that of the Tanimoto similarity searching method using the same dataset and molecular descriptors, and with the reported performance of similarity search methods trained by using ~100 active compounds (Ref 48) for identifying muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors. Known "active" chemical families refer to chemical families that contain at least one known active compound. Yields and false hit rates are the percent of testing active compounds identified as active.

Compound Diversity Category Defined in Ref 48	**Compound Biological Target Class (No of compounds) [average mean pair-wise similarity value computed in Ref 48]**	**Active Compounds in Training Set**		**Active Compounds in Testing Set**		**SVM Virtual Screening Performance (This Work)**		**Virtual Screening Performance of similarity searching methods reported in Ref 48**		**Virtual Screening Performance of Tanimoto similarity searching method (This work)**	
		No and Percent of Active Compounds	**No and Percent of Known "Active" Chemical Families Covered by Active Compounds**	**No and Percent of Active Compounds**	**No and Percent of Known "Active" Chemical Families Covered by Active Compounds**	**Yields**	**False Hit Rates**	**Yields**	**False Hit Rates**	**Yields**	**False Hit Rates**
High	Muscarinic M1 receptor agonists (983) [0.206]	100 (10.2%)	64 (31.5%)	883 (89.8%)	171 (84.2%)	49.5%	1.7%	27.4%~46.6%	4%	24.2%	3.9%
	NMDA receptor antagonists (1510) [0.199]	100 (6.6%)	83 (15.4%)	1410 (93.4%)	503 (93.5%)	26.7%	2.8%	15.7%~20.7%	4%	9.4%	4.4%

Intermediate	Thrombin inhibitors (1252) [0.321]	100 (8.0%)	46 (28.6%)	1152 (92.0%)	227 (91.7%)	60.0%	2.9%	44.5%~52.3%	4%	19.0%	4.3%
	HIV protease inhibitors (1054) [0.313]	100 (9.5%)	74 (26.3%)	954 (90.5%)	248 (88.3%)	67.3%	2.9%	51.6%~58.0%	4%	27.8%	4.4%
Low	Cephalosporins (1645) [0.501]	100 (6.1%)	43 (67.2%)	1545 (93.9%)	78 (82.5%)	82.1%	1.0%	NA	NA	39.3%	3.7%
	Rennin inhibitors (1241) [0.459]	100 (8.1%)	51 (37.0%)	1141 (91.9%)	121 (87.7%)	90.9%	1.8%	90.4%~94.7%	4%	38.4%	3.3%

4.3.2 Virtual screening performance of SVM trained by very sparse active datasets

The level of sparseness of the very sparse active datasets for the six biological target classes can be measured by the percentage of known actives in these training sets and the percentage of "active" families they occupy. As shown in **Table 4-2**, the percentage of known actives in the sparse active training sets is in the range of 2.4%~4.7%. The percentage of "active" families covered by the sparse active training sets is in the range of 6.7%~42.2% with five of these below 22.5%. Therefore, the level of sparseness of the very sparse active datasets is significantly higher than those of the regularly sparse active datasets.

The SVM VS tools developed by using the very sparse active datasets for identifying active compounds of the six biological target classes were tested by using three testing sets for each compound class. These testing sets are the active testing set for each class, 9.98 million distinct compounds from the PubChem, and the remaining 167K MDDR compounds outside the training sets of our developed SVM models. The performance of these SVM VS tools is given in **Table 4-2**. In spite of the use of very sparse active training sets of <4.7% of the actives covering 6.7%~42.2% of the "active" families, our SVM VS tools were able to achieve yields of 17.5%~35.5% for the high, 23.0%~48.1% for the intermediate, and 70.2%~92.4% for the low diversity classes. Therefore, our method appears to have some level of generalization capability in identifying a substantial amount of novel active compounds

outside the known active chemical families from a very sparse active training dataset.

In addition to the exhibition of effective hit selection performance, our SVM models appear to show substantially lower false-hit rates. In screening 9.997M PubChem compounds that exclude the known actives, without using top-ranked cut-off or additional filter, our SVM VS tools identified 398~2,336 compounds as active, representing 0.004%~0.01% of the 9.997M PubChem compounds. The estimated false-hit rates in screening 167K MDDR compounds of the other biological classes are in the range of 0.5%~1.6%. Even though a substantially larger number of compounds (9.997M vs. 98.4K~2.5M) were screened, these false-hit rates are comparable and in many cases better than those of 0.08%~3% by SBVS tools[53-55,59,65-67,69,168-170], 0.1%~5% by other reported ML models[48,49,62], 0.16%~82.% by clustering methods[29] , and 1.15%~26% by pharmacophore models.

Table 4-2 Dataset statistics and virtual screening performance of support vector machines developed by using very sparse active datasets of 40 active compounds for identifying muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors from PubChem and MDDR databases. Known "active" chemical families refer to chemical families that contain at least one known active compound.

Compound Diversity Category Defined in Ref 48	**Compound Biological Target Class (No of compounds) [average mean pairwise similarity value computed in Ref 48]**	**Active Compounds in Training Set**		**Active Compounds in Testing Set**		**Virtual Screening Performance**			
		No and Percent of Active Compounds	**No and Percent of Known "Active" Chemical Families Covered by Active Compounds**	**No and Percent of Active Compounds**	**No and Percent of Known "Active" Chemical Families Covered by Active Compounds**	**Yields**	**No and Percent of Identified Testing Active Compounds Outside Training Chemical Families**	**No and Percent of 9.997M PubChem Compounds identified as Active**	**No and Percent of the Remaining 167K MDDR Compounds as Active**
High	Muscarinic M1 receptor agonists (983) [0.206]	40 (4.1%)	34 (16.7%)	943 (95.9%)	191 (94.1%)	39.5%	149 (40.1%)	1,130 (0.01%)	1,618 (1.0%)
	NMDA receptor antagonists (1510) [0.199]	40 (2.7%)	36 (6.7%)	1470 (97.3%)	524 (97.4%)	17.5%	165 (64.2%)	2,336 (0.02%)	2,001 (1.2%)
Intermediate	Thrombin inhibitors (1252) [0.321]	40 (3.2%)	25 (15.5%)	1212 (96.8%)	237 (96.0%)	23.0%	102 (57.0%)	529 (0.005%)	1,198 (0.7%)
	HIV protease inhibitors (1054) [0.313]	40 (3.8%)	36 (12.8%)	1014 (96.2%)	269 (95.7%)	48.1%	301 (68.7%)	530 (0.005%)	2,658 (1.6%)
Low	Cephalosporins (1645) [0.501]	40 (2.4%)	27 (42.2%)	1605 (97.6%)	86 (89.7%)	92.4%	205 (13.8%)	770 (0.007%)	791 (0.5%)
	Rennin inhibitors (1241) [0.459]	40 (3.2%)	31 (22.5%)	1201 (96.8%)	130 (94.2%)	70.2%	410 (48.6%)	398 (0.004%)	2,220 (1.3%)

4.3.3 Evaluation of false-hit rates of SVM against inactives of similar molecular descriptors to the known actives

The subsets of MDDR compounds that are similar in molecular descriptors to at least one known active of the six biological target classes were selected by using the condition that the Tanimoto coefficient *sim(i,j)* is ≥ 0.9 with respect to at least one known active of each of these classes. A total of 19,495, 38,436, 32,037, 29,990, 29,127, and 24,166 inactives of similar molecular descriptors were collected for the muscarinic M1 receptor agonist, NMDA receptor antagonist, thrombin inhibitor, HIV protease inhibitor, cephalosporin, and rennin inhibitor classes respectively. Each of these six sets of inactives were used as the testing sets for evaluating the false-hit rates of our developed SVM VS tools against similarity compounds.

As shown in **Table 4-3**, against these similarity datasets, the false-hit rates of our SVM VS tools developed by using regularly sparse and very sparse active datasets are in the range of 4.6%~8.3% and 2.6%~6.3% respectively. Compared to the ranges of hit rates of 1.0%~2.9% and 0.5%~1.6% against the full set of the ~167K MDDR compounds of other biological target classes, our developed SVM VS tools appear to show fairly good performance in distinguishing the actives from the inactives that resemble the physicochemical properties of the known actives.

4.3.4 Evaluation of SVM identified false hits

Some of the false hits are known inhibitors that share structural frameworks with those of the studied biological target class. For instance, a number of

SVM identified "false" hits of HIV protease inhibitors are known rennin inhibitors. Some of the HIV protease inhibitors have been designed based on the transition state analogues of renin inhibitors[203]. Many of the SVM identified false hits of thrombin inhibitors are known peptidomimetic inhibitors of renin, HIV protease, farnesyltransferase, and trypsin. Peptidomimetic inhibitors arising from similar structural frameworks have been designed for renin, thrombin, HIV protease, Ras farnesyltransferase, and various other proteases[204]. Therefore, some of the false hits may partly arise from the mis-identification of compounds of similar structural frameworks. It cannot be ruled out that some of them may exhibit weak inhibitory activities due to the similar structural frameworks and thus were "correctly" identified by our SVM VS tools.

Examination of the false hits identified by SVM and other machine learning methods consistently suggests that the currently-used molecular descriptors are insufficient to adequately represent some of the compounds that contain complex structural or chemical configurations[130,149,181]. Examples of these agents are those with large rigid structure combined with a short flexible hydrophilic tail, compounds that contain multi-rings with various hetero atoms such as nitrogen, oxygen, sulphur, fluorine and chlorine. Due to the limited coverage of the number of bond links in a hetero-atom loop, the currently available topological descriptors are not yet capable of describing the special features of a complex multi-ring structure that contains multiple hetero atoms. It appears that none of the currently-available descriptors are capable of fully representing molecules containing a long flexible chain. Therefore, it might be

helpful to explore different combination of descriptors and to select more optimal set of descriptors by using more refined feature selection algorithms and parameters[130,195]. However, indiscriminate use of many existing topological descriptors, which are overlapping and redundant to each other, may introduce noise as well as extending the coverage of some the aspects of these special features. Thus, it may be necessary to introduce new descriptors for more appropriately representing these and other special features.

Table 4-3 Evaluation of support vector machines virtual screening tools for identifying muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors against the subset of inactive MDDR compounds that are similar to at least one known active compound in each respective active compound class. Similarity is defined by Tanimoto coefficient ≥0.9, which is computed by using molecular descriptors. The yields are given in Table 4-1 and Table 4-2 respectively.

Compound Diversity Category Defined in Ref 48	**Compound Biological Target Class (No of compounds)**	**No and Percent of Active Compounds in Training Set**	**No of Inactive Compounds Similar to an Active Compound (Testing Set)**	**SVM Virtual Screening Performance**	
				No of Inactive Compounds Predicted as Active	**False Hit Rate**
High	Muscarinic M1 receptor agonists (983)	40 (4.1%)	19,495	531	4.4%
		100 (10.2%)		1,068	7.8%
	NMDA receptor antagonists (1510)	40 (2.7%)	38,436	729	2.6%
		100(6.6%)		1,349	4.6%
Intermediate	Thrombin inhibitors (1252)	40 (3.2%)	32,037	1,535	5.7%
		100(8.0%)		1,267	6.4%
	HIV protease inhibitors (1054)	40 (3.8%)	29,990	603	3.3%
		100(9.5%)		1,398	6.4%
Low	Cephalosporins (1645)	40 (2.4%)	29,127	181	5.8%
		100(6.1%)		612	7.6%
	Rennin inhibitors (1241)	40 (3.2%)	24,166	637	6.3%
		100(8.1%)		887	8.3%

4.3.5 Does SVM select active compounds or membership of compound families?

To further evaluate whether our SVM VS tools identify active compounds rather than membership of certain compound families, Compound family distribution of the identified actives and inactives for the six biological target classes were analyzed. As shown in **Table 4-2**, 40.1%, 64.2%, 57.0%, 68.7%, 13.8%, and 48.6% of the identified muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors belong to the families that contain no known active. For those families that contain at least one known active, >70% of the compounds (>90% in majority cases) in each of these families were predicted as inactive by our SVM VS tools. These results suggest that our SVM VS tools identify active compounds rather than membership to certain compound families. Some of the identified actives not in the family of known active compounds may serve as potential "novel" active compounds. Therefore, as in the case shown by earlier studies[108,191], SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

4.4 Further perspective

SVM VS tools developed by using highly sparse active datasets show some level of capability in identifying novel active compounds at comparable and in many cases substantially lower false-hit rates than those of typical SBVS and LBVS tools reported in the literatures. The performance of SVM is significantly better than that of Tanimoto-based similarity search method

based on the same datasets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating lead discovery from sparse active datasets[65,69,71].

Chapter 5 Virtual Screening of Selective Kinase Inhibitors

*High performance virtual screening tools we built in **Chapter 3** can be applied for searching novel ligands for many targets whose ligands are available. The aim of this chapter is to investigate the applicability of our virtual screening method in predicting and searching potential c-Src (Section 5.1) and VEGFR-2 (Section 5.2) selective kinase inhibitors. c-Src and VEGFR-2 are two important kinases that play various roles in tumour progression, invasion, metastasis, angiogenesis and survival. New inhibitors for c-Src and VEGFR-2 are necessary for pharmaceutical research of cancer treatment.*

5.1 Virtual screening of c-Src kinase inhibitors

5.1.1 c-Src, c-Src inhibitors and cancer

Src promotes tumour invasion and metastasis, facilitates VEGF-mediated angiogenesis and survival in endothelial cells, and enhances growth factor driven proliferation in fibroblasts[205]. It is one of the multiple kinase targets of a number of multi-target kinase inhibitors effective in the clinical treatment of leukemia and in clinical trials of other cancers[86,206,207]. The successes and problems of these inhibitors have raised significant interest and efforts in discovering new Src inhibitors[208-210]. Several *in-silico* methods have been used for facilitating the search and design of Src inhibitors, which include pharmacophore[211], QSAR[212], and molecular docking[209].

While these *in-silico* methods have shown impressive capability in the identification of potential Src inhibitors, their applications may be affected by such problems as the vastness and sparse nature of chemical space that needs to be searched, complexity and flexibility of target structures, difficulties in accurately estimating binding affinity and solvation effects, and limited diversity of training active compounds[33,37,168]. It is desirable to explore other *in-silico* methods that complement these methods by expanded coverage of chemical space, increased screening speed, and reduced false-hit rates without necessarily relying on the modelling of target structural flexibility, binding affinity and salvation effects.

In this work, we developed a SVM VS model for identifying Src inhibitors, and evaluated its performance by both 5-fold cross validation test and large compound database screening test. In 5-fold cross validation test, a dataset of Src inhibitors and non-inhibitors was 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. In large database screening test, a SVM VS tool was developed by using Src inhibitors published before 2008, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using Src inhibitors reported since 2008 and not included in the training datasets, virtual-hit rate and false-hit rate in searching large libraries were evaluated by using 13.56M PubChem, 168K MDDR, and 9,305 MDDR compounds similar in structural and physicochemical properties to the known Src inhibitors.

PubChem and MDDR contain high percentages of inactive compounds significantly different from the known Src inhibitors, and the easily distinguishable features may make VS enrichments artificially good[202]. Therefore, VS performance may be more strictly tested by using subsets of compounds that resemble the physicochemical properties of the known Src inhibitors so that enrichment is not simply a separation of trivial physicochemical features[158]. To further evaluate whether our SVM VS tool predict Src inhibitors and non-inhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

5.1.2 Virtual screening model development

5.1.2.1 Compound collections and construction of training and testing datasets

We collected 1,020 Src inhibitors, with IC50<50μM, from the literatures[213-217] and the BindingDB database[110]. Our collected Src inhibitors are distributed in 288 families. The inhibitor selection criterion of IC50<50μM was used because it covers most of the reported HTS and VS hits[218]. The structures of representative Src inhibitors are shown in **Figure 5-1**. As few non-inhibitors have been reported, putative non-inhibitors were generated by using our method for generating putative inactive compounds[108,219] (please refer to **Chapter 3 Section 3.2.2**).

In the database screening test, 60.1% of families that contain Src inhibitors reported since 2008 are not covered by the Src inhibitor training dataset

(inhibitors reported before 2008), and the representative compounds of these families were deliberately placed into the inactive training sets as these inhibitors are not supposed to be known in our study. As shown in earlier studies[108,219] and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing "non-inhibitor" families were predicted as inhibitors by our SVM VS tool. Moreover, a small percentage of the compounds in these putative non-inhibitor datasets are expected to be un-reported and un-discovered inhibitors, their presence in these datasets is not expected to significantly affect the estimated false hit rate of SVM.

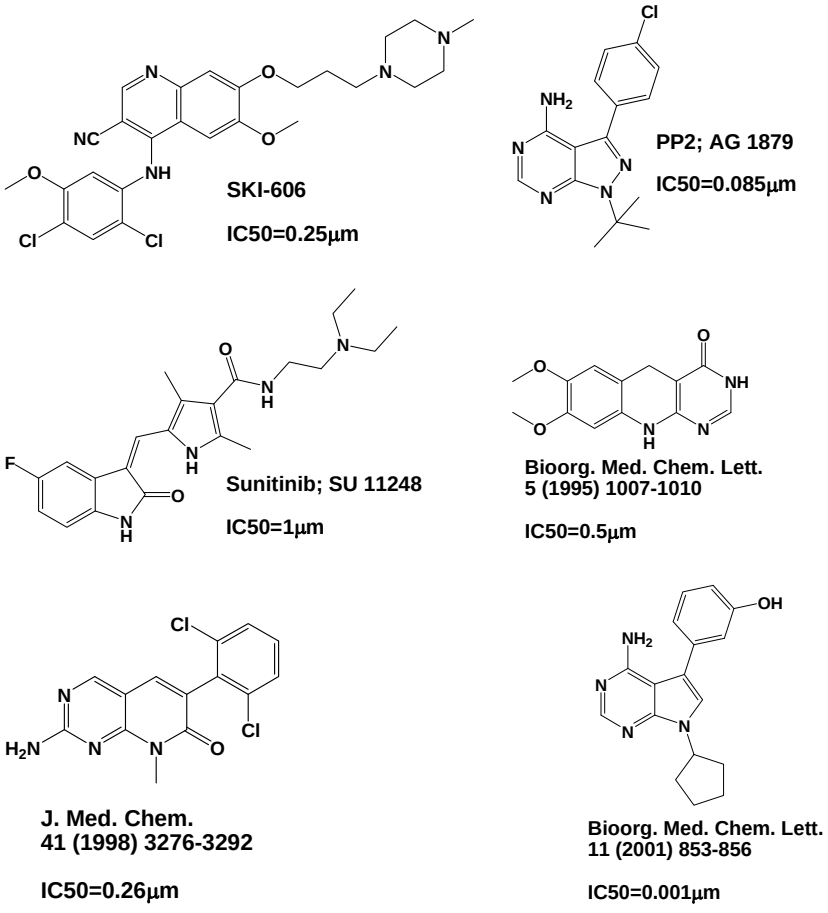


Figure 5-1 The structures of representative c-Src inhibitors

5.1.2.2 Molecular descriptors and computational model

A total of 98 important descriptors were chosen from the chemical descriptors calculated by our program MODEL which were used in this work. The details of molecular descriptors are explained in **Chapter 2 Section 2.2**.

Computational model for virtual screening is developed by using SVM.

5.1.3 Results and Discussion

5.1.3.1 Performance of SVM identification of Src inhibitors based on 5-fold cross validation test

Table 5-1 shows the 5-fold cross validation test results of SVM identification of Src inhibitors and putative non-inhibitors. The inhibitor and non-inhibitor prediction accuracies are 87.8%~93.1% and 99.75%~99.88% respectively. The overall prediction accuracy *Q* and Matthews correlation coefficient *C* are 99.61%~99.77% and 0.759~0.857 respectively. The inhibitor accuracies of our SVM are comparable to or slightly better than the reported accuracies of 58.3%~67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods[36], 83% for Lck inhibitors by SVM method[218], and 74%~87% for inhibitors of any of the 8 kinases (3 Ser/Thr and 5 Tyr kinases) by SVM, ANN, GA/kNN, and RP methods[94]. The non-inhibitor accuracies are comparable to the value of 99.9% for Lck inhibitors[218] and substantially better than the typical values of 77%~96% of other studies[36,94]. Caution needs to be exercised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good capability in identifying Src inhibitors at low false-hit rates.

Similar prediction accuracies were also found from two additional 5-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters.

Table 5-1 Performance of support vector machines for identifying Src inhibitors and non-inhibitors evaluated by 5-fold cross validation study.

Cross – Validation	Src inhibitors				Src non-inhibitors				Q (%)	C
	No of training/ testing inhibitors	TP	FN	SE(%)	No of training/ testing non-inhibitors	TN	FP	SP(%)		
1	816/204	189	15	92.65%	51966/12992	12959	33	99.75%	99.64%	0.786
2	816/204	184	20	90.20%	51966/12992	12975	17	99.87%	99.72%	0.823
3	816/204	179	25	87.75 %	51966/12992	12965	27	99.79%	99.61%	0.759
4	816/204	190	14	93.14%	51967/12991	12959	32	99.75%	99.65%	0.794
5	816/204	190	14	93.14%	51967/12991	12975	16	99.88%	99.77%	0.857
average				**91.47%**				**99.81%**	**99.68%**	**0.804**
SD				0.0212				0.000557	0.000605	0.0336
SE				0.0095				0.00025	0.00027	0.0150

5.1.3.2 Virtual screening performance of SVM in searching Src inhibitors from large compound libraries

As outlined in the methods section, we developed a SVM VS tool for searching Src inhibitors from large were developed by using Src kinases reported before 2008. The VS performance of SVM in identifying Src inhibitors reported since 2008 and in searching MDDR and PubChem databases is summarised in **Table 5-2**. The yield in searching Src inhibitors reported since 2008 is 66.2%, which is comparable to the reported 50%~94%

yields of various VS tools[220]. Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies.

Table 5-2 Virtual screening performance of support vector machines for identifying Src inhibitors from large compound libraries

Inhibitors in Training Set	Number of Inhibitors	1020
	Number of Chemical Families Covered by Inhibitors	288
Inhibitors in Testing Set	Number of Inhibitors	133
	Number of Chemical Families Covered by Inhibitors	65
	Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set	39.9%
Virtual Screening Performance	Yield	66.2%
	Number and Percent of Identified True Inhibitors Outside Training Chemical Families	43 (32.3%)
	Number and Percent of 13.56M PubChemCompounds Identified as Inhibitors	44,843 (0.33%)
	Number and Percent of the 168K MDDR Compounds Identified as Inhibitors	1,496 (0.89%)
	Number and Percent of the 9,305 MDDR Compounds Similar to the Known Inhibitors Identified as Inhibitors	719 (7.73%)

We also evaluated virtual-hit rates and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the known Src inhibitors by using 9,305 MDDR compounds similar to an Src inhibitor in the training dataset. Similarity was defined by Tanimoto similarity coefficient ≥0.9 between a MDDR compound and its closest inhibitor[219]. SVM

identified 719 virtual-hits from these 9,305 MDDR similarity compounds (virtual-hit rate 7.73%), which suggests that SVM has some level of capability in distinguishing Src inhibitors from non-inhibitor similarity compounds. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 1,496 and 0.89% respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 44,843 and 0.33% respectively.

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, signal transduction inhibitors, antiangiogenic, and antiarthritic (**Table 5-3**, details in next section). As some of these virtual-hits may be true Src inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rate is <7.73% in screening 9,305 MDDR similarity compounds, <0.89% in screening 168K MDDR compounds, and <0.33% in screening 13.56M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of 0.0054%~8.3% of SVM[89,219], 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[220].

5.1.3.3 Evaluation of SVM identified MDDR virtual-hits

SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 5-3** gives the MDDR classes that contain higher percentage (⩾3%) of SVM virtual-hits and the percentage values. We found that 623 (41.6%) of the 1,496 virtual-hits belong to the antineoplastic class, which represent 2.9% of the 21,557 MDDR compounds in the class. In particular, 231 (15.4%) of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 19.6% of the 1,181 MDDR compounds in the class. Moreover, 194 (13.0%) and 75 (5.0%) of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 9.5% and 4.6% of the 2,037 and 1,629 members in these classes respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction and angiogensis pathways. While some of these kinase inhibitors might be true Src inhibitors, a significant percentage of them are expected to arise from false selection of inhibitors of other kinases.

Table 5-3 MDDR classes that contain higher percentage (≥3%) of virtual-hits identified by SVMs in screening 168K MDDR compounds for Src inhibitors. The total number of SVM identified virtual hits is 1,496.

MDDR Classes that Contain Higher Percentage (≥3%) of Virtual Hits	**No of Virtual Hits in Class**	**Percentage of Class Members Selected as Virtual Hits**
Antineoplastic	623 (41.6%)	2.9%
Tyrosine-Specific Protein Kinase Inhibitor	231 (15.4%)	19.6%
Signal Transduction Inhibitor	194 (13.0%)	9.5%
Antiarthritic	176 (11.8%)	1.5%
Antiallergic/Antiasthmatic	83 (5.5%)	0.8%
Antihypertensive	76 (5.1%)	0.7%
Antiangiogenic	75 (5.0%)	4.6%
Treatment for Osteoporosis	55 (3.68%)	2.2%
Antidepressant	49 (3.27%)	0.8%

176 (11.8%) of the SVM virtual-hits belong to the antiarthritic class. A primary feature of rheumatoid arthritis in synovial tissues is the abnormal stimulation of fibrin deposition, angiogenesis and proinflammatory processes, which are promoted by thrombin increased IL-6 production via the PAR1 receptor/PI-PLC/PKC alpha/c-Src/NF-kappaB and p300 signaling pathways[221]. Therefore, Src inhibitors may have some effects against arthritis via interference with some of these processes. Moreover, several other kinases have been implicated in arthritis. An Abl inhibitor Gleevec has been reported to be effective in treatment of arthritis, which is probably due to its inhibition

of other related kinases such as c-kit and PDGFR[222]. EGFR-like receptor stimulates synovial cells and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis[89]. VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis[223]. FGFR may partly mediates osteoarthritis[224]. PDGF-like factors stimulates the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells[225]. Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma[226]. Therefore, some of the SVM virtual-hits in the antiarthritic class may be inhibitors of these kinases or their kinase-likes capable of producing antiarthritic activities.

Moreover, 83 (5.5%), 76 (5.1%), 55 (3.7%) and 49 (3.3%) of the SVM virtual hits are in the antiallergic/antiasthmatic, antihypertensive, osteoporosis treatment and antidepressant classes respectively. Src or Src family kinases have been implicated in and the respective inhibitors have shown observable effects against these diseases. For instance, Src family kinases and lipid mediators have been found to partly control allergic inflammation[227]. Inhibition of Src family kinase-dependent signaling cascades in mast cells may exert anti-allergic activity[228]. Up-regulation of Src signaling has been suggested to be important in the profibrotic and proinflammatory actions of aldosterone in a genetic model of hypertension, which can be significantly reduced by mineralocorticoid receptor blocker and Src inhibitor[229]. Src signalling pathways play critical roles in osteoclasts and osteoblasts, and Src inhibitors have been developed as therapeutic agents for bone diseases[230,231].

Src-family protein tyrosine kinases negatively regulate cerebellar long-term depression, which can be recovered by the application of Src-family protein tyrosine kinase inhibitors[232]. Therefore, some of the SVM virtual hits in these four MDDR classes may be Src inhibitors or Src family kinase inhibitors capable of regulating allergic inflammation, hypertension, osteoporosis and depression respectively.

5.1.3.4 Comparison of Virtual Screening Performance of SVM with Tanimoto-Based Similarity Searching Method

To evaluate whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of the Tanimoto-based similarity searching method (please refer to **Chapter 2 Section 2.3.4**) based on the same molecular descriptors, training dataset of Src inhibitors reported before 2008, and the testing dataset of Src inhibitors reported since 2008 and 168K MDDR compounds. The yield and maximum possible false-hit rate of the Tanimoto-based similarity searching method is 36.84% and 5.54% respectively. Compared to these results, the yield of SVM is smaller than but still comparable to that of the Tanimoto-based similarity searching method, and the false-hit rate of SVM is significantly reduced by ~10 fold. This suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used, and SVM is capable of achieving comparable yield at significantly reduced false-hit rate as compared to Tanimoto similarity-based approach.

5.1.3.5 Does SVM select Src inhibitors or membership of compound families?

To further evaluate whether SVM identifies Src inhibitors rather than membership of certain compound families, Compound family distribution of the identified Src inhibitors and non-inhibitors were analyzed. 48.9% of the identified inhibitors belong to the families that contain no known Src inhibitors. For those families that contain at least one known Src inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-inhibitor by SVM. These results suggest that SVM identify Src inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential "novel" Src inhibitors. Therefore, as in the case shown by earlier studies[108], SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

5.1.4 Further perspective

Our study suggested that SVM is capable of identifying Src inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures. It can be used for searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates without the need to define an applicability domain, i.e. it has a broad applicability domain that covers the whole chemical space defined by the current versions of PubChem and MDDR databases. The performance of SVM is substantially improved against Tanimoto-based similarity search method based on the same datasets

and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Because of its high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of Src inhibitors and other active compounds[65,69,71].

5.2 Virtual screening of VEGFR-2 kinase inhibitors

5.2.1 VEGFR, VEGFR inhibitors and cancer

VEGFR regulates angiogenesis, growth, migration and survival[233]. There are 3 main VEGFR subtypes, VEGFR-2 mediates almost all of the known cellular responses to VEGF, VEGFR-1 modulates VEGFR-2 signaling and acts as a dummy/decoy receptor, and VEGFR-3 mediates lymphangiogenesis in response to VEGF-C and VEGF-D[233]. VEGFR inhibitors have been successfully used for cancer treatments[86,234]. While increasing number of VEGFR inhibitors have been developed and tested, several problems limit the scope of their practical applications. These problems include increased toxicity partly due to the targeting of multiple kinases, acquired resistances, and reduced tumor responses (VEGFR inhibitors can cause extensive tumor necrosis without a marked decrease in tumor size) [235]. Moreover, on-target toxicity against specific VEGFR subtypes in various tissues is also a significant problem for the applications of VEGFR inhibitors[236]. The successes of VEGFR inhibitors and the encountered problems have led to further efforts for discovering new inhibitors[86,234].

In-silico methods such as pharmacophore[237], QSAR[63,238,239], fragment-based method[240], molecular docking[241,242], and their combinations[237,239] have been used for facilitating the search and design of VEGFR inhibitors, which have shown impressive capability in the identification of potential VEGFR inhibitors. In this work, SVM was tested for its capability in searching VEGFR-2 inhibitors from large compound libraries. Our focus on inhibitors of VEGFR-2 subtype was based on the availability of reported inhibitors of the subtype and the consideration that VEGFR-2 mediates almost all of the known cellular responses to VEGF[233]. The performance of SVM was evaluated by both 5-fold cross validation test and large database screening test. In 5-fold cross validation test, VEGFR-2 inhibitors and non-inhibitors was 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. In large database screening test, SVM was developed by using VEGFR-2 inhibitors published before 2008, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using VEGFR-2 inhibitors reported since 2008 and not included in the training datasets, virtual-hit rate and false-hit rate of the SVM in searching large libraries were evaluated by using 13.56M PubChem, 168K MDDR, and 13,872 MDDR compounds similar in structural and physicochemical properties to the known VEGFR-2 inhibitors.

Databases such as PubChem and MDDR contain high percentages of inactive compounds significantly different from VEGFR-2 inhibitors, and the easily

distinguishable features may make VS enrichments artificially good[202]. Therefore, VS performance may be more strictly tested by using subsets of compounds that resemble the physicochemical properties of the known VEGFR-2 inhibitors so that enrichment is not simply a separation of trivial physicochemical features[158]. To further evaluate whether SVM predict VEGFR-2 inhibitors and non-inhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed. Moreover, VS performance of SVM for screening MDDR compounds was compared with that of Tanimoto similarity search method on the same molecular descriptors, training dataset to determine whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used.

5.2.2 Virtual screening model development

5.2.2.1 Compound collection, training and testing datasets, molecular descriptors

Using the inhibitor selection criterion of IC50<10μM, which covers most of the reported HTS and VS hits[243,244], we collected 1,293 VEGFR-2 inhibitors regardless of their activities against other VEGFR subtypes from the literatures[245-255] and the BindingDB database[110]. The structures of representative VEGFR-2 inhibitors are shown in **Figure 5-2**. Our collected VEGFR-2 inhibitors are distributed in 433 families. As few non-inhibitors have been reported, putative non-inhibitors were generated by using our method for generating putative inactive compounds[108,219] (please refer to **Chapter 3 Section 3.2.2**).

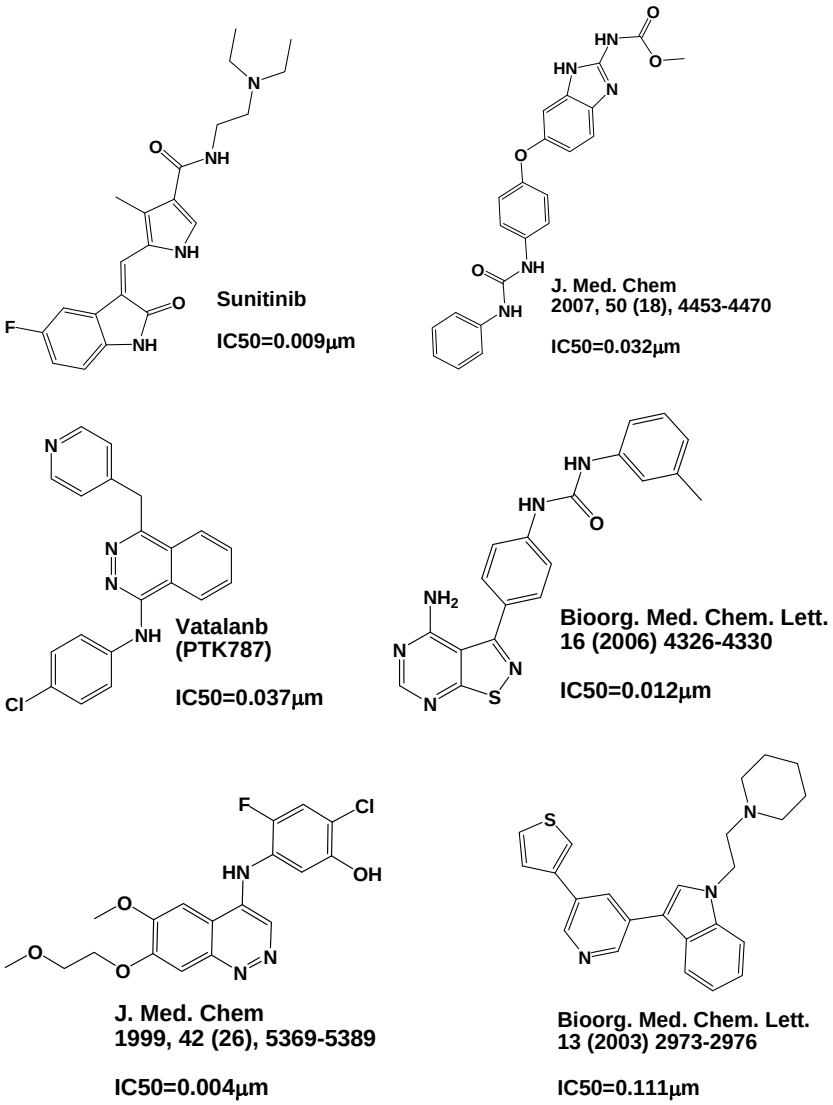


Figure 5-2 The structures of representative VEGFR-2 inhibitors

In conducting large database screening test, 1293 VEGFR-2 inhibitors reported before 2008 were used as a training dataset for developing SVM and 372 VEGFR-2 inhibitors reported since 2008 were used as an independent testing dataset for testing SVM. Only 27.6% of the families that contain VEGFR-2 inhibitors reported since 2008 are covered in the families that contain at least one VEGFR-2 inhibitor reported before 2008, and the representative compounds of these families were deliberately placed into the

inactive training sets as these inhibitors are not supposed to be known in our study. As shown in earlier studies[256] and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing "non-inhibitor" families were predicted as inhibitors by SVM. Moreover, a small percentage of the compounds in these putative non-inhibitor datasets are expected to be un-reported and un-discovered inhibitors, their presence in these datasets is not expected to significantly affect the estimated false hit rate of SVM.

5.2.2.2 Molecular Descriptors and computational model

A total of 98 important descriptors were chosen from the chemical descriptors calculated by our program MODEL which were used in this work. The detail about molecular descriptors is explained in **Chapter 2 Section 2.2**. Computational model for virtual screening is developed by using SVM.

5.2.3 Results and Discussion

5.2.3.1 VEGFR-2 Inhibitor prediction Performance of SVM evaluated by 5-fold cross validation test

Table 5-4 gives the 5-fold cross validation test results of SVM in identifying VEGFR-2 inhibitors and non-inhibitors. The accuracies for predicting inhibitors and non-inhibitors are 86.0%~90.0% and 99.62%~99.73% respectively. The overall prediction accuracy Q and Matthews correlation coefficient C are 99.40%~99.47% and 0.7236~0.7548 respectively. The inhibitor accuracies of our SVM are comparable to or better than the reported accuracies of 58.3%~67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods[36], 83% for Lck inhibitors by SVM method[218], and 74%~87%

for inhibitors of any of the 8 kinases (3 Ser/Thr and 5 Tyr kinases) by SVM, ANN, GA/kNN, and RP methods[94]. The non-inhibitor accuracies are comparable to the value of 99.9% for Lck inhibitors[218] and substantially better than the typical values of 77%~96% of other studies[36,94]. These are consistent with the result of a study of the comparison of SVM with 16 classification methods and 9 regression methods, which has shown that SVMs showed mostly good performances both on classification and regression tasks but other methods proved to be very competitive[256]. Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the datasets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good prediction capability in identifying VEGFR-2 inhibitors at low false-hit rates. Similar prediction accuracies are also found from two additional 5-fold cross validation studies conducted by using training-testing sets separately generated from different random number seed parameters.

Table 5-4 Performance of support vector machines for identifying VEGFR-2 inhibitors and non-inhibitors evaluated by 5-fold cross validation study.

Cross - Validation	VEGFR-2 inhibitors				VEGFR-2 non-inhibitors				Q (%)	C
	No of training/ testing inhibitors	TP	FN	SE(%)	No of training/ testing non-inhibitors	TN	FP	SP(%)		
1	1034/259	227	32	87.64%	51038/12760	12714	46	99.64%	99.40%	0.7236
2	1034/259	231	28	89.19%	51038/12760	12712	48	99.62%	99.42%	0.7334
3	1034/259	233	26	89.96 %	51038/12759	12715	43	99.66%	99.47%	0.7548
4	1035/258	229	41	88.76%	51039/12759	12718	41	99.68%	99.46%	0.7481
5	1035/258	222	36	86.05%	51039/12759	12725	34	99.73%	99.46%	0.7415
Average				**88.32%**				**99.67%**	**99.44%**	**0.7403**
SD				0.0152				0.000422	0.000303	0.0122
SE				0.0068				0.000189	0.000136	0.0055

5.2.3.2 Virtual screening performance of SVM in searching VEGFR-2 inhibitors from large compound libraries

A SVM in searching VEGFR-2 inhibitors from large libraries was developed by using VEGFR-2 inhibitors reported before 2008. The VS performance of this SVM in identifying VEGFR-2 inhibitors reported since 2008 and in searching MDDR and PubChem databases is summarised in **Table 5-5**. The yield in searching VEGFR-2 inhibitors reported since 2008 is 57.3%, which is comparable to the reported 50%~94% yields of various VS tools[220]. Strictly

speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition and diversity of compounds screened, and in the molecular descriptors, VS tools and their parameters used. The comparison cannot go beyond the statistics of accuracies as the reports are not detailed enough to address questions of whether all methods detect the same hit.

Table 5-5 Virtual screening performance of support vector machines for identifying VEGFR-2 inhibitors from large compound libraries.

Inhibitors in Training Dataset	No of Inhibitors	1293
	No of Chemical Families Covered by Inhibitors	433
Inhibitors in Testing Dataset	No of Inhibitors	372
	No of Chemical Families Covered by Inhibitors	152
	Percent of Inhibitors in Chemical Families Covered by Inhibitors in Training Set	27.63%
Virtual Screening Performance	Yield	57.26%
	No and Percent of Identified True Inhibitors Outside Training Chemical Families	114 (53.5%)
	No and Percent of 13.56M PubChemCompounds Identified as Inhibitors	89,572 (0.66%)
	No and Percent of the 168K MDDR Compounds Identified as Inhibitors	2,717 (1.62%)
	No and Percent of the 13,872 MDDR Compounds Similar to the Known Inhibitors Identified as Inhibitors	1,714 (12.36%)

Virtual-hit rates and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the VEGFR-2 inhibitors were evaluated by using 13,872 MDDR compounds similar to a VEGFR-2 inhibitor in the training dataset. Similarity was defined by

Tanimoto similarity coefficient ≥0.9 between a MDDR compound and its closest dual-inhibitor[124]. SVM identified 1,714 virtual-hits from these 13,872 MDDR similarity compounds (virtual-hit rate 12.4%), which suggests that SVM has some level of capability in distinguishing VEGFR-2 inhibitors from similarity non-inhibitors. Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168K MDDR and 13.56M PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 2,717 and 1.62% respectively. The numbers of virtual-hits and virtual-hit rates in screening 3.56M PubChem compounds are 89,572 and 0.66% respectively.

Many of the 2,717 MDDR virtual-hits belong to the classes of antineoplastic (45.3%), tyrosine-specific protein kinase inhibitor (12.7%), signal transduction inhibitor (12.7%), antiarthritic (11.0%), and antiangiogenic (9.3%), antihypertensive (5.1%), antiallergic/antiasthmatic (4.3%), and antidepressant (3.4%) (**Table 5-6**, details in next section). As some of these virtual-hits may be true VEGFR inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence the false-hit rate is ≤12.36% in screening 13,872 MDDR similarity compounds, ≤1.62% in screening 168K MDDR compounds, and ≤0.66% in screening 13.56M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of 0.0054%~8.3% of SVM[257], 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[258].

Table 5-6 MDDR classes that contain higher percentage (≥3%) of virtual-hits identified by SVMs in screening 168K MDDR compounds for VEGFR-2 inhibitors. The total number of SVM identified virtual hits is 2,717.

MDDR Classes that Contain Higher Percentage (>3%) of Virtual Hits	**No and Percentage of Virtual Hits in Class**	**Percentage of Class Members Selected as Virtual Hits**
Antineoplastic	1230 (45.3%)	5.7%
Tyrosine-Specific Protein Kinase Inhibitor	346 (12.7%)	29.3%
Signal Transduction Inhibitor	345 (12.7%)	16.9%
Antiarthritic	300 (11.0%)	2.6%
Antiangiogenic	256 (9.3%)	15.7%
Antihypertensive	139 (5.1%)	1.3%
Antiallergic/Antiasthmatic	118 (4.3%)	1.1%
Antidepressant	93 (3.4%)	1.5%

5.2.3.3 Evaluation of SVM identified MDDR virtual-hits

SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. Table 4 gives the MDDR classes that contain higher percentage (⩾3%) of SVM virtual-hits and the percentage values. We found that 1,230 or 45.3% of the 2,717 virtual-hits belong to the antineoplastic class, which represent 5.7% of the 21,557 MDDR compounds in the class. In particular, 346 or 12.7% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 29.3% of the 1,181 MDDR compounds in the class. Moreover, 12.7% and 9.4% of

the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 16.9% and 15.7% of the 2,037 and 1,629 members in the two classes respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis and other cancer-related pathways. Some of these SVM selected kinase inhibitors might have VEGFR inhibitory activities, and others were expectedly selected due to false selection of inhibitors of other kinases (at ≤1.62%~12.36% false-hit rates).

Substantial percentages of the SVM virtual-hits belong to the antiarthritic (11.0%), antihypertensive (5.1%), and antiallergic/antiasthmatic (4.3%) therapeutic classes. Some VEGFR inhibitors have been reported to show respective therapeutic effects. VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis[223]. Both VEGFR-1 and VEGFR-2 are expressed in human osteoarthritic cartilage[259]. VEGFR-2 and VEGFR-3 are present in most of the sublining blood vessels in arthritic synovium[260]. A VEGFR-2 inhibitor, PTK787/ZK222584, has been reported to cause significant anti-arthritic effects in models of rheumatoid arthritis via anti-angiogenic actions[124]. Hypertension is characterized by the development of a hyperdynamic circulation which can be markedly inhibited by EGFR-2 inhibitor (e.g. SU5416) blockade of the VEGF signaling pathway, leading to the consideration of modulation of angiogenesis for the treatment of hypertension[257]. VEGFR-2 and VEGFR-1 have been shown to be involved in the pathogenesis of the contact hypersensitivity reaction, and both the

induction and elicitation phases of contact hypersensitivity can be inhibited by VEGFR inhibitor PTK787/ZK222584[258]. Therefore, some of the SVM virtual-hits in the antiarthritic, antihypertensive, and antiallergic/antiasthmatic classes may be VEGFR inhibitors capable of producing the respective therapeutic effects.

Moreover, 93 (3.4%) of the SVM virtual hits are in the antidepressant class. It has been reported that depressive episodes in the context of borderline personality disorder may be accompanied by increased serum concentrations of VEGF and FGF-2[261]. VEGF has been implicated in neuronal survival, neuroprotection, regeneration, growth, differentiation, and axonal outgrowth, which is involved in the pathophysiology of major depressive disorder and the higher expression levels of VEGF in the peripheral leukocytes are associated with the depressive state[262]. Therefore, there is a possibility that inhibition of VEGFR signalling may have some level of antidepressant effect or act as enhancer of other antidepressant agents[263], and some of the SVM virtual hits in the antidepressant class may be possible VEGFR inhibitors that partly explain their antidepressant activities.

5.2.3.4 Comparison of Virtual Screening Performance of SVM with Tanimoto-Based Similarity Searching Method

To evaluate whether the performance of SVM is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of the Tanimoto-based similarity searching method (please refer to **Chapter 2 Section 2.3.4**) based on the same molecular descriptors, training

dataset of VEGFR-2 inhibitors reported before 2008, and the testing dataset of VEGFR-2 inhibitors reported since 2008 and 168K MDDR compounds. The yield and false-hit rate of the Tanimoto-based similarity searching method is 39.3% and 4.4% respectively. Compared to these results, the yield of SVM is significantly improved and the false-hit rate of SVM is substantially reduced. This suggests that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used.

5.2.3.5 Does SVM select VEGFR inhibitors or membership of compound families?

To further evaluate whether SVM identifies VEGFR-2 inhibitors rather than membership of certain compound families, Compound family distribution of the identified VEGFR-2 inhibitors and non-inhibitors were analyzed. A total of 53.5% of the identified VEGFR-2 inhibitors belong to the families that contain no known VEGFR-2 inhibitors. For those families that contain at least one known inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as non-inhibitor by SVM. These results suggest that SVM identifies VEGFR-2 inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential "novel" VEGFR-2 inhibitors. Therefore, as in the case shown by earlier studies[108], SVM has certain capacity for identifying novel active compounds from sparse as well as regular-sized active datasets.

5.2.4 Further perspective

By using training dataset of more diverse spectrum of inactive compounds as well as substantial number of literature-reported VEGFR-2 inhibitors, SVM shows substantial capability in identifying VEGFR-2 inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures. It is capable of searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates without the need to define an applicability domain, i.e. it has a broad applicability domain that covers the whole chemical space defined by the PubChem and MDDR databases. The performance of SVM is significantly better than that of Tanimoto-based similarity search method based on the same datasets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than the molecular descriptors used. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of VEGFR inhibitors and other active compounds[65,69,71]. It is also possible to discover dual kinase inhibitor of c-Src and VEGFR based on our developed models in our further study.

Chapter 6 Virtual Screening of Selective Multi-Target Kinase Inhibitors

Multi-target agents have been increasingly explored for enhancing efficacy and reducing counter-target activities and toxicities. Efficient virtual screening (VS) tools for searching selective multi-target agents are desired. In this chapter, combinatorial support vector machines (C-SVMs) were tested as VS tools for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). This is another application of our high performance virtual screening tool in drug discovery.

6.1 Introduction

Large percentage of drugs in development, which are typically directed at an individual target, frequently show reduced efficacies and undesired safety and resistance profiles due to network robustness[76], redundancy[77], crosstalk[78], compensatory and neutralizing actions[79], anti-target and counter-target activities[80], and on-target and off-target toxicities[81]. Multi-target agents and drug-combinations have been increasingly explored[76,82] for enhancing therapeutic efficacies and improving safety and resistance profiles by selectively modulating the elements of these counter-target and toxicity activities[83]. In particular, multi-target kinase inhibitors are among the most successful clinical anticancer drugs (e.g. sunitinib against PDGFR and VEGFR, dasatinib against Abl and Src, sorafenib against Braf and VEGFR, and lapatinib against EGFR and HER2) and have been actively pursued in

current drug discovery efforts[85,86]. Methods for efficient search of multi-target agents are highly desired.

Virtual screening (VS) methods have been widely explored for facilitating lead discovery against individual targets[37,89,219]. In particular, molecular docking[91], pharmacophore[92], QSAR[93], machine learning[94], and combination methods[95] have been extensively used for VS of single-target kinase inhibitors, but few multi-target VS studies have been reported[264,265]. An interesting strategy for identifying multi-target kinase inhibitors is to use experimentally obtained small-scale profiles for predicting inhibitors of a larger kinase set[265]. In principle, single-target VS tools may be combined to collectively identify multi-target agents, which is practically useful if the individual VS tools have sufficiently high yields and low false-hit rates. High yields 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 are needed for high enrichment factors in searching multi-target agents that are significantly fewer in numbers and more sparsely distributed in the chemical space than non-dual inhibitors (**Table 6-1**).

An extensively-used machine learning method, support vector machines (SVM), may be potentially explored as multi-target VS tools because it has shown high yields and low false-hit rates in searching single-target agents[108] sometimes based on sparsely distributed active compounds[219]. SVM identifies

active compounds in fast-speed by differentiating physicochemical profiles rather than structural similarity to active compounds *per se*, and requires no knowledge of target structure and no computation of structural flexibility, activity-related features, solvation effects and binding affinities. Multi-target VS performance of combinatorial SVMs (C-SMV), which combine the prediction of two separate SVM classifier for each the multiple kinases, was tested by using them to search dual-inhibitors of combinations of 9 anticancer kinase targets EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3. **Figure 6-1** shows the illustration of using combinatorial support vector machines method (C-SVM) for searching multi-target inhibitors. These kinase targets were selected because of their therapeutic relevance and the availability of sufficient number of the known inhibitors and dual-inhibitors. The first six kinases belong to the protein kinase group PTK group and the last three belong to the CMGC group respectively.

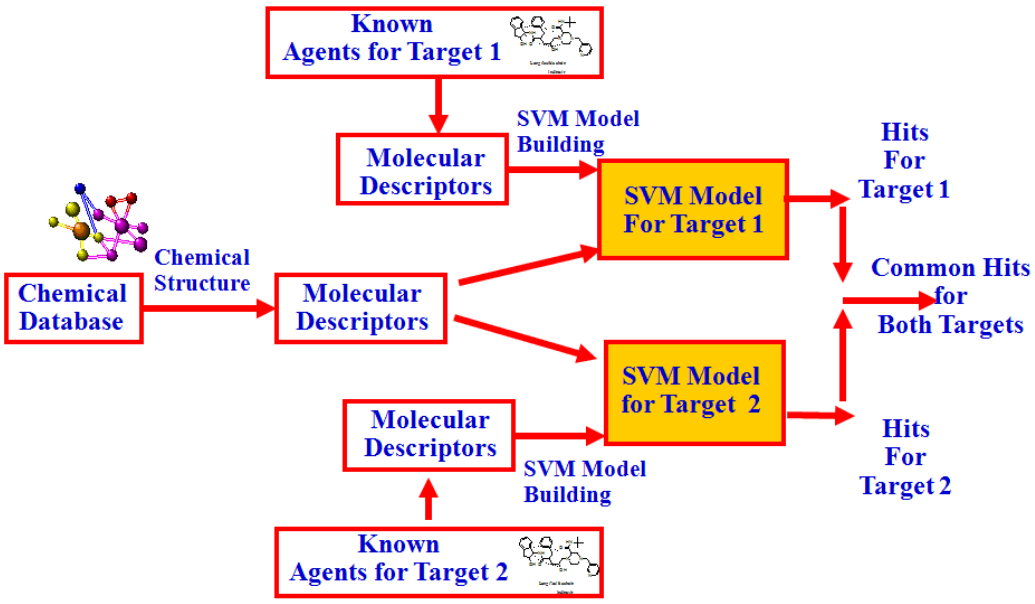


Figure 6-1 Illustration of using combinatorial support vector machines method (C-SVM) for searching multi-target inhibitors for searching multi-target inhibitors.

Based on dual-inhibitor availability, we focused on 11 kinase-pairs EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, Src-Lck, CDK1-CDK2, CDK1-GSK3, CDK2-GSK3, and CDK1-VEGFR. The first 7 kinase-pairs are intra-PTK group, the 8th to 10th and intra-CMGC group, and the 11th are inter-PTK-CMGC group kinase-pairs respectively, representative of different types of kinase-pairs. These kinase-pairs are frequently co-expressed or co-activated in various cancers[266,267], and targeted by multi-target agents[85,86] with good anticancer efficacies. Inhibitors of growth factor receptor tyrosine kinases EGFR, VEGFR, PDGFR and FGFR have been successfully used for cancer treatments[86,234,268-271]. EGFR promotes proliferation and survival[268]. VEGFR regulates angiogenesis and survival[234]. PDGFR modulates angiogenesis and growth, and is one of the multi-targets of several approved and clinical trial drugs[86,270]. FGFR regulates angiogenesis and cancer progression, and is one of the multi-targets of several clinical trial drugs[86,271]. Src-family kinases Src and Lck modulate multiple pathways of cell growth, differentiation, migration and survival, and are part of the multi-targets of several marketed and clinical trial drugs[86,272]. CDKs promote cell cycle progression, their inhibition severely limits the aberrant cell-cycle process in tumor and induces apoptosis, and CDK inhibitors are being developed and tested in clinical trials for anticancer therapeutics[273]. GSK3 modulates glucose metabolism and the function of various proteins, and is associated with neurodegenerative diseases, stroke, bipolar disorder, diabetes and cancer[274]. GSK3 inhibitors have started to reach clinical development for the treatment of various disorders[274].

Multi-target VS performance was tested by a rigorous method that assumes no explicit knowledge of known multi-target agents, because the number of known multi-target agents are generally small for many target-pairs. SVM of each kinase was trained by using non-dual inhibitors of that kinase. The collective yield of C-SVM of each kinase-pair (percent of known dual-inhibitors identified as dual-inhibitors) was estimated by using known dual-inhibitors of each kinase-pair. Target selectivity of each C-SVM was assessed by using non-dual inhibitors of the kinas-pair and inhibitors of the other 7 kinases, out of the 9 evaluated kinases, not included in the kinase-pair. Virtual-hit rates and false-hit rates in searching large compound libraries were evaluated by using 13.56 million PubChem, 168 thousand compounds from the MDL Drug Data Report (MDDR) database, and 276-3,806 MDDR compounds similar in structural and physicochemical properties to the known dual-kinase inhibitors. MDDR contains biologically relevant compounds (active against individual molecular target or biological assay) and well-defined derivatives reported in the patent literature, journals, meetings and congresses. PubChem and MDDR contain high percentages of inactive or active compounds significantly different from the dual-inhibitors, and the easily distinguishable features may make VS enrichments artificially good[202]. Therefore, VS performance is more strictly tested by using subset of MDDR compounds similar to the dual-inhibitors so that enrichment is not simply a separation of trivial physicochemical features[158].

VS performance of C-SVM was further compared with those of three VS methods, which include a popular molecular docking software DOCK version 3.5.54 at the DOCK Blaster server[275], a similarity-based statistical learning method k nearest neighbour (kNN) [73], and a machine-learning method probabilistic neural networks (PNN) [276] against the same sets of dual- and non-dual kinase inhibitors and 1.02 million Zinc clean-leads dataset (Zinc-CLD) [111]. The specific indicators to be compared are the dual-inhibitor yields for both intra-group and inter-group dual-kinase inhibitors, and the false-hit rates for non-dual kinase inhibitors and the Zinc-CLD dataset, which enable objective assessment of the capability of C-SVM with respect to those of the popular as well as machine learning based VS methods.

6.2 Materials and methods

6.2.1 Compound collection, training and testing datasets, molecular descriptors

A total of 233-1,316 non-dual inhibitors of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3, and 41-230 dual inhibitors of EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, Src-Lck, CDK1-CDK2, CDK1-GSK3, CDK2-GSK3, and CDK1-VEGFR, each with IC50≤10μM, were collected from the literature[277-286] and the BindingDB database[110]. Dual-inhibitors and non-dual inhibitors of a kinase-pair refer to inhibitors of both and one of the two kinases respectively regardless of their activities against other kinases. **Table 6-1** summarises the statistics of these inhibitors and MDDR compounds similar to at least one dual-inhibitor. **Figure 6-2** shows the Venn graph of our collected dual-

inhibitors the 11 evaluated kinase pairs and non-dual-inhibitors of the 9 evaluated kinases. 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[89,219]. First, 13.56 million PubChem and 168 thousand MDDR compounds were clustered into 8,993 compound families of similar molecular descriptors[189], 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[3], and 2,851 clusters for 171,045 natural products[190]. 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.

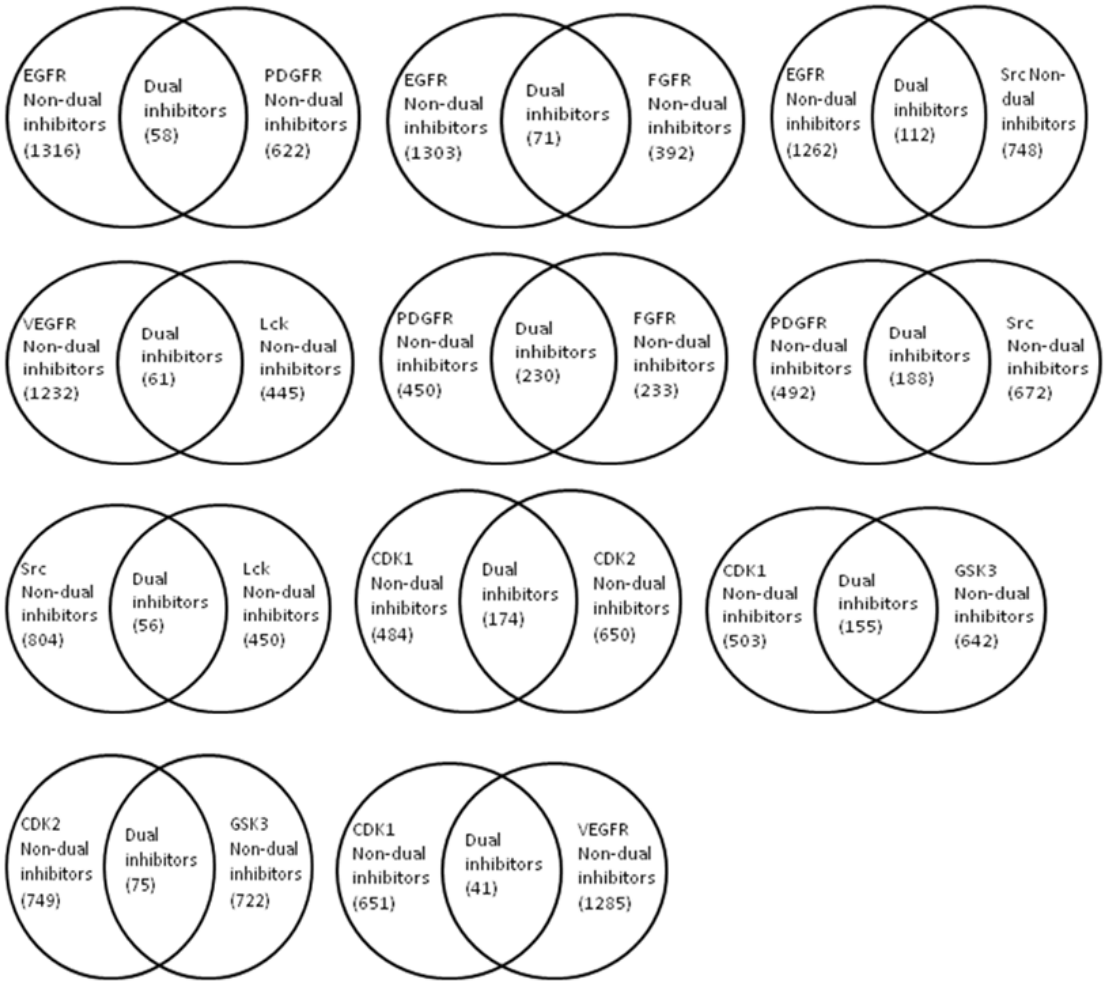


Figure 6-2 The Venn graph of the collected dual-inhibitors the 11 evaluated kinase-pairs and non-dual-inhibitors of the 9 evaluated kinases.

Table 6-1 Datasets of dual-inhibitors and non-dual-inhibitors of the kinase-pairs used for developing and testing combinatorial SVM dual-inhibitor 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				Inhibitors of other 7 kinases	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 7 kinases studied in this work	No (%) of dual-inhibitors of A and B as inhibitor of more than 2 of the other 7 kinases studied in this work	No of inhibitors	No of Compounds
EGFR-PDGFR	1316 (384)	336 (70)	100 (19)	622 (202)	251 (70)	153 (23)	58 (40)	22 (37.9%)	50 (86.2%)	3 (5.2%)	4097	3806
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%)	4327	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%)	3971	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%)	4355	413
PDGFR-FGFR	450 (168)	100 (29)	118 (27)	233 (90)	89 (29)	79 (25)	230 (78)	90 (39.1%)	214 (93.0%)	3 (1.3%)	5180	3614
PDGFR-Src	492 (174)	237 (53)	144 (24)	672 (213)	206 (53)	170 (38)	188 (67)	71 (37.8%)	184 (97.9%)	3 (1.6%)	4741	2893
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%)	4783	276
CDK1-CDK2	484 (199)	183 (52)	99 (28)	650 (251)	178 (52)	68 (34)	174 (84)	53 (30.5%)	24 (13.8%)	0 (0.0%)	4785	2629
CDK1-GSK3	503 (224)	140 (45)	38 (20)	642 (266)	143 (45)	83 (22)	155 (51)	49 (31.6%)	17 (11.0%)	0 (0.0%)	4793	3279
CDK2-GSK3	749 (280)	226 (62)	58 (23)	722 (275)	249 (62)	107 (24)	75 (44)	31 (41.3%)	17 (22.7%)	0 (0.0%)	4547	1617
CDK1-VEGFR	651 (251)	250 (75)	23 (8)	1285(434)	251 (75)	70 (17)	41 (25)	7 (17.1%)	0 (0.0%)	0 (0.0%)	4149	427

The collected non-dual and dual inhibitors of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3 are distributed in 431, 456, 246, 170, 284, 192, 255, 301, and 295 families respectively, which is consistent with reported 191 unique scaffolds (154 clusters and 43 singletons) for 565 kinase inhibitors[94]. Because of the extensive efforts in searching kinase inhibitors, the number of undiscovered "inhibitor" families for each kinase in PubChem and MDDR is expected to be relatively small, most likely no more than several hundred families. The ratio of the "inhibitor" and "inactive" families for each kinase (hundreds families vs 8,534-8,823 families contained in PubChem and MDDR at present) is expected to be no more than ~999/8500, which is <13%. Therefore, putative non-inhibitor training dataset can be generated by extracting a few representative compounds from each of the families that contain no known inhibitor, with a maximum possible "wrong" classification rate of <13% even in the extreme and unlikely cases that all of the undiscovered inhibitors are misplaced into the non-inhibitor class (please refer to **Chapter 3 Section 3.2.2**). The noise level generated by up to 13% "wrong" negative family represntation is expected to be substantially smaller than the maximum 50% false-negative noise level tolerated by SVM[47]. It is noted that 40%-62.2% of the dual-inhibitor families contain no non-dual inhibitor of the same kinase-pair, whose representative compounds were included in the inactive training datasets as dual-inhibitors are supposed to be unknown in our study. A substantial percentage of the dual-inhibitors in these "non-inhibitor" families were non-the-less identified as dual-inhibitors by our C-SVMs.

A total of 98 important descriptors were chosen from the chemical descriptors calculated by our program MODEL which were used in this work. The detail about molecular descriptors is explained in **Chapter 2 Section 2.2**.

6.2.2 Computational models

SVM is based on the structural risk minimization principle of statistical learning theory[145]. It consistently shows outstanding classification performance, is less penalized by sample redundancy, has lower risk for over-fitting, is capable of accommodating large and structurally diverse training and testing datasets, and is fast in performing classification tasks[147,148]. However, the performance of SVM is critically dependent on the diversity of training datasets. Because of the limited knowledge of known inhibitors for many kinase targets, sufficiently good SVM VS tools may not be readily developed for these targets. Non-the-less, 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 detailed mathematical algorithms of SVM are described in **Chapter 2 Section 2.3.1**. Readers are referred to this section. Our 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.

6.3 Results and discussion

6.3.1 Dual-inhibitors and non-dual inhibitors of the studied kinase-pairs

As shown in **Table 6-1**, the numbers of dual-inhibitors and non-dual inhibitors of the kinase-pairs are 58, 1,316 and 622 for EGFR-PDGFR, 71, 1,303 and 392 for EGFR-FGFR, 112, 1,262 and 748 for EGFR-Src, 61, 1,232 and 445 for VEGFR-Lck, 230, 450 and 233 for PDGFR-FGFR, 188, 492 and 672 for PDGFR-Src, 56, 804 and 450 for Src-Lck, 174, 484, and 650 for CDK1-CDK2, 155, 503, and 642 for CDK1-GSK3, 75, 749, and 722 for CDK2-GSK3, and 41, 651, and 1285 for CDK1-VEGFR respectively. The dual-inhibitors and non-dual inhibitors are distributed in 17-84 and 90-427 families respectively. Hence, both the numbers and diversity of non-dual inhibitors and dual-inhibitors are at reasonable levels for developing and testing VS tools. The percentages of dual-inhibitors outside the common families of the non-dual inhibitors in the training datasets are 62.1% for EGFR-PDGFR, 57.9% for EGFR-FGFR, 58.9% for EGFR-Src, 52.5% for VEGFR-Lck, 60.9% for PDGFR-FGFR, 62.2% for PDGFR-Src, 58.9% for Src-Lck, 69.5% for CDK1-CDK2, 68.4% for CDK1-GSK3, 58.7% for CDK2-GSK3, and 82.9% for CDK1-VEGFR respectively. Therefore, these dual-inhibitors have substantial degree of novelty against non-dual inhibitors. Moreover, 0.0%-98.6% of the dual-inhibitors of the kinase-pairs are inhibitor of at least one of the other 7 kinases, but only up to 5.2% of the dual-inhibitors are inhibitor of at least 3 of the other 7 kinases. Hence, most of these dual-inhibitors are non-ubiquitous inhibitors and show some degree of kinase selectivity even-though the majority of them target more than 2 kinases.

Some distinguished features of dual-inhibitors may be probed by evaluating the top-6 scaffolds contained in higher percentages of the dual-inhibitors of the studied intra-PTK group kinase-pairs, which are shown in **Figure 6-3**. **Table 6-2** shows the distribution of these scaffolds in the dual-inhibitors and non-dual-inhibitors of the studied intra-PTK group kinase-pairs. Scaffold A is contained in 63.8% of EGFR-PDGFR, 76.1% of PDGFR-Src, 33.9% of EGFR-Src, 54.9% of EGFR-FGFR and 27.8% of VEGFR-Lck dual-inhibitors respectively; Scaffold B is contained in 57.1% of Src-Lck, 29.5% of VEGFR-Lck and 25.9% of EGFR-Src dual-inhibitors respectively. Scaffold A and scaffold B appear to be the backbone of majority of dual-inhibitors of the studied kinase-pairs. Scaffold C is mainly contained in 19.6% of EGFR-Src dual inhibitors. Scaffold D is mainly contained in 32.4% in EGFR-FGFR and 4.5% in EGFR-Src dual-inhibitors. Scaffold E is contained in 17.8% of PDGFR-FGFR, 8.6% of EGFR-PDGFR, 7.0% of EGFR-FGFR and 6.9% of PDGFR-Src dual-inhibitors. Scaffold F is contained in 37.5% of Src-Lck and 34.4% of VEGFR-Lck dual-inhibitors. These scaffolds are also contained, mostly at significantly lower percentage levels, in the non-dual inhibitors of at least one of the kinases of the respective kinase-pairs. Therefore, some specific variations of side-chain groups of these scaffolds appear to be sufficient to convert some dual-inhibitors into non-dual inhibitors, which suggest that physicochemical properties as well as structural features are important for distinguishing dual and non-dual inhibitors

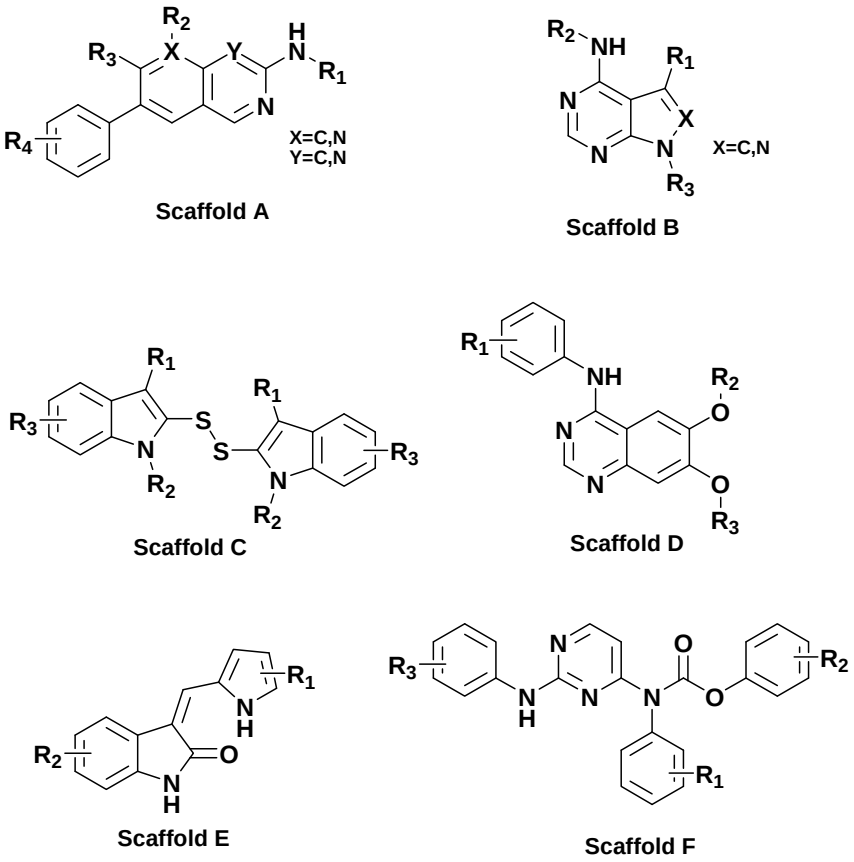


Figure 6-3 Top-6 scaffolds contained in higher percentages of the dual-inhibitors of the studied intra-PTK group kinase-pairs.

Table 6-2 Distribution of top-6 scaffolds in dual-inhibitors of 7 intra-PTK group kinase combinations of EGFR, VEGFR, PDGFR, FGFR, Src and Lck, and non-dual inhibitors of the constituent kinases

Kinase Pair	Datasets	Percentage of inhibitors containing scaffold A	Percentage of inhibitors containing scaffold B	Percentage of inhibitors containing scaffold C	Percentage of inhibitors containing scaffold D	Percentage of inhibitors containing scaffold E	Percentage of inhibitors containing scaffold F
EGFR-PDGFR	Dual inhibitors	63.8% (37/58)	0% (0/58)	0% (0/58)	1.7% (1/58)	8.6% (5/58)	0% (0/58)
	EGFR non-dual inhibitors	0.2% (3/1316)	6.3% (83/1316)	1.2% (16/1316)	7.7% (101/1316)	0% (0/1316)	0% (0/1316)
	PDGFR non-dual inhibitors	20.3% (126/622)	0% (0/622)	0% (0/622)	0% (0/622)	7.1% (44/622)	0% (0/622)
EGFR-FGFR	Dual inhibitors	54.9% (39/71)	0% (0/71)	0% (0/71)	32.4% (23/71)	7.0% (5/71)	0% (0/71)
	EGFR non-dual inhibitors	0.1% (1/1303)	6.4% (83/1303)	1.2% (16/1303)	6.1% (79/1303)	0% (0/1303)	0% (0/1303)
	FGFR non-dual inhibitors	25.5% (100/392)	0% (0/392)	0% (0/392)	2.3% (9/392)	10.0% (39/392)	0.3% (1/392)
EGFR-Src	Dual inhibitors	33.9% (38/112)	25.9% (29/112)	19.6% (22/112)	4.5% (5/112)	2.7% (3/112)	0% (0/112)
	EGFR non-dual inhibitors	0.2% (2/1262)	4.3% (54/1262)	1.6% (20/1262)	7.7% (97/1262)	0.2% (2/1262)	0% (0/1262)
	Src non-dual inhibitors	18.2% (136/748)	10.4% (78/748)	0.8% (6/748)	5.0% (37/748)	1.60% (12/748)	2.8% (21/748)
VEGFR-Lck	Dual inhibitors	27.9% (17/61)	29.5% (18/61)	0% (0/61)	0% (0/61)	0% (0/61)	34.4% (21/61)
	VEGFR non-dual inhibitors	0.7% (8/1232)	0.8% (10/1232)	0% (0/1232)	5.4% (66/1232)	4.7% (58/1232)	0% (0/1232)
	Lck non-dual	5.6% (25/445)	10.3% (46/445)	0% (0/445)	1.6% (7/445)	0% (0/445)	1.6% (7/445)

	inhibitors						
PDGFR-FGFR	Dual inhibitors	67.4% (155/230)	0% (0/230)	0% (0/230)	0% (0/230)	17.8% (41/230)	0% (0/230)
	PDGFR non-dual inhibitors	1.8% (8/450)	0% (0/450)	0% (0/450)	0.2% (1/450)	1.8% (8/450)	0% (0/450)
	FGFR non-dual inhibitors	11.2% (26/233)	0% (0/233)	0% (0/233)	13.7% (32/233)	1.3% (3/233)	0.4% (1/233)
PDGFR-Src	Dual inhibitors	76.1% (143/188)	0% (0/188)	0% (0/188)	0% (0/188)	6.9% (13/188)	0% (0/188)
	PDGFR non-dual inhibitors	2.9% (14/492)	0% (0/492)	0% (0/492)	0.2% (1/492)	7.3% (36/492)	0% (0/492)
	Src non-dual inhibitors	3.7% (25/672)	15.9% (107/672)	1.9% (13/672)	6.3% (42/672)	0.3% (2/672)	3.1% (21/672)
Src-Lck	Dual inhibitors	0% (0/56)	57.1% (32/56)	0% (0/56)	1.8% (1/56)	1.8% (1/56)	37.5% (21/56)
	Src non-dual inhibitors	21.6% (174/804)	9.3% (75/804)	1.6% (13/804)	5.1% (41/804)	1.9% (15/804)	0% (0/804)
	Lck non-dual inhibitors	5.9% (26/450)	7.8% (35/450)	0% (0/450)	1.3% (6/450)	0% (0/450)	2% (9/450)

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

The VS performance of C-SVMs in identifying dual-inhibitors of the 11 kinase-pairs is summarised in **Table 6-3** and further shown in **Figure 6-4**. The parameters of the developed SVM classification models for the evaluated kinases are in the ranges of σ=0.5~0.8. The dual-inhibitor yields are 27.6% for EGFR-PDGFR, 40.9% for EGFR-FGFR, 26.8% for EGFR-Src, 52.6% for VEGFR-Lck, 33.9% for PDGFR-FGFR, 38.3% for PDGFR-Src, 48.2% for Src-Lck, 52.3% for CDK1-CDK2, 49.0% for CDK1-GSK3, 57.3% for CDK2-GSK3, and 12.2% for CDK1-VEGFR respectively. The yields for the intra-PTK group and intra-CMGC group kinase pairs are comparable to the expected 25%-49% yields of combinations of good VS tools with individual yields of 50%-70%. Therefore, C-SVMs show reasonably good capability in identifying multi-target agents for kinase-pairs within a protein kinase group without requiring explicit knowledge of multi-target agents. However, the yield for the inter-PTK-CMGC kinase group CDK1-VEGFR kinase-pair is only 12.2%, which is significantly lower than those for the intra-PTK group and intra-CMGC group kinase-pairs. Structural analysis of the inhibitors of CDK1 and VEGFR binding sites has revealed that inhibitors generally make extensive favorable van der Waals contacts and several hydrogen bonds with Lys33, Leu83 and Asp86 at the hinge region of CDK1, and with Cys919, Asn923, Cys1045 and Asp1046 at the hinge region of VEGFR respectively, relatively small structural changes may easily reduce the optimal fit to the binding site, and some dual-inhibitors are able to bind to both kinases because of their structural flexibility to tolerate the different binding site geometry and

to form alternative hydrogen bonds[287]. In some cases, dual selectivity of inhibitors of inter-kinase-group kinase-pairs may require structural flexibility to fit in a hydrophobic pocket conserved in both kinase classes[288]. Such special structural features in dual-inhibitors of inter-kinase-group kinase-pairs are not necessarily needed and thus may not be found in non-dual inhibitors of individual kinases used in our training datasets, which likely be an important reason for the reduced yield of C-SVM in identifying CDK1-VEGFR dual-inhibitors. The smaller number of known CDK1-VEGFR dual-inhibitors may also affect the accurate assessment of VS outcome.

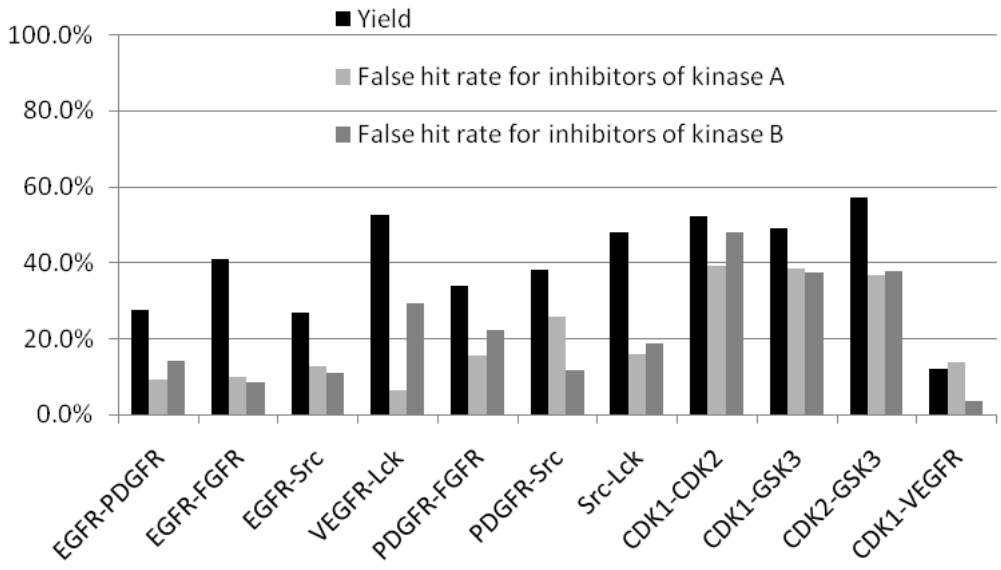


Figure 6-4 The VS performance of C-SVMs in identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3

Table 6-3 Virtual screening performance of combinatorial SVMs for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3

Kinase	**Virtual Screening Performance**								
	Dual inhibitors		Non-dual inhibitors of the same kinase pair		Inhibitors of other 7 kinases	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	False hit rate	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-PDGFR	27.60%	9 (15.5%)	9.20%	14.30%	1.88%	1.5% (57)	0.10% (175)	0.031% (4155)	0.025% (257)
EGFR-FGFR	40.90%	6 (8.5%)	10.10%	8.70%	1.06%	6.5% (65)	0.07% (126)	0.016% (2200)	0.004% (36)
EGFR-Src	26.80%	13 (11.6%)	12.90%	11.10%	1.49%	2.13% (24)	0.096% (162)	0.033% (4471)	0.007% (76)
VEGFR-Lck	52.60%	8 (13.1%)	6.60%	29.20%	2.80%	5.1% (21)	0.10% (170)	0.036% (4817)	0.011% (113)
PDGFR-FGFR	33.90%	35 (15.2%)	15.60%	22.30%	0.98%	1.4% (51)	0.057% (95)	0.013% (1746)	0.0008% (8)
PDGFR-Src	38.30%	30 (16.0%)	25.80%	11.60%	1.81%	2.9% (84)	0.104% (175)	0.021% (2799)	0.001% (14)
Src-Lck	48.20%	9 (16.1%)	15.80%	18.70%	0.98%	9.4% (26)	0.078% (131)	0.020% (2674)	0.002% (25)
CDK1-CDK2	52.30%	57 (32.8%)	39.20%	48.10%	3.39%	0.34% (9)	0.075% (126)	0.022% (2953)	0.014% (139)
CDK1-GSK3	49.00%	41 (26.5%)	38.40%	37.40%	4.30%	0.30% (10)	0.028% (47)	0.016% (2218)	0.016% (159)
CDK2-GSK3	57.30%	24 (32.0%)	36.80%	37.70%	2.99%	0.43% (7)	0.085% (142)	0.021% (2901)	0.020% (203)
CDK1-VEGFR	12.20%	0 (0.0%)	14.00%	3.70%	4.77%	0.0% (0)	0.007% (12)	0.023% (3113)	0.002% (19)

Target selectivity was tested by using C-SVMs to screen the 233-1,316 non-dual inhibitors of the 11 kinase-pairs, which misidentified 9.2% and 14.3% of the non-dual inhibitors of the kinase-pair as dual-inhibitors for EGFR-PDGFR, 10.1% and 8.7% for EGFR-FGFR, 12.9% and 11.1% for EGFR-Src, 6.6% and 29.2% for VEGFR-Lck, 15.6% and 22.3% for PDGFR-FGFR, 25.8% and 11.6% for PDGFR-Src, 15.8% and 18.7% for Src-Lck, 39.2% and 48.1% for CDK1-CDK2, 38.4% and 37.4% for CDK1-GSK3, 36.8% and 37.7% for CDK2-GSK3, and 14.0% and 3.7% for CDK1-VEGFR respectively. Therefore, C-SVMs are reasonably selective in distinguishing dual-inhibitors from non-dual inhibitors. There are two possible reasons for the misidentification of a substantial percentage of non-dual inhibitors as dual-inhibitors. First, SVMs were trained by non-dual inhibitors only, which may not fully distinguish dual and non-dual inhibitors. Secondly, some of the misidentified non-dual inhibitors are probably true dual-inhibitors not yet experimentally tested for multi-target activities. It is noted that "mistaken" selection of these non-dual inhibitors is still useful for searching single-target leads.

Target selectivity was further tested by using C-SVMs to screen the 3,971-5,180 inhibitors of the other 7 kinases not included in a particular kinase-pair. We found that 1.88% of these inhibitors were misidentified as dual-inhibitors for EGFR-PDGFR, 1.06% for EGFR-FGFR, 1.49% for EGFR-Src, 2.80% for VEGFR-Lck, 0.98% for PDGFR-FGFR, 1.81% for PDGFR-Src, 0.98% for Src-Lck, 3.39% for CDK1-CDK2, 4.30% for CDK1-GSK3, 2.99% for CDK2-GSK3, and 4.77% for CDK1-VEGFR respectively. These showed that C-

SVMs are fairly selective in separating inhibitors of specific kinase-pair from those of other kinases.

Virtual-hit rates and false-hit rates of C-SVMs in screening compounds that resemble the structural and physicochemical properties of the training datasets were evaluated by using 276-3,614 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[219]. C-SVMs identified 57 virtual-hits from 3,806 MDDR similarity compounds (virtual-hit rate 1,5%) for EGFR-PDGFR, 65 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, 51 from 3,614 MDDR compounds (1.4%) for PDGFR-FGFR, 84 from 2,893 MDDR compounds (2.9%) for PDGFR-Src, 26 from 276 MDDR compounds (9.4%) for Src-Lck, 9 from 2,629 MDDR compounds (0.34%) for CDK1-CDK2, 10 from 3,279 MDDR compounds (0.30%) for CDK1-GSK3, 7 from 1,617 MDDR compounds (0.43%) for CDK2-GSK3, and 0 from 505 MDDR compounds (0.0%) for CDK1-VEGFR respectively.

Significantly lower virtual-hit rates and thus false-hit rates were found in screening large libraries of 168 thousand MDDR and 13.56 million PubChem compounds. The numbers of virtual-hits and virtual-hit rates in screening 168 thousand MDDR compounds are 175 and 0.1% for EGFR-PDGFR, 126 and 0.07% for EGFR-FGFR, 162 and 0.096% for EGFR-Src, 170 and 0.1% for VEGFR-Lck, 95 and 0.057% for PDGFR-FGFR, 175 and 0.104% for

PDGFR-Src, and 131 and 0.078% for Src-Lck, 126 and 0.075% for CDK1-CDK2, 47 and 0.028% for CDK1-GSK3, 142 and 0.085% for CDK2-GSK3 and 12 and 0.007% for CDK1-VEGFR respectively. The numbers of virtual-hits and virtual-hit rates in screening 13.56M PubChem compounds are 4,155 and 0.031% for EGFR-PDGFR, 2,200 and 0.015% for EGFR-FGFR, 4,471 and 0.033% for EGFR-Src, 4,817 and 0.036% for VEGFR-Lck, 1,746 and 0.013% for PDGFR-FGFR, 2,799 and 0.021% for PDGFR-Src, 2,674 and 0.02% for Src-Lck, 2,953 and 0.022% for CDK1-CDK2, 2,218 and 0.016% for CDK1-GSK3, 2,901 and 0.021% for CDK2-GSK3, and 3,113 and 0.023% for CDK1-VEGFR respectively.

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, tyrosine-specific protein kinase inhibitors, and signal transduction inhibitors (**Table 6-5**, details in next section). 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 our combinatorial SVMs are at most equal to and likely less than the virtual-hit rates. Hence the false-hit rates are ≤1.4%-9.4% in screening 276-3,614 MDDR similarity compounds, ≤0.057%-0.104% in screening 168 thousand MDDR compounds, and ≤0.013%-0.036% 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[89,219], 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[220].

6.3.3 Comparison of the performance of Combinatorial SVM with other virtual screening methods

The VS performance of C-SVMs was further compared with DOCK 3.5.54 at the DOCK Blaster server[275], kNN[73], and PNN[276] by using the common testing datasets composed of 41~230 dual-inhibitors of the 11 evaluated kinase-pairs (set-1), 3,971~5,180 non-dual inhibitors of the 9 evaluated kinases (set-2), and 1.02 million Zinc clean-leads dataset (Zinc-CLD) [111] (set-3) respectively. DOCK VS studies were conducted against the protein crystal structures typically used in DOCK Blaster VS studies[275]. Specifically, the PDB entry for EGFR, FGFR, c-Src, VEGFR, CDK2, Lck, and GSK3 are 3BEL, 3C4F, 1YOL, 1Y6B, 2A4L, 2OG8, and 1Q5K respectively[275]. Moreover, a modelled 3D structure of PDGFR in the well-established molecular docking benchmarking sets[158] was used for PDGFR. CDK1 was not evaluated because we were unable to find a published experimental or modelled 3D structure.

In DOCK studies, the dual-inhibitor yield was estimated based on the screening results of set-1 and set-2 compounds, which is the percentage of the known dual-inhibitors made to the top-50% of the successfully docked set-1 and set-2 compounds for every kinase of a kinase-pair, the false-hit rate for misidentifying inhibitors of other 7 kinases as dual-inhibitors of a kinase-pair is the percentage of these inhibitors made to the top-50% of the successfully docked set-1 and set-2 compounds for every kinase of that kinase-pair, and the virtual-hit rate for the Zinc-CLD compounds is the percentage of these compounds made to the top-2% of the successfully docked set-3 compounds

for every kinase of that kinase-pair. The kNN and PNN methods and software used in this study were described in **Chapter 2 Section 2.3.2 and Section 2.3.3**. The training datasets of kNN and PNN and the methods for estimating the yield and virtual hit rate are the same as those of SVM. The parameters of the developed k-NN and PNN classification models for the evaluated kinases are in the ranges of k=1 or 3, and δ=0.003~0.11 respectively. The CPU time is ~0.12, ~8, and ~5.5 hours per kinase target of SVM, kNN, and PNN models in screening the 1.02 million Zinc clean-leads dataset respectively. The classification speed of SVM is faster than that of k-NN and PNN due to the fact that SVM typically uses 0.007~0.017% of the training dataset as support vectors for classification, whereas k-NN and PNN use the whole training dataset. It took ~ 2 weeks to get the docking results from the DOCK Blaster server for screening the whole Zinc clean-leads dataset per kinase target.

Table 6-4 and **Figure 6-5** shows the comparison of the performance of C-SVMs with the other three VS methods for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3 from the three common testing datasets. Overall, the yields of all VS methods are comparable, mostly in the ranges of 21.3%~57.3% for the intra-PTK and intra-CMGC group kinase-pairs and 12.2%~19.5% for the inter-PTK-CMGC group kinase-pair. C-SVM, kNN and PNN also produced comparable false hit-rates, at 0.98%~6.05%, for misidentifying inhibitors of other 7 kinases as dual-inhibitors of the evaluated kinase-pairs, with SVM showing slightly lower false hit-rates for the majority of the evaluated kinase-pairs.

For the 8 kinase-pairs with available 3D structure, DOCK produced higher false hit-rates than other three evaluated VS methods in misidentifying inhibitors of other 7 kinases as dual-inhibitors. These false-hit rates may be significantly reduced by adjusting the docking cut-off values for individual kinases, e.g. from top-50% to top-10%, which may however lead to significantly reduced yields. High false-positive rates has been a common issue in structure-based VS, and the false-positives in kinase docking studies arise partly from the inability to favourably score certain key hydrogen-binding interactions required for kinase binding and to discriminate conformational artifacts of docked ligands[60]. False-hit rates can be significantly reduced by such strategies as the incorporation of the reported kinase binding features into docking constrains[60], consensus scoring using multiple ligand information and maximum common binding modes for multiple kinases[105], and combining docking with pharmacophore filtering[289].

C-SVM produced substantially lower virtual-hit rates (0.008%~0.025%) than those (0.009%~0.348%) of the other three VS methods for identifying the Zinc-CLD compounds as virtual dual-inhibitors of the evaluated kinase-pairs. The numbers of Zinc-CLD compounds identified as virtual-hits by C-SVM are in the range of 8~203, compared to those of 1439~3963, 96~1406, and 332~2830 by DOCK, kNN, and PNN respectively. The numbers of un-discovered dual-inhibitors of the evaluated kinase-pairs in the Zinc-CLD are un-known. It is noted that only 12.1% of the known dual-inhibitors of the evaluated kinase-pairs and 14.0% the known non-dual inhibitors of the

evaluated kinases satisfy the criteria used for assembling the Zinc-CLD. Therefore, the numbers of un-discovered dual-inhibitors in the Zinc-CLD are expected to be very small, most likely fewer than 100. Based on this estimate, the minimum and maximum numbers of false-hits of C-SVM, DOCK, kNN, and PNN are 0~103 and 8~203, 1339~3863 and 1439~3963, 0~1306 and 96~1406, and 232~2730 and 332~2839 respectively. C-SVM appears to show substantially lower false-hit rates than those of the other three VS methods in screening a large compound database.

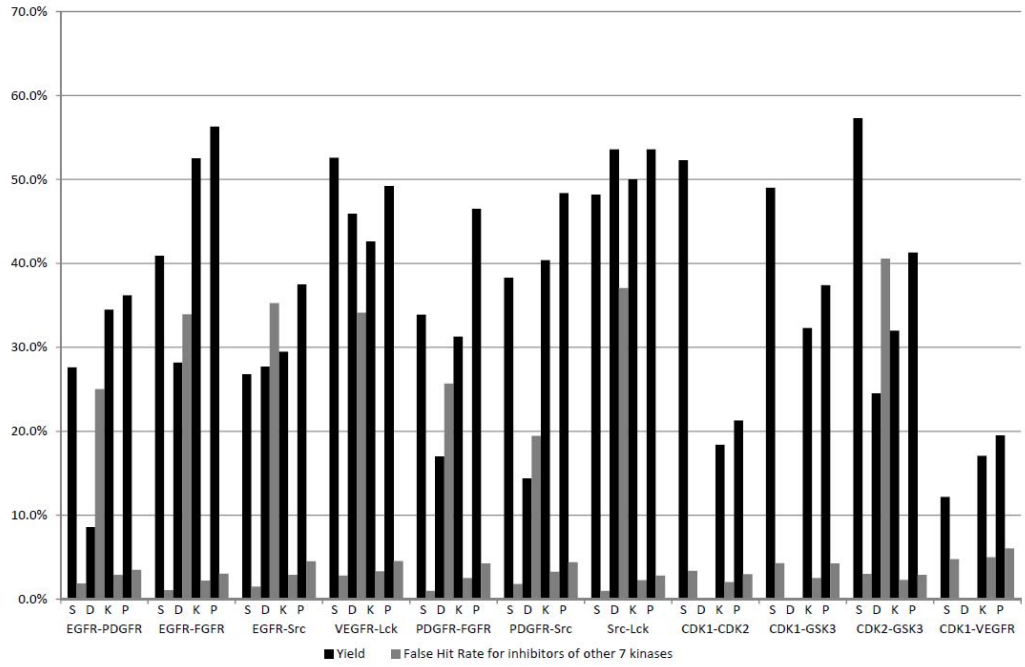


Figure 6-5 The comparison of the performance of C-SVMs with the other three VS methods DOCK, kNN and PNN for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3. The labels S, D, K, P beneath the performance bars represent C-SVM, DOCK, kNN, and PNN respectively.

Table 6-4 Comparison of the performance of combinatorial SVMs with other virtual screening methods for identifying dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3.

Kinase Pair	**Virtual Screening Performance**											
	SVM			KNN			PNN			DOCK		
	Yield of Dual inhibitors	False Hit Rate for Predicting Inhibitors of Other 7 Kinases as Dual Inhibitor of the Kinase Pair	Virtual hit rate (No of virtual hits) for screening 1.02 million Zinc clean-leads dataset	Yield of Dual inhibitors	False Hit Rate for Predicting Inhibitors of Other 7 Kinases as Dual Inhibitor of the Kinase Pair	Virtual hit rate (No of virtual hits) for screening 1.02 million Zinc clean-leads dataset	Yield of Dual inhibitors	False Hit Rate for Predicting Inhibitors of Other 7 Kinases as Dual Inhibitor of the Kinase Pair	Virtual hit rate (No of virtual hits) for screening 1.02 million Zinc clean-leads dataset	Yield of Dual inhibitors	False Hit Rate for Predicting Inhibitors of Other 7 Kinases as Dual Inhibitor of the Kinase Pair	Virtual hit rate (No of virtual hits) for screening 1.02 million Zinc clean-leads dataset
EGFR-PDGFR	27.60%	1.88%	0.025% (257)	34.50%	2.88%	0.112% (1144)	36.20%	3.49%	0.217% (2211)	8.60%	25.04%	0.141% (1439)
EGFR-FGFR	40.90%	1.06%	0.004% (36)	52.50%	2.22%	0.057% (579)	56.30%	3.03%	0.095% (971)	28.20%	33.93%	0.247% (2516)
EGFR-Src	26.80%	1.49%	0.007% (76)	29.50%	2.90%	0.081% (824)	37.50%	4.53%	0.107% (1095)	27.70%	35.28%	0.158% (1615)
VEGFR-Lck	52.60%	2.80%	0.011% (113)	42.60%	3.33%	0.091% (927)	49.20%	4.55%	0.167% (1700)	45.90%	34.14%	0.236% (2404)

PDGFR-FGFR	33.90%	0.98%	0.0008% (8)	31.30%	2.51%	0.009% (96)	46.50%	4.27%	0.033% (332)	17.00%	25.68%	0.291% (2968)
PDGFR-Src	38.30%	1.81%	0.001% (14)	40.40%	3.27%	0.048% (494)	48.40%	4.41%	0.105% (1070)	14.40%	19.45%	0.144% (1468)
Src-Lck	48.20%	0.98%	0.002% (25)	50.00%	2.26%	0.029% (294)	53.60%	2.82%	0.037% (376)	53.60%	37.07%	0.348% (3542)
CDK1-CDK2	52.30%	3.39%	0.014% (139)	18.40%	2.03%	0.135% (1377)	21.30%	2.97%	0.367% (3738)	N.A	N.A	N.A
CDK1-GSK3	49.00%	4.30%	0.016% (159)	32.30%	2.51%	0.131% (1331)	37.40%	4.27%	0.281% (2865)	N.A	N.A	N.A
CDK2-GSK3	57.30%	2.99%	0.020% (203)	32.00%	2.31%	0.118% (1203)	41.30%	2.88%	0.245% (2498)	24.50%	40.55%	0.389% (3963)
CDK1-VEGFR	12.20%	4.77%	0.002% (19)	17.10%	5.01%	0.138% (1409)	19.50%	6.05%	0.278% (2830)	N.A	N.A	N.A

6.3.4 Evaluation of Combinatorial SVM identified MDDR virtual-hits

C-SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. **Table 6-5** gives the MDDR classes that contain higher percentage (≥9%) of C-SVM virtual-hits and the percentage values. We found that 58-110 or 50%-62% of the 95-175 virtual-hits belong to the antineoplastic class, which represent 0.30%-0.51% of the 21,557 MDDR compounds in the class. In particular, 34-71 or 21%-40% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 2.9%-6.0% of the 1,181 MDDR compounds in the class. Moreover, 13%-28% and 9%-14% of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 0.83%-2.4% and 0.98%-1.5% of the 2,037 and 1,629 members in the two classes respectively. Therefore, many of the C-SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis and other cancer-related pathways. While some of these kinase inhibitors might be true dual-inhibitors of specific kinase-pairs, the majority of them are expected to arise from false selection of non-dual inhibitors of the same kinase-pairs (at 6.6%-29.2% false-hit rates) and inhibitors of other kinases (at 0.2%-12.7% false-hit rates).

Some of the C-SVM virtual-hits belong to the antiarthritic class. Five of our evaluated kinases or their kinase-likes 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[89]. VEGF

has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis[223]. FGFR may partly mediates osteoarthritis[224]. PDGF-like factors stimulates the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells[225]. Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma[226]. Therefore, some of the C-SVM virtual-hits in the antiarthritic class may be inhibitors of our evaluated kinases or their kinase-likes capable of producing antiarthritic activities.

Moreover, some of the C-SVM virtual-hits for PDGFR-FGFR belong to the atherosclerosis therapy class. Both kinases have been implicated in atherosclerosis. PDGF drives pathological mesenchymal responses in such vascular disorders as atherosclerosis, restenosis, pulmonary hypertension, retinal diseases, and fibrotic diseases[290]. Multiple FGFRs are elevated in therosclerotic 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[291]. Therefore, some of the C-SVM virtual-hits in the atherosclerosis therapy may be the inhibitors of the two kinases.

Table 6-5 MDDR classes that contain higher percentage (≥9%) of virtual-hits identified by combinatorial SVMs in screening 168 thousand MDDR compounds for dual-inhibitors of 11 combinations of EGFR, VEGFR, PDGFR, FGFR, Src, Lck, CDK1, CDK2, and GSK3.

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-PDGFR	175	Antineoplastic	110	0.50%
		Tyrosine-Specific Protein Kinase Inhibitor	71	6.00%
		Signal Transduction Inhibitor	39	2.00%
		Antiangiogenic	25	1.50%
		Antiarthritic	21	0.20%
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%

PDGFR-FGFR	95	Antineoplastic	58	0.30%
		Tyrosine-Specific Protein Kinase Inhibitor	27	2.30%
		Signal Transduction Inhibitor	22	1.10%
		Atherosclerosis Therapy	10	0.90%
		Antiarthritic	10	0.10%
PDGFR-Src	175	Antineoplastic	103	0.50%
		Signal Transduction Inhibitor	49	2.40%
		Tyrosine-Specific Protein Kinase Inhibitor	40	3.40%
		Antiangiogenic	16	1.00%
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%
CDK1-CDK2	126	Antineoplastic	87	0.40%
		Protein Kinase C Inhibitor	23	4.02%
		Antiviral	20	0.51%
		Tyrosine-Specific Protein Kinase Inhibitor	19	1.61%
		Signal Transduction Inhibitor	14	0.69%
CDK1-GSK3	47	Antineoplastic	27	0.13%
		Tyrosine-Specific Protein Kinase Inhibitor	10	0.85%
		Antihypertensive	5	0.05%
		Protein Kinase C Inhibitor	5	0.87%
		Antidepressant	4	0.06%
CDK2-GSK3	142	Antineoplastic	100	0.46%
		Protein Kinase C Inhibitor	28	4.90%

		Antihypertensive	21	0.19%
		Antiviral	20	0.51%
		Signal Transduction Inhibitor	18	0.88%
CDK1-VEGFR	12	Antineoplastic	5	0.02%
		Tyrosine-Specific Protein Kinase Inhibitor	3	0.25%
		Neuronal Injury Inhibitor	2	0.04%
		Antiangiogenic	2	0.12%
		Antiarthritic	2	0.02%

6.3.5 Does Combinatorial SVM select kinase inhibitors or membership of compound families?

To further evaluate whether C-SVMs identify kinase inhibitors rather than membership of certain compound families, Compound family distribution of the identified dual-inhibitors of the 7 intra-PTK group kinase-pairs were analyzed. As shown in **Table 6-4**, 15.5%, 8.5%, 11.6%, 13.1%, 15.2%, 16.0% and 16.1% of the identified EGFR-PDGFR, EGFR-FGFR, EGFR-Src, VEGFR-Lck, PDGFR-FGFR, PDGFR-Src, and Src-Lck dual-inhibitors are outside the families that contain at least one pair of non-dual inhibitors of the two kinases of the kinase-pair (i.e., at least one inhibitor for kinase A and one inhibitor for kinase B). For those families that contain at least one pair of non-dual inhibitors of the two kinases of a kinase-pair, 17.2%-68.2% of the compounds (>40.0% in majority cases) in each of these families were predicted as non-dual inhibitors by C-SVMs. These results suggest that C-SVMs identify dual-inhibitors not solely based on membership to certain compound families.

6.3.6 Molecular features important for selecting dual-kinase inhibitors

The molecular features important for selecting dual-kinase inhibitors were preliminarily analyzed by testing the VS performance with varying sets of molecular descriptors. Our analysis suggested that the VS performance is critically dependent on a proper combination of multiple simple molecular property descriptors that reflect ring and hydrogen binding features, chemical

property descriptors that represent hydrophobicity and molecular polarizability, molecular connectivity and shape profile descriptors that define the structural and flexibility features, and electro-topological state descriptors that determine the molecular skeletons, structural frameworks and their electronic properties. Our analysis is consistent with the reported structural analysis of the inhibitors of CDK1 and VEGFR that shows the importance of molecular structures for making extensive van der Waals contacts, hydrogen bonding with specific residues in both kinases, and structural flexibility to accommodate the different binding site geometry and to allow the formation of alternative hydrogen bonds. Our analysis is also consistent with another report that dual-kinase binding may require a combination of structural flexibility and the favourable hydrophobic interactions at specific pocket conserved in both kinase classes. Moreover, many dual-inhibitors adopt specific scaffolds, such as those illustrated in **Figure 6-3**, that enable them to more easily fit to the particular regions of the ATP site, which may be partly captured by the electro-topological state descriptors. A more comprehensive analysis using structural-based and feature selection methods may shed more light on the detailed molecular features of dual-kinase inhibition as well as single kinase inhibition.

6.4 Further perspective

Combinatorial SVM VS tools developed by using non-dual inhibitors show good capability in identifying dual-inhibitors of several anticancer target kinase-pairs at comparable and in many cases substantially lower false-hit rates than those of typical VS tools reported in the literatures. 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. With the discovery of increasing number of selective multi-target agents 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. These multi-target VS tools may be combined with structure-based filters for enhanced target selectivity. Because of their high computing speed and generalization capability, combinatorial SVM can be potentially explored to develop 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 7 Concluding Remarks

This last chapter summarizes the major findings and contributions of this study (Section 7.1). Limitation of present study and suggestion on possible areas for further studies are discussed in Section 7.2.

7.1 Major findings and contributions

Machine learning methods have been explored for developing such alternative VS tools because of their high-CPU speed and capability for covering highly diverse spectrum of compounds. However, while exhibiting equally good hit selection performance in screening extremely-large and large libraries, the currently developed machine learning tools tend to show lower hit-rate and, in some cases, lower enrichment factor than the best performing SBVS tools. This work selected the most popular ML method support vector machine to test whether the performance of SVM can be improved by using training-sets of diverse inactive compounds. Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds. This approach was applied for generating putative inactive compounds. An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the "inactive" chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. In retrospective database screening of active compounds from large libraries such as PubChem, MDDR and ZINC, The hit-rates of our methods are comparable and the enrichment factors are substantially better than the best results of other

VS tools. The putative negatives generation method plays an important role in it. This method greatly increased the performance of VS without losing much positive accuracy. It showed that at the study of chemistry and biological problems, certain assumption could be made to solve the problems although sometimes it may lead to certain degree of noise.

This work also evaluated the performance of SVM trained by sparsely distributed actives (regularly sparse and very sparse actives) in six MDDR biological target classes composed of high number of known actives (983~1,645) of high, intermediate, and low structural diversity (muscarinic M1 receptor agonists, NMDA receptor antagonists, thrombin inhibitors, HIV protease inhibitors, cephalosporins, and rennin inhibitors). Comparing the results with those of data fusion method, the yields of our regularly sparse SVM models are slightly improved for the high and intermediate classes, and the false-hit rates of our SVM models are substantially reduced for all three classes. These results suggest that, by using the equally small number of active compounds as training data, SVM is capable of producing equally good or slightly better yields and generalization capability at substantially reduced false-hit rates than those of the data fusion method. It was also found that our SVM models have substantial capability in identifying novel active compounds from sparse active datasets at low false-hit rates. An important feature of these SVM virtual screening methods is that they have generalization capability for covering highly diverse spectrum compounds. Even based on the sparse active datasets, SVM also can be potentially used to

develop useful VS (virtual screening) tools or be used as part of integrated VS tools in facilitating lead discovery.

By using training dataset of more diverse spectrum of inactive compounds as well as substantial number of literature-reported c-Src and VEGFR-2 inhibitors, the results of SVM based virtual screening shows substantial capability in identifying c-Src and VEGFR-2 inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in the literatures. It is capable of searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates without the need to define an applicability domain, i.e. it has a broad applicability domain that covers the whole chemical space defined by the PubChem and MDDR databases. Because of their high computing speed and generalization capability for covering highly diverse spectrum compounds, Our SVM models can be applied to discover the potential leads of c-Src and VEGFR-2 inhibitors for pharmaceutical purposes.

This work on the prediction of multi-target kinase inhibitors pioneers the applicaton of SVM based virtual screening. Combinatorial support vector machines (C-SVMs) were tested as VS tools for searching dual-inhibitors of 11 combinations of 9 anticancer kinase targets (EGFR, VEGFR, PDGFR, Src, FGFR, Lck, CDK1, CDK2, GSK3). C-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. Compared with other methods,

Combinatorial SVM VS tools show good capability in identifying dual-inhibitors of several anticancer target kinase-pairs at comparable and in many cases substantially lower false-hit rates. Therefore, C-SVMs models are potentially useful to discover multi-target agents for enhancing efficacy and reducing counter-target activities and toxicities.

7.2 Limitations and suggestions for future studies

The SVM models developed using our putative negative dataset are not perfect. There are still some false hits that cannot be ruled out easily. These false hits are "correctly" identified by our SVM models due to the similar structural frameworks with real active compounds. Our molecular descriptors used in the SVM model are insufficient to adequately differentiate the compounds with similar structural frameworks. Therefore, it is necessary to explore different combinations of descriptors and to select any more optimal sets of descriptors by using more refined feature selection algorithms and parameters in future work. Also it may be helpful to introduce new descriptors for more appropriate representations of compounds or descriptors which can be used to describe the interaction between proteins and their ligands.

The putative negatives generation method helps a lot in improving the performance of SVM based virtual screening. However, 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. As will be demonstrated, such an adverse

effect is expected to be relatively small for many biological target classes. On the other hand, the clustering of chemical space also can affect the generation of putative negative dataset. Chemical space clustering is a difficult area in cheminformatics. The clustering method, distance metrix selection and descriptors are three important factors for clustering. K-means clustering method used in this work is not the best clustering method but is suitable and computable for large chemical spaces. In future studies, new clustering algorithm can be developed for improving the accuracy of chemical space clustering. The selection of correlation coefficients and other chemical descriptors such as fingerprint also can be the direction of improvement.

The good performance of our SVM based VS system has been showed in several projects. However, the good performance of virtual screening is not only in screening hits, yield and enrichment factors but also a good potential in terms of prediction of novel structure. Experimental studies are necessary to do for validating our high performance virtual screening tools. Based on this, we have formed extensive collaborations with several research groups on drug development. Some compounds are selected and sent to our collaborators for further study.

The capability of the combinatorial SVMs in identifying multi-kinase inhibitors and other multi-target agents need be further enhanced by incorporating knowledge of multi-target agents into VS tool development processes. With the discovery of increasing number of selective multi-target agents 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.

These years have seen plenty of debate aimed to define which of the many VS approaches is the best. However, this question remains with no conclusive answer. Each approach has its own advantages and drawbacks, and the choice of one or others depends on the particular question faced by the medicinal chemist. In terms of performance, ligand based methods tend to present better enrichment factors and higher speed serving as a more efficient methodologies to remove non active compounds but target based method provides a more straightforward picture of interactions between the drug and molecular target and a better prediction in terms of novel structures. Now a synergistic, rational, synthetic combination of different approaches is a trend. Combined VS approach tends to include less costly approaches, usually ligand based VS, at the first stage, while the most demanding methods, usually docking, for the last stage when the original large compound library has been reduced to a manageable size.

BIBLIOGRAPHY

1. Horvath D. A virtual screening approach applied to the search for trypanothione reductase inhibitors. J Med Chem 1997;40(15):2412-23.
2. Bohacek RS, McMartin C, Guida WC. The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 1996;16(1):3-50.
3. Reymond TFaJ-L. 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 2007;(published on Web 01/30/2007).
4. Rarey M, Stahl M. Similarity searching in large combinatorial chemistry spaces. J Comput Aided Mol Des 2001;15(6):497-520.
5. Cavasotto CN, Orry AJ. Ligand docking and structure-based virtual screening in drug discovery. Curr Top Med Chem 2007;7(10):1006-14.
6. Lyne PD. Structure-based virtual screening: an overview. Drug Discov Today 2002;7(20):1047-55.
7. Kroemer RT. Structure-based drug design: docking and scoring. Curr Protein Pept Sci 2007;8(4):312-28.
8. Sun H. Pharmacophore-based virtual screening. Curr Med Chem 2008;15(10):1018-24.
9. Xue L, Godden JW, Stahura FL, Bajorath J. Similarity search profiles as a diagnostic tool for the analysis of virtual screening calculations. J Chem Inf Comput Sci 2004;44(4):1275-81.
10. Guido RV, Oliva G, Andricopulo AD. Virtual screening and its integration with modern drug design technologies. Curr Med Chem 2008;15(1):37-46.
11. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol 2008;153 Suppl 1:S7-26.
12. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49(20):5912-31.
13. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 2003;9(1):47-57.
14. Kim R, Skolnick J. Assessment of programs for ligand binding affinity prediction. J Comput Chem 2008;29(8):1316-31.
15. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. 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 2008;22(3-4):213-28.
16. Sheridan RP, McGaughey GB, Cornell WD. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. J Comput Aided Mol Des 2008;22(3-4):257-65.
17. Jain AN. Bias, reporting, and sharing: computational evaluations of docking methods. J Comput Aided Mol Des 2008;22(3-4):201-12.
18. Hawkins PC, Skillman AG, Nicholls A. Comparison of shape-matching and docking as virtual screening tools. J Med Chem 2007;50(1):74-82.
19. Wolber G, Seidel T, Bendix F, Langer T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. Drug Discov Today 2008;13(1-2):23-9.
20. Carosati E, Budriesi R, Ioan P, Ugenti MP, Frosini M, Fusi F, Corda G, Cosimelli B, Spinelli D, Chiarini A, Cruciani G. Discovery of novel and cardioselective diltiazem-like calcium channel blockers via virtual screening. J Med Chem 2008;51(18):5552-65.

21. Moffat K, Gillet VJ, Whittle M, Bravi G, Leach AR. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. J Chem Inf Model 2008;48(4):719-29.
22. McGaughey GB, Sheridan RP, Bayly CI, Culberson JC, Kreatsoulas C, Lindsley S, Maiorov V, Truchon JF, Cornell WD. Comparison of topological, shape, and docking methods in virtual screening. J Chem Inf Model 2007;47(4):1504-19.
23. Stahura FL, Bajorath J. New methodologies for ligand-based virtual screening. Curr Pharm Des 2005;11(9):1189-202.
24. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers and Chemistry 2001;26(1):5-14.
25. Manallack DT, Livingstone DJ. Neural networks in drug discovery: have they lived up to their promise? European Journal of Medicinal Chemistry 1999;34(3):195-208.
26. Trotter MWB, Holden SB. Support vector machines for ADME property classification. QSAR & Combinatorial Science 2003;22(5):533-548.
27. Lengauer T, Lemmen C, Rarey M, Zimmermann M. Novel technologies for virtual screening. Drug Discov Today 2004;9(1):27-34.
28. Oprea TI, Matter H. Integrating virtual screening in lead discovery. Curr Opin Chem Biol 2004;8(4):349-58.
29. Bocker A, Schneider G, Teckentrup A. NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. J Chem Inf Model 2006;46(6):2220-9.
30. Schuster D, Maurer EM, Laggner C, Nashev LG, Wilckens T, Langer T, Odermatt A. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. J Med Chem 2006;49(12):3454-66.
31. Steindl T, Laggner C, Langer T. Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening. J Chem Inf Model 2005;45(3):716-24.
32. Schroeter T, Schwaighofer A, Mika S, Laak AT, Suelzle D, Ganzer U, Heinrich N, Muller KR. Machine Learning Models for Lipophilicity and Their Domain of Applicability. Mol Pharm 2007;4(4):524-538.
33. Li H, Yap CW, Ung CY, Xue Y, Li ZR, Han LY, Lin HH, Chen YZ. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. J Pharm Sci 2007;(Published Online).
34. Fox T, Kriegl JM. Machine learning techniques for in silico modeling of drug metabolism. Curr Top Med Chem 2006;6(15):1579-91.
35. Duch W, Swaminathan K, Meller J. Artificial intelligence approaches for rational drug design and discovery. Curr Pharm Des 2007;13(14):1497-508.
36. Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, Lewell XQ, Greenidge P, Stiefl N. Evaluation of machine-learning methods for ligand-based virtual screening. J Comput Aided Mol Des 2007.
37. Shoichet BK. Virtual screening of chemical libraries. Nature 2004;432(7019):862-5.
38. Stichtenoth DO, Frolich JC. The second generation of COX-2 inhibitors: what advantages do the newest offer? Drugs 2003;63(1):33-45.
39. Linkins LA, Weitz JI. Pharmacology and clinical potential of direct thrombin inhibitors. Current Pharmaceutical Design 2005;11(30):3877-3884.
40. Ribeiro S, Horuk R. The clinical potential of chemokine receptor antagonists. Pharmacology & Therapeutics 2005;107(1):44-58.
41. Spaltenstein A, Kazmierski WM, Miller JF, Samano V. Discovery of next generation inhibitors of HIV protease. Current topics in medicinal chemistry 2005;5(16):1589-1607.
42. Fabbro D, Ruetz S, Buchdunger E, Cowan-Jacob SW, Fendrich G, Liebetanz J, Mestan J, O'Reilly T, Traxler P, Chaudhuri B, Fretz H, Zimmermann J, Meyer T, Caravatti G, Furet P, Manley PW. Protein kinases as targets for anticancer agents: from inhibitors to useful drugs. Pharmacology & Therapeutics 2002;93(2-3):79-98.

43. Kumar R, Singh VP, Baker KM. Kinase inhibitors for cardiovascular disease. Journal of Molecular and Cellular Cardiology 2006;doi:10.1016/j.yjmcc.2006.09.005.
44. Rotella DP. Phosphodiesterase 5 inhibitors: current status and potential applications. Nature reviews Drug discovery 2002;1(9):674-682.
45. Pacher P, Kecskemeti V. Trends in the development of new antidepressants. Is there a light at the end of the tunnel? Current Medicinal Chemistry 2004;11(7):925-943.
46. Jorissen RN, Gilson MK. Virtual screening of molecular databases using a support vector machine. J Chem Inf Model 2005;45(3):549-61.
47. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW. 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 2006;46(1):193-200.
48. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A. 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 2006;46(2):462-70.
49. Franke L, Byvatov E, Werz O, Steinhilber D, Schneider P, Schneider G. Extraction and visualization of potential pharmacophore points using support vector machines: application to ligand-based virtual screening for COX-2 inhibitors. J Med Chem 2005;48(22):6997-7004.
50. Wilton DJ, Harrison RF, Willett P, Delaney J, Lawson K, Mullier G. Virtual screening using binary kernel discrimination: analysis of pesticide data. Journal of Chemical Information and Modeling 2006;46(2):471-477.
51. Chen B, Harrison RF, Pasupa K, Willett P, Wilton DJ, Wood DJ, Lewell XQ. Virtual screening using binary kernel discrimination: effect of noisy training data and the optimization of performance. Journal of Chemical Information and Modeling 2006;46(2):478-486.
52. Mozziconacci JC, Arnoult E, Bernard P, Do QT, Marot C, Morin-Allory L. Optimization and validation of a docking-scoring protocol; application to virtual screening for COX-2 inhibitors. J Med Chem 2005;48(4):1055-68.
53. Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP. Comparison of automated docking programs as virtual screening tools. J Med Chem 2005;48(4):962-76.
54. Evers A, Klabunde T. Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor. J Med Chem 2005;48(4):1088-97.
55. Lorber DM, Shoichet BK. Hierarchical docking of databases of multiple ligand conformations. Curr Top Med Chem 2005;5(8):739-49.
56. Alvarez JC. High-throughput docking as a source of novel drug leads. Curr Opin Chem Biol 2004;8(4):365-70.
57. Schapira M, Raaka BM, Das S, Fan L, Totrov M, Zhou Z, Wilson SR, Abagyan R, Samuels HH. Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking. Proc Natl Acad Sci U S A 2003;100(12):7354-9.
58. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2001;46(1-3):3-26.
59. Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J Med Chem 2003;46(13):2656-62.
60. Perola E. Minimizing false positives in kinase virtual screens. Proteins 2006;64(2):422-35.
61. Harper G, Bradshaw J, Gittins JC, Green DV, Leach AR. Prediction of biological activity for high-throughput screening using binary kernel discrimination. J Chem Inf Comput Sci 2001;41(5):1295-300.
62. Lepp Z, Kinoshita T, Chuman H. Screening for new antidepressant leads of multiple activities by support vector machines. J Chem Inf Model 2006;46(1):158-67.

63. J. Cui LYH, 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 2007;44:866-877.
64. Yamazaki K, Kusunose N, Fujita K, Sato H, Asano S, Dan A, Kanaoka M. Identification of phosphodiesterase-1 and 5 dual inhibitors by a ligand-based virtual screening optimized for lead evolution. Bioorganic & Medicinal Chemistry Letters 2006;16(5):1371-1379.
65. Vidal D, Thormann M, Pons M. A novel search engine for virtual screening of very large databases. J Chem Inf Model 2006;46(2):836-43.
66. Enyedy IJ, Ling Y, Nacro K, Tomita Y, Wu X, Cao Y, Guo R, Li B, Zhu X, Huang Y, Long YQ, Roller PP, Yang D, Wang S. Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. J Med Chem 2001;44(25):4313-24.
67. Doman TN, McGovern SL, Witherbee BJ, Kasten TP, Kurumbail R, Stallings WC, Connolly DT, Shoichet BK. Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. J Med Chem 2002;45(11):2213-21.
68. Wang JL, Liu D, Zhang ZJ, Shan S, Han X, Srinivasula SM, Croce CM, Alnemri ES, Huang Z. 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 2000;97(13):7124-9.
69. Stiefl N, Zaliani A. A knowledge-based weighting approach to ligand-based virtual screening. J Chem Inf Model 2006;46(2):587-96.
70. Pirard B, Brendel J, Peukert S. The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. J Chem Inf Model 2005;45(2):477-85.
71. Rella M, Rushworth CA, Guy JL, Turner AJ, Langer T, Jackson RM. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. J Chem Inf Model 2006;46(2):708-16.
72. Giaccone G, Herbst RS, Manegold C, Scagliotti G, Rosell R, Miller V, Natale RB, Schiller JH, Von Pawel J, Pluzanska A, Gatzemeier U, Grous J, Ochs JS, Averbuch SD, Wolf MK, Rennie P, Fandi A, Johnson DH. Gefitinib in combination with gemcitabine and cisplatin in advanced non-small-cell lung cancer: a phase III trial--INTACT 1. J Clin Oncol 2004;22(5):777-84.
73. Herbst RS, Giaccone G, Schiller JH, Natale RB, Miller V, Manegold C, Scagliotti G, Rosell R, Oliff I, Reeves JA, Wolf MK, Krebs AD, Averbuch SD, Ochs JS, Grous J, Fandi A, Johnson DH. Gefitinib in combination with paclitaxel and carboplatin in advanced non-small-cell lung cancer: a phase III trial--INTACT 2. J Clin Oncol 2004;22(5):785-94.
74. Dancey J, Sausville EA. Issues and progress with protein kinase inhibitors for cancer treatment. Nat Rev Drug Discov 2003;2(4):296-313.
75. Cohen P. Protein kinases--the major drug targets of the twenty-first century? Nat Rev Drug Discov 2002;1(4):309-15.
76. Smalley KS, Haass NK, Brafford PA, Lioni M, Flaherty KT, Herlyn M. Multiple signaling pathways must be targeted to overcome drug resistance in cell lines derived from melanoma metastases. Mol Cancer Ther 2006;5(5):1136-44.
77. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001;29(2):153-9.
78. Muller R. Crosstalk of oncogenic and prostanoid signaling pathways. J Cancer Res Clin Oncol 2004;130(8):429-44.
79. Sergina NV, Rausch M, Wang D, Blair J, Hann B, Shokat KM, Moasser MM. Escape from HER-family tyrosine kinase inhibitor therapy by the kinase-inactive HER3. Nature 2007;445(7126):437-41.
80. Christopher M., Overall, Kleifeld O. Validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy. Nature Reviews Cancer 2006;6:227-239.
81. Force T, Krause DS, Van Etten RA. Molecular mechanisms of cardiotoxicity of tyrosine kinase inhibition. Nat Rev Cancer 2007;7(5):332-44.
82. Keith CT, Borisy AA, Stockwell BR. Multicomponent therapeutics for networked systems. Nat Rev Drug Discov 2005;4(1):71-8.

83. Larder BA, Kemp SD, Harrigan PR. Potential mechanism for sustained antiretroviral efficacy of AZT-3TC combination therapy. Science 1995;269(5224):696-9.
84. Zhang X, Crespo A, Fernandez A. Turning promiscuous kinase inhibitors into safer drugs. Trends Biotechnol 2008;26(6):295-301.
85. Krug M, Hilgeroth A. Recent advances in the development of multi-kinase inhibitors. Mini Rev Med Chem 2008;8(13):1312-27.
86. Gill AL, Verdonk M, Boyle RG, Taylor R. 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 2007;7(14):1408-22.
87. Nahta R, Yu D, Hung MC, Hortobagyi GN, Esteva FJ. Mechanisms of disease: understanding resistance to HER2-targeted therapy in human breast cancer. Nat Clin Pract Oncol 2006;3(5):269-80.
88. Tabernero J. The role of VEGF and EGFR inhibition: implications for combining anti-VEGF and anti-EGFR agents. Mol Cancer Res 2007;5(3):203-20.
89. Yamane S, Ishida S, Hanamoto Y, Kumagai K, Masuda R, Tanaka K, Shiobara N, Yamane N, Mori T, Juji T, Fukui N, Itoh T, Ochi T, Suzuki R. Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients. J Inflamm (Lond) 2008;5:5.
90. Maris JM, Courtright J, Houghton PJ, Morton CL, Kolb EA, Lock R, Tajbakhsh M, Reynolds CP, Keir ST, Wu J, Smith MA. Initial testing (stage 1) of sunitinib by the pediatric preclinical testing program. Pediatr Blood Cancer 2008;51(1):42-8.
91. Gozalbes R, Simon L, Froloff N, Sartori E, Monteils C, Baudelle R. Development and experimental validation of a docking strategy for the generation of kinase-targeted libraries. J Med Chem 2008;51(11):3124-32.
92. Deng XQ, Wang HY, Zhao YL, Xiang ML, Jiang PD, Cao ZX, Zheng YZ, Luo SD, Yu LT, Wei YQ, Yang SY. Pharmacophore modelling and virtual screening for identification of new Aurora-A kinase inhibitors. Chem Biol Drug Des 2008;71(6):533-9.
93. Deanda F, Stewart EL, Reno MJ, Drewry DH. Kinase-Targeted Library Design through the Application of the PharmPrint Methodology. J Chem Inf Model 2008;48(12):2395-403.
94. Briem H, Gunther J. Classifying "kinase inhibitor-likeness" by using machine-learning methods. Chembiochem 2005;6(3):558-66.
95. Gundla R, Kazemi R, Sanam R, Muttineni R, Sarma JA, Dayam R, Neamati N. Discovery of novel small-molecule inhibitors of human epidermal growth factor receptor-2: combined ligand and target-based approach. J Med Chem 2008;51(12):3367-77.
96. Morphy R, Rankovic Z. The physicochemical challenges of designing multiple ligands. J Med Chem 2006;49(16):4961-70.
97. Morphy R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds. J Med Chem 2006;49(10):2969-78.
98. Jia J, Zhu F, Ma X, Cao Z, Li Y, Chen YZ. Mechanisms of drug combinations: interaction and network perspectives. Nat Rev Drug Discov 2009;8(2):111-28.
99. Vina D, Uriarte E, Orallo F, Gonzalez-Diaz H. Alignment-free prediction of a drug-target complex network based on parameters of drug connectivity and protein sequence of receptors. Mol Pharm 2009;6(3):825-35.
100. Prado-Prado FJ, Uriarte E, Borges F, Gonzalez-Diaz H. Multi-target spectral moments for QSAR and Complex Networks study of antibacterial drugs. Eur J Med Chem 2009;44(11):4516-21.
101. Gonzalez-Diaz H, Prado-Prado FJ. Unified QSAR and network-based computational chemistry approach to antimicrobials, part 1: multispecies activity models for antifungals. J Comput Chem 2008;29(4):656-67.
102. Gonzalez-Diaz H, Prado-Prado FJ, Santana L, Uriarte E. Unify QSAR approach to antimicrobials. Part 1: predicting antifungal activity against different species. Bioorg Med Chem 2006;14(17):5973-80.

103. Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW. "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 2006;46(6):2445-56.
104. Givehchi A, Bender A, Glen RC. Analysis of activity space by fragment fingerprints, 2D descriptors, and multitarget dependent transformation of 2D descriptors. J Chem Inf Model 2006;46(3):1078-83.
105. Renner S, Derksen S, Radestock S, Morchen F. Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. J Chem Inf Model 2008;48(2):319-32.
106. Erhan D, L'Heureux P J, Yue SY, Bengio Y. Collaborative filtering on a family of biological targets. J Chem Inf Model 2006;46(2):626-35.
107. Dragos H, Gilles M, Alexandre V. Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model 2009;49(7):1762-76.
108. Han LY, Ma XH, Lin HH, Jia J, Zhu F, Xue Y, Li ZR, Cao ZW, Ji ZL, Chen YZ. 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 2008;26(8):1276-86.
109. Liu XH, Ma XH, Tan CY, Jiang YY, Go ML, Low BC, Chen YZ. Virtual screening of Abl inhibitors from large compound libraries by support vector machines. J Chem Inf Model 2009;49(9):2101-10.
110. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 2007;35(Database issue):D198-201.
111. Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. J Chem Inf Model 2005;45(1):177-82.
112. Scior T, Medina-Franco JL, Do QT, Martinez-Mayorga K, Yunes Rojas JA, Bernard P. How to recognize and workaround pitfalls in QSAR studies: a critical review. Curr Med Chem 2009;16(32):4297-313.
113. Susnow RG, Dixon SL. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. J Chem Inf Comput Sci 2003;43(4):1308-15.
114. Perez JJ. Managing molecular diversity. Chemical Society Reviews. Volume 34: Royal Society of Chemistry; 2005. pp 143-152.
115. Willett P, Barnard JM, Downs GM. Chemical Similarity Searching. J Chem Inf Comput Sci 1998;38(6):983-996.
116. Fang H, Tong W, Shi LM, Blair R, Perkins R, Branham W, Hass BS, Xie Q, Dial SL, Moland CL, Sheehan DM. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. Chemical Research in Toxicology 2001;14:280-294.
117. Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. Environmental Health Perspectives 2004;112(12):1249-1254.
118. Hu JY, Aizawa T. Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. Water Research 2003;37(6):1213-1222.
119. Jacobs MN. In silico tools to aid risk assessment of endocrine disrupting chemicals. Toxicology 2004;205(1-2):43-53.
120. Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of Chemical Information and Computer Sciences 2003;43(6):1882-1889.
121. Doniger S, Hofman T, Yeh J. Predicting CNS Permeability of Drug Molecules:Comparison of Neural Network and Support Vector Machine Algorithms. Journal of Computational Biology 2002;9(6):849-864.

122. He L, Jurs PC, Custer LL, Durham SK, Pearl GM. Predicting the Genotoxicity of Polycyclic Aromatic Compounds from Molecular Structure with Different Classifiers. Chemical Research in Toxicology 2003;16(12):1567-1580.
123. Snyder RD, Pearl GS, Mandakas G, Choy WN, Goodsaid F, Rosenblum IY. Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. Environmental and Molecular Mutagenesis 2004;43(3):143-158.
124. Grosios K, Wood J, Esser R, Raychaudhuri A, Dawson J. Angiogenesis inhibition by the novel VEGF receptor tyrosine kinase inhibitor, PTK787/ZK222584, causes significant anti-arthritic effects in models of rheumatoid arthritis. Inflamm Res 2004;53(4):133-42.
125. Hall LH KG, Haney DN. *Molconn-Z*: eduSoft LC: Ashland VA; 2002.
126. Li ZR, Han LY, Xue Y, Yap CW, Li H, Jiang L, Chen YZ. MODEL - Molecular descriptor lab: A web-based server for computing structural and physicochemical features of compounds. Biotechnology and Bioengineering 2007;97(2):389-396.
127. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 2003;43(2):493-500.
128. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr Pharm Des 2006;12(17):2111-20.
129. Wegner JK. JOELib/JOELib2. Department of Computer Science,University of Tübingen: Germany; 2005.
130. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. J Chem Inf Comput Sci 2004;44(5):1630-8.
131. Hemmer MC, Steinhauer V, Gasteiger J. Deriving the 3D structure of organic molecules from their infrared spectra. Vibrational Spectroscopy 1999;19(1):151-164.
132. Rücker G, Rücker C. Counts of all walks as atomic and molecular descriptors. Journal of Chemical Information and Computer Sciences 1993;33(5):683-695.
133. Schuur JH, Setzer P, Gasteiger J. 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 1996;36(2):334-344.
134. Pearlman RS, Smith KM. Metric validation and the receptor-relevant subspace concept. Journal of Chemical Information and Computer Sciences 1999;39(1):28-35.
135. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A. 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 1997;11(1):79-92.
136. Galvez J, Garcia R, Salabert MT, Soler R. Charge indexes. New topological descriptors. Journal of Chemical Information and Computer Sciences 1994;34(3):520-525.
137. Consonni V, Todeschini R, Pavan M. 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 2002;42(3):682-692.
138. Randic M. Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. Tetrahedron 1975;31(11-12):1477-1481.
139. Randic M. Molecular profiles. Novel geometry-dependent molecular descriptors. New Journal of Chemistry 1995;19:781-791.
140. Kier LB, Hall LH. Molecular structure description: The electrotopological state. San Diego: Academic Press; 1999.
141. Platts JA, Butina D, Abraham MH, Hersey A. Estimation of molecular free energy relation descriptors using a group contribution approach. Journal of Chemical Information and Computer Sciences 1999;39(5):835-845.

142. Sadowski J, Gasteiger J, Klebe G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. J Chem Inf Comput Sci 1994;34:1000-1008.
143. Livingstone DJ. Data analysis for chemists: Applications to QSAR and chemical product design. Oxford: Oxford University Press; 1995.
144. Eriksson L, Johansson E, Kettaneh-Wold N, Wade KM. Multi- and megavariate data analysis - Principles and applications. Umea, Sweden: Umetrics, AB; 2001.
145. Vapnik VN. The nature of statistical learning theory. New York: Springer; 1995.
146. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 1998;2(2):127-167.
147. Pochet N, De Smet F, Suykens JA, De Moor BL. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. Bioinformatics 2004;20:3185-3195.
148. Li F, Yang Y. Analysis of recursive gene selection approaches from microarray data. Bioinformatics 2005;21:3741-3747.
149. Yap CW, Chen YZ. Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. J Pharm Sci 2005;94(1):153-68.
150. Yap CW, Chen YZ. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. J Chem Inf Model 2005;45(4):982-92.
151. Grover II, Singh II, Bakshi II. Quantitative structure-property relationships in pharmaceutical research - Part 2. Pharm Sci Technol Today 2000;3(2):50-57.
152. Trotter MWB, Buxton BF, Holden SB. Support vector machines in combinatorial chemistry. Meas Control 2001;34(8):235-239.
153. Czerminski R, Yasri A, Hartsough D. Use of support vector machine in pattern classification: Application to QSAR studies. Quantitative Structure-Activity Relationships 2001;20(3):227-240.
154. Johnson RA, Wichern DW. Applied multivariate statistical analysis. Englewood Cliffs, NJ: Prentice Hall; 1982.
155. Fix E, Hodges JL. Discriminatory analysis: Non-parametric discrimination: Consistency properties. Texas: USAF School of Aviation Medicine; 1951.
156. Fujishima S, Takahashi Y. Classification of dopamine antagonists using TFS-based artificial neural network. J Chem Inf Comput Sci 2004;44(3):1006-9.
157. Bostrom J, Hogner A, Schmitt S. Do structurally similar ligands bind in a similar fashion? J Med Chem 2006;49(23):6716-25.
158. Huang N, Shoichet BK, Irwin JJ. Benchmarking sets for molecular docking. J Med Chem 2006;49(23):6789-801.
159. Mosier PD, Jurs PC. QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 2002;42(6):1460-70.
160. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci 2004;44(1):1-12.
161. Wold S, Eriksson L. Statistical validation of QSAR results. In: Van de Waterbeemd H, editor. Chemometric methods in molecular design. Weinheim; New York: Wiley-VCH; 1995. pp 309-318.
162. Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model 2002;20(4):269-76.
163. Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405(2):442-51.
164. Davies JW, Glick M, Jenkins JL. Streamlining lead discovery by aligning in silico and high-throughput screening. Curr Opin Chem Biol 2006;10(4):343-51.
165. Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 2006;11(23-24):1046-53.
166. van de Waterbeemd H, Gifford E. ADMET in silico modelling: towards prediction paradise? Nat Rev Drug Discov 2003;2(3):192-204.
167. Li H, Ung CY, Yap CW, Xue Y, Li ZR, Chen YZ. Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods. J Mol Graph Model 2006;25(3):313-23.

168. Ghosh S, Nie A, An J, Huang Z. Structure-based virtual screening of chemical libraries for drug discovery. Curr Opin Chem Biol 2006;10(3):194-202.
169. Shoichet BK, McGovern SL, Wei B, Irwin JJ. Lead discovery using molecular docking. Curr Opin Chem Biol 2002;6(4):439-46.
170. Jansen JM, Martin EJ. Target-biased scoring approaches and expert systems in structure-based virtual screening. Curr Opin Chem Biol 2004;8(4):359-64.
171. Li H, Yap CW, Xue Y, Li ZR, Ung CY, Han LY, Chen YZ. Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic or toxicological properties of pharmaceutical agents. Drug Development Research 2006;66(4):245-259.
172. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. Nature 2004;432(7019):855-61.
173. Then RL. Antimicrobial dihydrofolate reductase inhibitors--achievements and future options: review. J Chemother 2004;16(1):3-12.
174. McGuire JJ. Anticancer antifolates: current status and future directions. Curr Pharm Des 2003;9(31):2593-613.
175. Linares GE, Ravaschino EL, Rodriguez JB. Progresses in the field of drug design to combat tropical protozoan parasitic diseases. Curr Med Chem 2006;13(3):335-60.
176. Serretti A, De Ronchi D, Lorenzi C, Berardi D. New antipsychotics and schizophrenia: a review on efficacy and side effects. Curr Med Chem 2004;11(3):343-58.
177. Adler CH, Kumar R. Pharmacological and surgical options for the treatment of cervical dystonia. Neurology 2000;55(12 Suppl 5):S9-14.
178. Hain TC, Uddin M. Pharmacological treatment of vertigo. CNS Drugs 2003;17(2):85-100.
179. Demol P, Ruoff HJ, Weihrauch TR. Rational pharmacotherapy of gastrointestinal motility disorders. Eur J Pediatr 1989;148(6):489-95.
180. H.P. Rang MMD, J.M. Ritter. Pharmacology: Churchill Livingstone; 2001.
181. Li H, Ung C, Yap C, Xue Y, Li Z, Cao Z, Chen Y. Prediction of Genotoxicity of Chemical Compounds by Statistical Learning Methods. Chemical Research in Toxicology 2005;18(6):1071-1080.
182. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ. Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci 2004;44(4):1497-505.
183. Sutherland JJ, O'Brien LA, Weaver DF. Spline-fitting with a genetic algorithm: a method for developing classification structure-activity relationships. J Chem Inf Comput Sci 2003;43(6):1906-15.
184. Bostrom J, Bohm M, Gundertofte K, Klebe G. A 3D QSAR study on a set of dopamine D4 receptor antagonists. J Chem Inf Comput Sci 2003;43(3):1020-7.
185. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 2003;31(13):3692-7.
186. Han LY, Cai CZ, Ji ZL, Cao ZW, Cui J, Chen YZ. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. Nucleic Acids Res 2004;32(21):6437-44.
187. Lin HH, Han LY, Cai CZ, Ji ZL, Chen YZ. Prediction of transporter family from protein sequence by support vector machine approach. Proteins 2006;62(1):218-31.
188. L.Y. Han CJZ, 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 2007;(accepted).
189. Oprea TI, Gottfries J. Chemography: the art of navigating in chemical space. J Comb Chem 2001;3(2):157-66.
190. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci U S A 2005;102(48):17272-7.

191. Ung CY, Li H, Yap CW, Chen YZ. In silico prediction of pregnane X receptor activators by machine learning approaches. Mol Pharmacol 2007;71(1):158-68.
192. Whittle M, Gillet VJ, Willett P, Loesel J. Analysis of data fusion methods in virtual screening: similarity and group fusion. J Chem Inf Model 2006;46(6):2206-19.
193. MICROMEDEX. MICROMEDEX. Greenwood Village, Colorado: MICROMEDEX; Edition expires 12/2003.
194. Bethesda. AHFS drug information: American Society of Health-System Pharmacists, Inc; 2001.
195. Li H, Yap CW, Ung CY, Xue Y, Cao ZW, Chen YZ. Effect of Selection of Molecular Descriptors on the Prediction of Blood-Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. Journal of Chemical Information and Modeling 2005;45(5):1376-1384.
196. Clader JW, Wang Y. Muscarinic receptor agonists and antagonists in the treatment of Alzheimer's disease. Curr Pharm Des 2005;11(26):3353-61.
197. Lipton SA. Pathologically-activated therapeutics for neuroprotection: mechanism of NMDA receptor block by memantine and S-nitrosylation. Curr Drug Targets 2007;8(5):621-32.
198. De Kock MF, Lavand'homme PM. The clinical role of NMDA receptor antagonists for the treatment of postoperative pain. Best Pract Res Clin Anaesthesiol 2007;21(1):85-98.
199. Lepor NE. Anticoagulation for acute coronary syndromes: from heparin to direct thrombin inhibitors. Rev Cardiovasc Med 2007;8 Suppl 3:S9-17.
200. Page MG. Emerging cephalosporins. Expert Opin Emerg Drugs 2007;12(4):511-24.
201. Sepehrdad R, Frishman WH, Stier CT, Jr., Sica DA. Direct inhibition of renin as a cardiovascular pharmacotherapy: focus on aliskiren. Cardiol Rev 2007;15(5):242-56.
202. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44(3):793-806.
203. Eder J, Hommel U, Cumin F, Martoglio B, Gerhartz B. Aspartic proteases in drug discovery. Curr Pharm Des 2007;13(3):271-85.
204. Ripka AS, Rich DH. Peptidomimetic design. Curr Opin Chem Biol 1998;2(4):441-52.
205. Brunton VG, Frame MC. Src and focal adhesion kinase as therapeutic targets in cancer. Curr Opin Pharmacol 2008;8(4):427-32.
206. Lee D, Gautschi O. Clinical development of SRC tyrosine kinase inhibitors in lung cancer. Clin Lung Cancer 2006;7(6):381-4.
207. Hiscox S, Nicholson RI. Src inhibitors in breast cancer therapy. Expert Opin Ther Targets 2008;12(6):757-67.
208. Lin LG, Xie H, Li HL, Tong LJ, Tang CP, Ke CQ, Liu QF, Lin LP, Geng MY, Jiang H, Zhao WM, Ding J, Ye Y. Naturally occurring homoisoflavonoids function as potent protein tyrosine kinase inhibitors by c-Src-based high-throughput screening. J Med Chem 2008;51(15):4419-29.
209. Lee K, Kim J, Jeong KW, Lee KW, Lee Y, Song JY, Kim MS, Lee GS, Kim Y. Structure-based virtual screening of Src kinase inhibitors. Bioorg Med Chem 2009;17(8):3152-61.
210. Farard J, Lanceart G, Loge C, Nourrisson MR, Cruzalegui F, Pfeiffer B, Duflos M. Design, synthesis and evaluation of new 6-substituted-5-benzyloxy-4-oxo-4H-pyran-2-carboxamides as potential Src inhibitors. J Enzyme Inhib Med Chem 2008;23(5):629-40.
211. Alfaro-Lopez J, Yuan W, Phan BC, Kamath J, Lou Q, Lam KS, Hruby VJ. Discovery of a novel series of potent and selective substrate-based inhibitors of p60c-src protein tyrosine kinase: conformational and topographical constraints in peptide design. J Med Chem 1998;41(13):2252-60.
212. Chen P, Doweyko AM, Norris D, Gu HH, Spergel SH, Das J, Moquin RV, Lin J, Wityak J, Iwanowicz EJ, McIntyre KW, Shuster DJ, Behnia K, Chong S, de Fex H, Pang S, Pitt S, Shen DR, Thrall S, Stanley P, Kocy OR, Witmer MR, Kanner SB, Schieven GL, Barrish JC. Imidazoquinoxaline Src-family kinase p56Lck inhibitors: SAR, QSAR, and the discovery of (S)-N-(2-chloro-6-methylphenyl)-2-(3-methyl-1-

piperazinyl)imidazo- [1,5-a]pyrido[3,2-e]pyrazin-6-amine (BMS-279700) as a potent and orally active inhibitor with excellent in vivo antiinflammatory activity. J Med Chem 2004;47(18):4517-29.
213. Altmann E, Missbach M, Green J, Susa M, Wagenknecht HA, Widler L. 7-Pyrrolidinyl- and 7-piperidinyl-5-aryl-pyrrolo[2,3-d]pyrimidines--potent inhibitors of the tyrosine kinase c-Src. Bioorg Med Chem Lett 2001;11(6):853-6.
214. Widler L, Green J, Missbach M, Susa M, Altmann E. 7-Alkyl- and 7-cycloalkyl-5-aryl-pyrrolo[2,3-d]pyrimidines--potent inhibitors of the tyrosine kinase c-Src. Bioorg Med Chem Lett 2001;11(6):849-52.
215. Missbach M, Altmann E, Widler L, Susa M, Buchdunger E, Mett H, Meyer T, Green J. Substituted 5,7-diphenyl-pyrrolo[2,3d]pyrimidines: potent inhibitors of the tyrosine kinase c-Src. Bioorg Med Chem Lett 2000;10(9):945-9.
216. Klutchko SR, Hamby JM, Boschelli DH, Wu Z, Kraker AJ, Amar AM, Hartl BG, Shen C, Klohs WD, Steinkampf RW, Driscoll DL, Nelson JM, Elliott WL, Roberts BJ, Stoner CL, Vincent PW, Dykes DJ, Panek RL, Lu GH, Major TC, Dahring TK, Hallak H, Bradford LA, Showalter HD, Doherty AM. 2-Substituted aminopyrido[2,3-d]pyrimidin-7(8H)-ones. structure-activity relationships against selected tyrosine kinases and in vitro and in vivo anticancer activity. J Med Chem 1998;41(17):3276-92.
217. Noronha G, Barrett K, Boccia A, Brodhag T, Cao J, Chow CP, Dneprovskaia E, Doukas J, Fine R, Gong X, Gritzen C, Gu H, Hanna E, Hood JD, Hu S, Kang X, Key J, Klebansky B, Kousba A, Li G, Lohse D, Mak CC, McPherson A, Palanki MS, Pathak VP, Renick J, Shi F, Soll R, Splittgerber U, Stoughton S, Tang S, Yee S, Zeng B, Zhao N, Zhu H. Discovery of [7-(2,6-dichlorophenyl)-5-methylbenzo [1,2,4]triazin-3-yl]-[4-(2-pyrrolidin-1-ylethoxy)phenyl]amine--a potent, orally active Src kinase inhibitor with anti-tumor activity in preclinical assays. Bioorg Med Chem Lett 2007;17(3):602-8.
218. Liew CY, Ma XH, Liu X, Yap CW. SVM Model for Virtual Screening of Lck Inhibitors. J Chem Inf Model 2009.
219. Ma XH, Wang R, Yang SY, Li ZR, Xue Y, Wei YC, Low BC, Chen YZ. Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. J Chem Inf Model 2008;48(6):1227-37.
220. Ma XH, Jia J, Zhu F, Xue Y, Li ZR, Chen YZ. Comparative Analysis of Machine Learning Methods in Ligand Based Virtual Screening of Large Compound Libraries. Comb Chem High Throughput Screen 2009:(accepted).
221. Chiu YC, Fong YC, Lai CH, Hung CH, Hsu HC, Lee TS, Yang RS, Fu WM, Tang CH. Thrombin-induced IL-6 production in human synovial fibroblasts is mediated by PAR1, phospholipase C, protein kinase C alpha, c-Src, NF-kappa B and p300 pathway. Mol Immunol 2008;45(6):1587-99.
222. Paniagua RT, Sharpe O, Ho PP, Chan SM, Chang A, Higgins JP, Tomooka BH, Thomas FM, Song JJ, Goodman SB, Lee DM, Genovese MC, Utz PJ, Steinman L, Robinson WH. Selective tyrosine kinase inhibition by imatinib mesylate for the treatment of autoimmune arthritis. J Clin Invest 2006;116(10):2633-42.
223. Carvalho JF, Blank M, Shoenfeld Y. Vascular endothelial growth factor (VEGF) in autoimmune diseases. J Clin Immunol 2007;27(3):246-56.
224. Daouti S, Latario B, Nagulapalli S, Buxton F, Uziel-Fusi S, Chirn GW, Bodian D, Song C, Labow M, Lotz M, Quintavalla J, Kumar C. Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis. Osteoarthritis Cartilage 2005;13(6):508-18.
225. Remmers EF, Sano H, Wilder RL. Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue pathology of rheumatoid arthritis. Semin Arthritis Rheum 1991;21(3):191-9.
226. Meyn MA, 3rd, Smithgall TE. Small molecule inhibitors of Lck: the search for specificity within a kinase family. Mini Rev Med Chem 2008;8(6):628-37.
227. Rivera J, Olivera A. Src family kinases and lipid mediators in control of allergic inflammation. Immunol Rev 2007;217:255-68.

228. Lee JH, Kim JW, Ko NY, Mun SH, Kim do K, Kim JD, Won HS, Shin HS, Kim HS, Her E, Kim YM, Choi WS. Mast cell-mediated allergic response is suppressed by Sophorae flos: inhibition of SRC-family kinase. Exp Biol Med (Maywood) 2008;233(10):1271-9.
229. Callera GE, Montezano AC, Yogi A, Tostes RC, He Y, Schiffrin EL, Touyz RM. c-Src-dependent nongenomic signaling responses to aldosterone are increased in vascular myocytes from spontaneously hypertensive rats. Hypertension 2005;46(4):1032-8.
230. Metcalf CA, 3rd, van Schravendijk MR, Dalgarno DC, Sawyer TK. Targeting protein kinases for bone disease: discovery and development of Src inhibitors. Curr Pharm Des 2002;8(23):2049-75.
231. Shakespeare WC, Wang Y, Bohacek R, Keenan T, Sundaramoorthi R, Metcalf C, 3rd, Dilauro A, Roeloffzen S, Liu S, Saltmarsh J, Paramanathan G, Dalgarno D, Narula S, Pradeepan S, van Schravendijk MR, Keats J, Ram M, Liou S, Adams S, Wardwell S, Bogus J, Iuliucci J, Weigele M, Xing L, Boyce B, Sawyer TK. SAR of carbon-linked, 2-substituted purines: synthesis and characterization of AP23451 as a novel bone-targeted inhibitor of Src tyrosine kinase with in vivo anti-resorptive activity. Chem Biol Drug Des 2008;71(2):97-105.
232. Tsuruno S, Kawaguchi SY, Hirano T. Src-family protein tyrosine kinase negatively regulates cerebellar long-term depression. Neurosci Res 2008;61(3):329-32.
233. Hicklin DJ, Ellis LM. Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. J Clin Oncol 2005;23(5):1011-27.
234. Zhong H, Bowen JP. Molecular design and clinical development of VEGFR kinase inhibitors. Curr Top Med Chem 2007;7(14):1379-93.
235. van Cruijsen H, van der Veldt A, Hoekman K. Tyrosine kinase inhibitors of VEGF receptors: clinical issues and remaining questions. Front Biosci 2009;14:2248-68.
236. Roodhart JM, Langenberg MH, Witteveen E, Voest EE. The molecular basis of class side effects due to treatment with inhibitors of the VEGF/VEGFR pathway. Curr Clin Pharmacol 2008;3(2):132-43.
237. Yu H, Wang Z, Zhang L, Zhang J, Huang Q. The discovery of novel vascular endothelial growth factor receptor tyrosine kinases inhibitors: pharmacophore modeling, virtual screening and docking studies. Chem Biol Drug Des 2007;69(3):204-11.
238. Sharma BK, Sharma SK, Singh P, Sharma S. A quantitative structure-activity relationship study of novel, potent, orally active, selective VEGFR-2 and PDGFRalpha tyrosine kinase inhibitors: derivatives of N-phenyl-N'-{4-(4-quinolyloxy)phenyl}urea as antitumor agents. J Enzyme Inhib Med Chem 2008;23(2):168-73.
239. Du J, Lei B, Qin J, Liu H, Yao X. Molecular modeling studies of vascular endothelial growth factor receptor tyrosine kinase inhibitors using QSAR and docking. J Mol Graph Model 2009;27(5):642-54.
240. Dakshanamurthy S, Kim M, Brown ML, Byers SW. In-silico fragment-based identification of novel angiogenesis inhibitors. Bioorg Med Chem Lett 2007;17(16):4551-6.
241. Vieth M, Cummins DJ. DoMCoSAR: a novel approach for establishing the docking mode that is consistent with the structure-activity relationship. Application to HIV-1 protease inhibitors and VEGF receptor tyrosine kinase inhibitors. J Med Chem 2000;43(16):3020-32.
242. Usui T, Ban HS, Kawada J, Hirokawa T, Nakamura H. Discovery of indenopyrazoles as EGFR and VEGFR-2 tyrosine kinase inhibitors by in silico high-throughput screening. Bioorg Med Chem Lett 2008;18(1):285-8.
243. Keseru GM, Makara GM. The influence of lead discovery strategies on the properties of drug candidates. Nat Rev Drug Discov 2009;8(3):203-12.
244. Keseru GM, Makara GM. Hit discovery and hit-to-lead approaches. Drug Discov Today 2006;11(15-16):741-8.

245. Kiselyov AS, Semenov VV, Milligan D. 4-(Azolylphenyl)-phthalazin-1-amines: Novel inhibitors of VEGF receptors I and II. Chem Biol Drug Des 2006;68(6):308-13.
246. Fraley ME, Arrington KL, Hambaugh SR, Hoffman WF, Cunningham AM, Young MB, Hungate RW, Tebben AJ, Rutledge RZ, Kendall RL, Huckle WR, McFall RC, Coll KE, Thomas KA. Discovery and evaluation of 3-(5-thien-3-ylpyridin-3-yl)-1H-indoles as a novel class of KDR kinase inhibitors. Bioorg Med Chem Lett 2003;13(18):2973-6.
247. Kuo GH, Prouty C, Wang A, Emanuel S, Deangelis A, Zhang Y, Song F, Beall L, Connolly PJ, Karnachi P, Chen X, Gruninger RH, Sechler J, Fuentes-Pesquera A, Middleton SA, Jolliffe L, Murray WV. Synthesis and structure-activity relationships of pyrazine-pyridine biheteroaryls as novel, potent, and selective vascular endothelial growth factor receptor-2 inhibitors. J Med Chem 2005;48(15):4892-909.
248. Thompson AM, Delaney AM, Hamby JM, Schroeder MC, Spoon TA, Crean SM, Showalter HD, Denny WA. Synthesis and structure-activity relationships of soluble 7-substituted 3-(3,5-dimethoxyphenyl)-1,6-naphthyridin-2-amines and related ureas as dual inhibitors of the fibroblast growth factor receptor-1 and vascular endothelial growth factor receptor-2 tyrosine kinases. J Med Chem 2005;48(14):4628-53.
249. Nakamura H, Sasaki Y, Uno M, Yoshikawa T, Asano T, Ban HS, Fukazawa H, Shibuya M, Uehara Y. Synthesis and biological evaluation of benzamides and benzamidines as selective inhibitors of VEGFR tyrosine kinases. Bioorg Med Chem Lett 2006;16(19):5127-31.
250. Heyman HR, Frey RR, Bousquet PF, Cunha GA, Moskey MD, Ahmed AA, Soni NB, Marcotte PA, Pease LJ, Glaser KB, Yates M, Bouska JJ, Albert DH, Black-Schaefer CL, Dandliker PJ, Stewart KD, Rafferty P, Davidsen SK, Michaelides MR, Curtin ML. Thienopyridine urea inhibitors of KDR kinase. Bioorg Med Chem Lett 2007;17(5):1246-9.
251. Ruel R, Thibeault C, L'Heureux A, Martel A, Cai ZW, Wei D, Qian L, Barrish JC, Mathur A, D'Arienzo C, Hunt JT, Kamath A, Marathe P, Zhang Y, Derbin G, Wautlet B, Mortillo S, Jeyaseelan R, Sr., Henley B, Tejwani R, Bhide RS, Trainor GL, Fargnoli J, Lombardo LJ. Discovery and preclinical studies of 5-isopropyl-6-(5-methyl-1,3,4-oxadiazol-2-yl)-N-(2-methyl-1H-pyrrolo[2,3-b]pyridin-5-yl)pyrrolo[2,1-f][1,2,4]triazin-4-amine (BMS-645737), an in vivo active potent VEGFR-2 inhibitor. Bioorg Med Chem Lett 2008;18(9):2985-9.
252. Peifer C, Selig R, Kinkel K, Ott D, Totzke F, Schachtele C, Heidenreich R, Rocken M, Schollmeyer D, Laufer S. Design, synthesis, and biological evaluation of novel 3-aryl-4-(1H-indole-3yl)-1,5-dihydro-2H-pyrrole-2-ones as vascular endothelial growth factor receptor (VEGF-R) inhibitors. J Med Chem 2008;51(13):3814-24.
253. Hennequin LF, Thomas AP, Johnstone C, Stokes ES, Ple PA, Lohmann JJ, Ogilvie DJ, Dukes M, Wedge SR, Curwen JO, Kendrew J, Lambert-van der Brempt C. Design and structure-activity relationship of a new class of potent VEGF receptor tyrosine kinase inhibitors. J Med Chem 1999;42(26):5369-89.
254. Hasegawa M, Nishigaki N, Washio Y, Kano K, Harris PA, Sato H, Mori I, West RI, Shibahara M, Toyoda H, Wang L, Nolte RT, Veal JM, Cheung M. Discovery of novel benzimidazoles as potent inhibitors of TIE-2 and VEGFR-2 tyrosine kinase receptors. J Med Chem 2007;50(18):4453-70.
255. Ji Z, Ahmed AA, Albert DH, Bouska JJ, Bousquet PF, Cunha GA, Glaser KB, Guo J, Li J, Marcotte PA, Moskey MD, Pease LJ, Stewart KD, Yates M, Davidsen SK, Michaelides MR. Isothiazolopyrimidines and isoxazolopyrimidines as novel multi-targeted inhibitors of receptor tyrosine kinases. Bioorg Med Chem Lett 2006;16(16):4326-30.
256. David M, Friedrich L, Kurt H. The support vector machine under test. Neurocomputing 2003;55(1-2):169-186.
257. Fernandez M, Mejias M, Angermayr B, Garcia-Pagan JC, Rodes J, Bosch J. Inhibition of VEGF receptor-2 decreases the development of hyperdynamic splanchnic circulation and portal-systemic collateral vessels in portal hypertensive rats. J Hepatol 2005;43(1):98-103.

258. Yamamoto A, Watanabe H, Sueki H, Nakanishi T, Yasuhara H, Iijima M. Vascular endothelial growth factor receptor tyrosine kinase inhibitor PTK787/ZK 222584 inhibits both the induction and elicitation phases of contact hypersensitivity. J Dermatol 2007;34(7):419-29.
259. Enomoto H, Inoki I, Komiya K, Shiomi T, Ikeda E, Obata K, Matsumoto H, Toyama Y, Okada Y. Vascular endothelial growth factor isoforms and their receptors are expressed in human osteoarthritic cartilage. Am J Pathol 2003;162(1):171-81.
260. Paavonen K, Mandelin J, Partanen T, Jussila L, Li TF, Ristimaki A, Alitalo K, Konttinen YT. Vascular endothelial growth factors C and D and their VEGFR-2 and 3 receptors in blood and lymphatic vessels in healthy and arthritic synovium. J Rheumatol 2002;29(1):39-45.
261. Kahl KG, Bens S, Ziegler K, Rudolf S, Kordon A, Dibbelt L, Schweiger U. Angiogenic factors in patients with current major depressive disorder comorbid with borderline personality disorder. Psychoneuroendocrinology 2009;34(3):353-7.
262. Iga J, Ueno S, Yamauchi K, Numata S, Tayoshi-Shibuya S, Kinouchi S, Nakataki M, Song H, Hokoishi K, Tanabe H, Sano A, Ohmori T. Gene expression and association analysis of vascular endothelial growth factor in major depressive disorder. Prog Neuropsychopharmacol Biol Psychiatry 2007;31(3):658-63.
263. Warner-Schmidt JL, Duman RS. VEGF as a potential target for therapeutic intervention in depression. Curr Opin Pharmacol 2008;8(1):14-9.
264. Prado-Prado FJ, de la Vega OM, Uriarte E, Ubeira FM, Chou KC, Gonzalez-Diaz H. Unified QSAR approach to antimicrobials. 4. Multi-target QSAR modeling and comparative multi-distance study of the giant components of antiviral drug-drug complex networks. Bioorg Med Chem 2008.
265. Zhang X, Fernandez A. In silico drug profiling of the human kinome based on a molecular marker for cross reactivity. Mol Pharm 2008;5(5):728-38.
266. Gockel I, Moehler M, Frerichs K, Drescher D, Trinh TT, Duenschede F, Borschitz T, Schimanski K, Biesterfeld S, Herzer K, Galle PR, Lang H, Junginger T, Schimanski CC. Co-expression of receptor tyrosine kinases in esophageal adenocarcinoma and squamous cell cancer. Oncol Rep 2008;20(4):845-50.
267. Stommel JM, Kimmelman AC, Ying H, Nabioullin R, Ponugoti AH, Wiedemeyer R, Stegh AH, Bradner JE, Ligon KL, Brennan C, Chin L, DePinho RA. Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. Science 2007;318(5848):287-90.
268. Speake G, Holloway B, Costello G. Recent developments related to the EGFR as a target for cancer chemotherapy. Curr Opin Pharmacol 2005;5(4):343-9.
269. Moasser MM. Targeting the function of the HER2 oncogene in human cancer therapeutics. Oncogene 2007;26(46):6577-92.
270. Lewis NL. The platelet-derived growth factor receptor as a therapeutic target. Curr Oncol Rep 2007;9(2):89-95.
271. Rusnati M, Presta M. Fibroblast growth factors/fibroblast growth factor receptors as targets for the development of anti-angiogenesis strategies. Curr Pharm Des 2007;13(20):2025-44.
272. Benati D, Baldari CT. SRC family kinases as potential therapeutic targets for malignancies and immunological disorders. Curr Med Chem 2008;15(12):1154-65.
273. Schwartz GK, Shah MA. Targeting the cell cycle: a new approach to cancer therapy. J Clin Oncol 2005;23(36):9408-21.
274. Medina M, Castro A. Glycogen synthase kinase-3 (GSK-3) inhibitors reach the clinic. Curr Opin Drug Discov Devel 2008;11(4):533-43.
275. Irwin JJ, Shoichet BK, Mysinger MM, Huang N, Colizzi F, Wassam P, Cao Y. Automated docking screens: a feasibility study. J Med Chem 2009;52(18):5712-20.
276. Derksen S, Rau O, Schneider P, Schubert-Zsilavecz M, Schneider G. Virtual screening for PPAR modulators using a probabilistic neural network. ChemMedChem 2006;1(12):1346-50.
277. Noble ME, Endicott JA, Johnson LN. Protein kinase inhibitors: insights into drug design from structure. Science 2004;303(5665):1800-5.

278. Vema A, Panigrahi SK, Rambabu G, Gopalakrishnan B, Sarma JA, Desiraju GR. Design of EGFR kinase inhibitors: a ligand-based approach and its confirmation with structure-based studies. Bioorg Med Chem 2003;11(21):4643-53.
279. Yu H, Wang Z, Zhang L, Zhang J, Huang Q. Pharmacophore modeling and in silico screening for new KDR kinase inhibitors. Bioorg Med Chem Lett 2007;17(8):2126-33.
280. Matsuno K, Ichimura M, Nakajima T, Tahara K, Fujiwara S, Kase H, Ushiki J, Giese NA, Pandey A, Scarborough RM, Lokker NA, Yu JC, Irie J, Tsukuda E, Ide S, Oda S, Nomoto Y. 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 2002;45(14):3057-66.
281. Thompson AM, Connolly CJ, Hamby JM, Boushelle S, Hartl BG, Amar AM, Kraker AJ, Driscoll DL, Steinkampf RW, Patmore SJ, Vincent PW, Roberts BJ, Elliott WL, Klohs W, Leopold WR, Showalter HD, Denny WA. 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 2000;43(22):4200-11.
282. Dalgarno D, Stehle T, Narula S, Schelling P, van Schravendijk MR, Adams S, Andrade L, Keats J, Ram M, Jin L, Grossman T, MacNeil I, Metcalf C, 3rd, Shakespeare W, Wang Y, Keenan T, Sundaramoorthi R, Bohacek R, Weigele M, Sawyer T. Structural basis of Src tyrosine kinase inhibition with a new class of potent and selective trisubstituted purine-based compounds. Chem Biol Drug Des 2006;67(1):46-57.
283. Abbott L, Betschmann P, Burchat A, Calderwood DJ, Davis H, Hrnciar P, Hirst GC, Li B, Morytko M, Mullen K, Yang B. Discovery of thienopyridines as Src-family selective Lck inhibitors. Bioorg Med Chem Lett 2007;17(5):1167-71.
284. Showalter HD, Sercel AD, Leja BM, Wolfangel CD, Ambroso LA, Elliott WL, Fry DW, Kraker AJ, Howard CT, Lu GH, Moore CW, Nelson JM, Roberts BJ, Vincent PW, Denny WA, Thompson AM. 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 1997;40(4):413-26.
285. Asano T, Yoshikawa T, Usui T, Yamamoto H, Yamamoto Y, Uehara Y, Nakamura H. Benzamides and benzamidines as specific inhibitors of epidermal growth factor receptor and v-Src protein tyrosine kinases. Bioorg Med Chem 2004;12(13):3529-42.
286. Caballero J, Fernandez M, Saavedra M, Gonzalez-Nilo FD. 2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-d]pyrimidine derivatives. Bioorg Med Chem 2008;16(2):810-21.
287. Kuo GH, Wang A, Emanuel S, Deangelis A, Zhang R, Connolly PJ, Murray WV, Gruninger RH, Sechler J, Fuentes-Pesquera A, Johnson D, Middleton SA, Jolliffe L, Chen X. Synthesis and discovery of pyrazine-pyridine biheteroaryl as a novel series of potent vascular endothelial growth factor receptor-2 inhibitors. J Med Chem 2005;48(6):1886-900.
288. Apsel B, Blair JA, Gonzalez B, Nazif TM, Feldman ME, Aizenstein B, Hoffman R, Williams RL, Shokat KM, Knight ZA. Targeted polypharmacology: discovery of dual inhibitors of tyrosine and phosphoinositide kinases. Nat Chem Biol 2008;4(11):691-9.
289. Peach ML, Nicklaus MC. Combining docking with pharmacophore filtering for improved virtual screening. J Cheminform 2009;1(1):6.
290. Andrae J, Gallini R, Betsholtz C. Role of platelet-derived growth factors in physiology and medicine. Genes Dev 2008;22(10):1276-312.
291. Raj T, Kanellakis P, Pomilio G, Jennings G, Bobik A, Agrotis A. Inhibition of fibroblast growth factor receptor signaling attenuates atherosclerosis in apolipoprotein E-deficient mice. Arterioscler Thromb Vasc Biol 2006;26(8):1845-51.